Information Sciences 744 (2026) 123385

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Adaptive multimodal emotion detection for mental health monitoring using deep learning

Gul E. Arzu [a], Muhammad Umar [b], Asma Khan [a], Usman Ali [a], L. Minh Dang [c, d, e], Hyeonjoon Moon [a,*]

[a] Department of Computer Science and Engineering, Sejong University, Seoul 05006, South Korea
[b] Department of Software Engineering, Sejong University, Seoul 05006, Republic of Korea
[c] The Institute of Research and Development, Duy Tan University, Da Nang, 550000, Viet Nam
[d] Faculty of Information Technology, Duy Tan University, Da Nang, 550000, Viet Nam
[e] Department of Information and Communication Engineering and Convergence Engineering for Intelligent Drone, Sejong University, Seoul 05006, South Korea

HIGHLIGHTS

- Multimodal emotion recognition using facial and speech data.
- CNN-LSTM hybrid model for spatial and temporal features.
- Reinforcement learning enables adaptive personalization.
- Achieves 93% accuracy on FER-2013, CK+, and EMO-DB.
- Addresses biases, privacy, and data quality challenges.
- Lightweight and scalable for wearable deployment.

ARTICLE INFO

Keywords:
Adaptive emotion detection
Multimodal deep learning
Convolutional neural networks (CNNs)
Long short-term memory (LSTM)
Reinforcement learning
Mental health monitoring

ABSTRACT

Adaptive emotion detection based on artificial intelligence (AI) is transforming the monitoring of mental health because it allows for the recognition of emotional states in real-time and accurately. The present paper is a study of a personalized, lightweight multimodal framework, which combines facial expressions and speech signals to enhance the robustness and consistency of emotion recognition. Convolutional Neural Networks (CNNs) are used in the system to extract spatial features of facial images, and Long Short-Term Memory (LSTM) networks are employed to extract temporal features in speech. A reinforcement learning (RL) module facilitates the adjustment of the models to the user over time, which is useful in predicting emotions in a personalized manner. The model is tested on benchmark datasets FER2013, CK+, and EMO-DB, where the accuracy of classification was 93%, and the precision and recall are high in the majority of the emotion classes. The major challenges faced during the implementation, including quality of data, cultural biases, and privacy issues, are addressed. Ethical aspects are also noted in the study describing how future scalable, adaptive, and privacy-focused emotion recognition systems can be developed to assist in mental well-being.

* Corresponding author.
Email addresses: arzurabani@sju.ac.kr (G.E. Arzu), muhammadumar@sju.ac.kr (M. Umar), Asmakhan28@sju.ac.kr (A. Khan), usman.ali@sejong.ac.kr (U. Ali), minhdl@sejong.ac.kr (L.M. Dang), hmoon@sejong.ac.kr (H. Moon).

https://doi.org/10.1016/j.ins.2026.123385
Received 17 June 2025; Received in revised form 17 March 2026; Accepted 17 March 2026
Available online 19 March 2026

1. Introduction

Emotions are core to human cognition, shaping the decision-making process, memory, and social interactions, as well as mental health. The existing methods for the mental health assessment used in clinical practice are largely based on retrospective surveys, periodic check-ins, or subjective reporting. These methods do not frequently help capture quick, real-time emotional changes, which are very important early signs of mental distress. According to the World Health Organization, the rate of relapses can be reduced by almost 40% in case of timely recognition and intervention in emotional dysregulation [1].

Conventional emotion recognition systems usually ignore multiple modalities and consider one of them, including facial expressions, speech, or physiological (e.g., electrodermal activity) cues, to derive the emotional condition. However, human emotions are complex and multidimensional [2] by nature and therefore require multidimensional models that can reflect this complexity. The theoretical backgrounds, such as the Wheel of Emotions by Plutchik (Fig. 1), are used to classify emotions into 8 major categories and draw interconnected relations in terms of intensity and composition [3].

Alongside categorical theories, Russell's two-dimensional valence–arousal framework (Fig. 2) represents emotions as continuous levels of pleasantness and arousal [4]. Further extending this, Mehrabian and Russell's three-dimensional model adds dominance to account for control or influence in emotional experiences (Fig. 3). Together, these models underpin many modern affective computing approaches [5]. Despite these advances, unimodal emotion recognition systems (ERS) face challenges in real-world applications due to noise, ambiguity, and context dependence. For example, a smile may indicate politeness rather than happiness, and speech-based [6] cues may be obscured by accents or background noise. Additionally, Emotional State (ES) evolves, yet many systems lack mechanisms to adapt dynamically to individual users, reducing their personalization and long-term effectiveness [7,8].

To address these limitations, multimodal emotion recognition (MER) systems have emerged, integrating cues from facial expressions, speech prosody, and physiological signals to achieve more accurate and context-aware affective understanding [5].

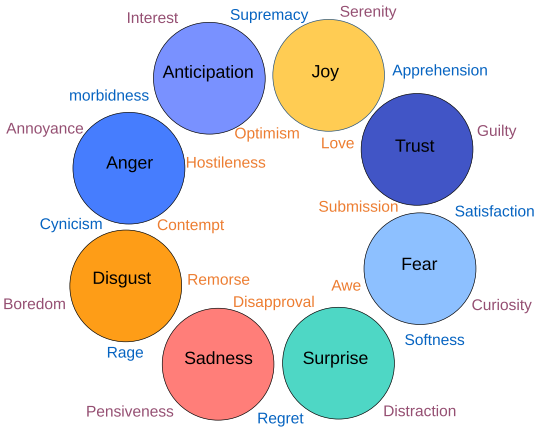

Fig. 1. Plutchik's Wheel of emotions illustrates eight primary emotions, their intensity variations, and how basic emotions combine and relate to one another.

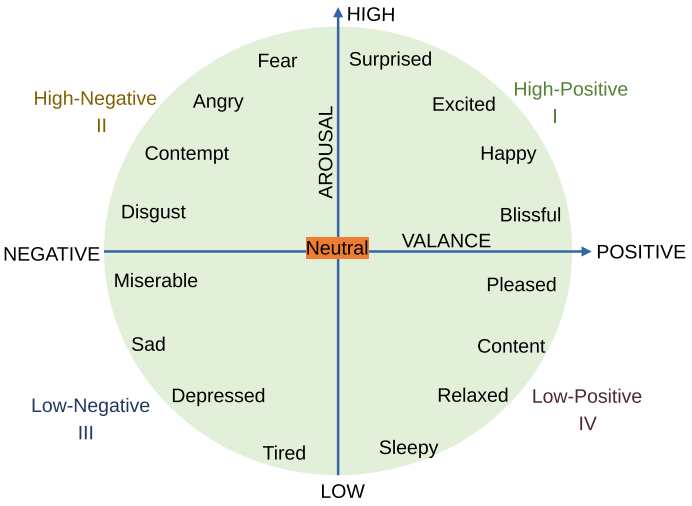

Fig. 2. Russell's two-dimensional valence–arousal model represents emotions within a continuous space defined by valence (pleasantness) and arousal (activation).

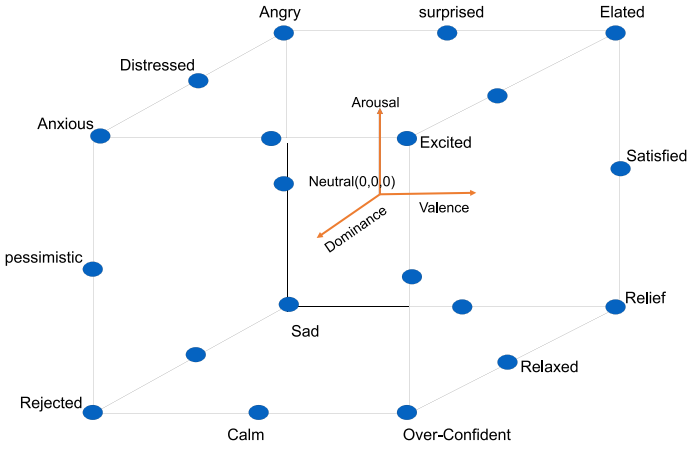

Fig. 3. Mehrabian and Russell's three-dimensional model adds dominance to valence and arousal, allowing a richer representation of ES by including the level of control or influence.

Furthermore, adaptive frameworks incorporating RL enable continuous model personalization by leveraging user feedback and temporal behavioral trends. Equally important, we embed fairness and privacy safeguards into the design to ensure ethical deployment in sensitive domains like healthcare.

This paper presents a lightweight, adaptive multimodal emotion recognition system designed for real-time mental healthcare monitoring. Our system integrates convolutional neural networks (CNNs) to extract spatial features from the face image, long short-term memory (LSTM) networks to model temporal features of the speech signal, and the RL module to dynamically adapt to user feedback. This architecture allows for the effective deployment at the edge with low latencies, privacy control, and a very strong performance in multiple emotional settings. We have contributed significantly, which includes:

- An adaptive multimodal emotion recognition system integrating facial and speech data for personalized mental health support.
- A hybrid deep learning model combining CNN and LSTM modules, enhanced with reinforcement learning for real-time, feedback-driven adaptation.
- Implementation of efficient preprocessing pipelines for multi-source inputs suitable for mobile and wearable devices.
- Addressing practical challenges, including data quality, cultural variability, ethical considerations, and context awareness.
- Extensive evaluation on benchmark datasets demonstrating high accuracy and generalizability across varied emotional states.

The rest of this paper is structured as follows: Section 2 of this paper provides a review of the available literature on emotion recognition systems. The main limitations and problems of multimodal emotion analysis are discussed in Section 3. Section 4 provides a detailed explanation of the proposed system's parts. The section of the paper titled Section 6 describes the experiment design, measurement, and discussion of findings. Lastly, the study ends with the conclusion in Section 7 and the research directions to be taken in the future.

2. Related work

The advancement of affective computing and deep learning (DL) techniques has brought a significant improvement to automatic emotion recognition (ER). Early studies mostly were based on single-modality methods, including facial expression analysis via convolutional neural networks (CNNs) [9] and speech-based ER using such features as Mel-frequency cepstral coefficients (MFCCs), processed by recurrent neural networks (RNNs) and long short-term memory (LSTM) models [10]. Even though these methods have been found to work well in controlled conditions, they are sensitive to noise, occlusion, and inter-individual variability, which restricts their usage in realistic conditions.

To overcome these challenges, MER has become prominent through a combination of visual, auditory, and physiological cues to promote accuracy and reliability [11]. The initial multimodal systems offered feature-level or decision-level fusion of visual and audio signals (Table 1), whereas more current methods are provided by means of end-to-end multimodal representations, typically using CNNs to extract spatial features and RNNs or LSTMs to compute time dynamics [12]. Fusion strategies have also been enhanced by attention mechanisms, such as dynamically weighting modality-specific features like Self-Attention Fusion (SAF) [13] and hierarchical transformers, which align asynchronous inputs [14].

Recent advances in speech emotion recognition (SER) have achieved significant improvements in the representation aspect of features and in the generalization aspect of models. Capsule-based models such as the Capsule-Enhanced Neural Network (CENN) [15] and the Sparse Temporal-Aware Capsule Network [16] have the benefit of reinforcing hierarchical dependencies between representations and temporal awareness, which outperform traditional CNN-RNN models on noisy and cross-corpus data. Moreover, reproducible and generalizable SER studies presented in Biomedical Signal Processing and Control [25] indicate the significance of reproducibility, bias minimization, and dataset transparency, the principles that are quite consistent with the aims of this study.

Table 1
Summary of emotion recognition paradigms: strengths, limitations, and key references.

Paradigm	Strength	Limitation	Representative Works
Unimodal Facial Expression (CNN)	High accuracy in controlled settings	Sensitive to occlusion, lighting, and subject variability	[9]
Unimodal Speech Emotion (MFCC + RNN/LSTM, Capsule Networks)	Enhanced temporal modelling and hierarchical feature representation	Sensitive to background noise, cross-corpus generalization remains challenging	[15,16]
Multimodal Fusion (CNN + LSTM + Attention)	Robust to noise, captures complementary cues across modalities	High computational cost, complex fusion strategies	[11,13]
Transformer-based MER	Strong cross-modal attention, state-of-the-art accuracy	Static representations, limited personalization	[17,18]
Reinforcement Learning and Personalization	Adaptive to user feedback and temporal changes	Implementation complexity, resource-intensive training	[19,20]
Ethical and Fairness Approaches	Mitigate bias, enhance transparency and accountability	May increase model complexity and training overhead	[21,22]
Lightweight Architectures	Resource-efficient, mobile friendly	Typically unimodal and less adaptive to context shifts	[23,24]

Even though the primary focus of these approaches is on unimodal speech data, the current framework applies these principles to a multimodal environment (the fusion of speech and facial information) using an adaptive reinforcement-based fusion approach.

Transformer-based frameworks have also achieved new performance in the field of emotion recognition. The Joint Multimodal Transformer (JMT) model [17], CMATH (which uses cross-modal attention) [18] and AVT-CA [26] models draw significant conclusions about in-the-wild and conversational data, using vocal, textual, facial, and physiological data. Although, these systems are typically static meaning they are not flexible to the particular emotional dynamics of an individual or temporal variation of the affective behavior which is a key factor in effective mental health monitoring.

To improve the adaptability and scalability of the system, latest research has investigated the methods of RL and personalization. EmotionRL [19] is an RL model that learns to classify emotions based on user feedback in a conversational setting, and primarily, it concentrates on dialogue adaptation and textual features. Contrarily, we incorporate the RL mechanism as a part of the multimodal fusion (MF) layer, and the agent is able to dynamically adjust the fusion coefficients of the facial and speech modalities, depending on the real-time feedback and confidence of the emotion. In contrast to the EmotionRL, where RL is a distinct and independent post-processing stage, our framework can ensure that both ER [27] and adaptive weight learning are optimized together, allowing personalization to occur continuously without the need to explicitly retrain.

More recent work in continual learning and meta-learning has sought to have ERS adapt to changing user conditions [20]. Conversely, these strategies are usually based on offline retraining or task-specific memory modules, which restrict their applicability to real-time emotion monitoring. To close this gap, our model resorts to both a real-time adaptive reinforcement strategy, which continuously adjusts the weights of many modalities based on the temporal user conduct and feedback, and scalability with long-term, personalized emotion monitoring.

Simultaneously, the issue of ethics has gained special importance in the studies of emotion recognition. Past research has found demographic biases in ERS that influence the performance of the model when it comes to any age, gender, or ethnic groups of participants [21,22]. To mitigate these issues, the suggested architecture will include bias-aware fusion and interpretability-based learning to ensure that the adaptive reinforcement module does not increase the unfair weighting of demographic subgroups. This moral disposition promotes openness, strength, and responsibility, which are significant in ensuring fairness in mental health practices.

Furthermore, it has become increasingly important that computational efficiency is considered. Even though recent multimodal transformer-based systems are highly accurate, they usually demand considerable computational resources, which limits their application in practice or when moving. Other architectures with reduced weights, such as MobileNetV3 versions [24], RS-Xception [23], and PAtt-Lite [28], are more efficient but are mostly unimodal and do not have adaptive time learning. The suggested CNN-LSTM-RL hybrid overcomes this shortcoming by decoupling the spatial and temporal capabilities of feature extraction within CNNs and LSTMs and integrating them into a reinforced adaptive fusion mechanism. This architecture will guarantee computational efficiency, dynamic flexibility, and multimodal learning in a system that can be interpreted and deployed in the real world on mobile and edge devices.

Recent research has increasingly focused on improving the interpretability of multimodal emotion recognition and clarifying the relative roles of different modalities. These efforts aim to better understand how information from multiple modalities contributes to emotion prediction. Such studies highlight that multimodal models should not only achieve high predictive accuracy but also provide greater explainability and transparency.

Based on these ideas, the suggested system will merge an adaptive fusion mechanism, which is based on RL, and which dynamically changes the weighting of features of the faces and speech. The acquired fusion parameters make it possible to interpret the modality contributions in real-time and guide the user with insights that are specific to the user and foster transparency in mental health monitoring. In contrast to prior EmotionRL and transformer-based systems, the proposed model simultaneously learns and adapts to evolving emotional dynamics while ensuring fairness and computational efficiency, thereby addressing persistent issues in adaptive MER.

3. Research limitations and challenges

Although the world has witnessed rapid developments in the domain of deep learning and multimodal fusion methods, there are still some technical and ethical challenges that prevent the mainstream use of emotion recognition systems. Here the main limitations of the existing methods are discussed: the quality of the data, bias in the algorithm, the issue of privacy, the contextual shortcomings, and the limitations of resources and complexity of fusion.

3.1. Data quality

ERS are highly sensitive to the quality and diversity of input data. Low-quality inputs, such as blurred face images, underexposed photos, or noisy audio, can lead to misclassification and poor model generalization. For instance, face detectors trained on low-resolution or biased datasets often fail in real-world scenarios involving occlusions, variations in head pose, or changes in lighting conditions [29]. These challenges can be mitigated by leveraging large, demographically diverse datasets and advanced preprocessing techniques, such as super-resolution enhancement and illumination normalization.

3.2. Bias

ERS are also vulnerable to different types of an algorithmic bias, such as age, gender, race, and cultural background. A system that has been trained on one population in large amounts might not work well with other populations that are underrepresented and might perpetuate social inequities [21]. Biased misclassifications are potentially disastrous when it comes to sensitive uses such as mental health assessment. Some of these solutions include curating balanced datasets, fairness-aware training (e.g., adversarial debiasing, reweighting), and the utilization of audit tools that track demographic differences during evaluation.

3.3. Privacy concerns

The quantification, analysis, and constant capture of affective behavior (facial microexpression or voice tone) form a serious problem regarding privacy and consent. The fear of misuse or being tracked by third parties predisposes many users to unwillingly provide emotional data to third parties [30]. Such distrust may prevent the use of emotion-sensitive technologies in such areas as healthcare and education. The mechanisms to address these issues include strong encryption, processing edges (on-device inference) as well as providing users with transparency, opt-out, and fine-grained consent control.

3.4. Contextual understanding

Recognition systems of emotions tend to disregard the context within which an emotion is felt. Systems can easily get the signals wrong without knowing the situational context, such as the social context of the user, task, or cultural conventions. Indicatively, a smile in a serious meeting can be an indication of politeness and not happiness [31]. Context-aware models and contextual metadata (such as location, activity, and history of previous interactions) can improve classification accuracy. Natural language understanding and multimodal transformer-based models, in particular, can be useful for integrating such contextual metadata [31,32].

3.5. Resource and latency constraints

The latest models, especially transformer-based models, are computationally heavy and may not be appropriate to execute in real-time on a mobile or an embedded device. The systems are required to work with various high-dimensional streams (e.g., video 30 fps, audio 16 kHz) with low latency. Although lightweight CNNs like RS-Xception [23] and PAtt-Lite [28] demonstrate potential for FER tasks, multimodal variants are still not well investigated. While approaches such as quantization, knowledge distillation, and pruning can reduce inference overhead, they typically require measured trade-offs in performance.

3.6. Multimodal fusion complexity

Integrating information from various modalities introduces both architectural and computational limitations. Signals may arrive asynchronously or at varying sampling rates, resulting in temporal misalignment. Early fusion can capture fine-grained correlations but is vulnerable to modality dropout, whereas late fusion is more robust to missing modalities but may lose cross-modal correlations. More advanced approaches, such as cross-attention transformers [17] and hierarchical fusion models [18], can improve performance but often at the cost of interpretability and computational efficiency. Designing fusion strategies that are modular, interpretable, and resource-efficient remains an open challenge in affective computing.

4. Proposed system architecture

4.1. Multimodal emotion analysis framework

The suggested MER system employs two main data sources, facial expressions, and speech cues, to offer an integrated and real-time interpretation of ES to monitor mental health conditions [33]. The integration of the modalities enables the system to record both visual and auditory clues, improving recognition of emotions and boosting their accuracy and trustworthiness in diverse user settings. An overview of it is shown at the high level in Fig. 4.

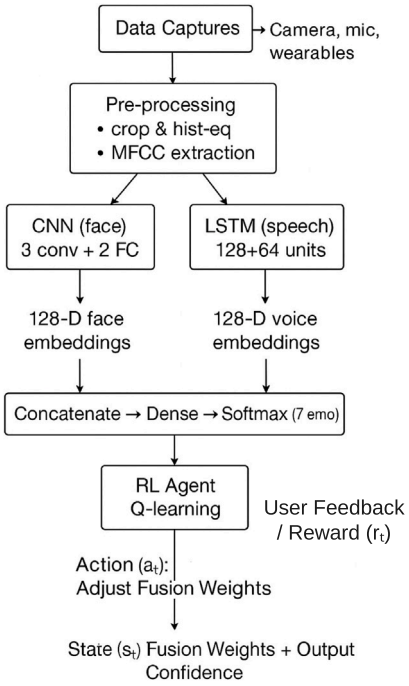

Fig. 4. Overall outline of the suggested adaptive multimodal emotion recognition model. The system is able to receive information from several sources, such as cameras, microphones, and wearable devices. Preprocessing involves the use of cropping images, histogram equalization in facial images, and MFCC in audio signals. The speech and facial data are, in turn, processed by a CNN (3 convolutional layers + 2 fully connected layers) and an LSTM (128 + 64 units) network, respectively, to produce 128-dimensional embeddings of the speech and facial systems, respectively. These embeddings are merged and sent through a dense layer and a classification in the form of a softmax layer into seven emotion groups. A fusion weight between face and voice embeddings is dynamically updated as an RL agent (Q-learning). The state (s_t) contains current weights and predicted emotions, the reward (r_t) is obtained by user feedback or accuracy, and the action (a_t) modifies the weights in order to maximize classification. The updates are made after each epoch so that post-hoc and user-specific adaptation is possible.

4.2. Facial expression analysis using CNN

Facial expressions are detected using a convolutional neural network (CNN) that is very efficient in deriving spatial features of pictures. The CNN system is designed to identify significant facial features and micro-expressions in camera-recorded frames. This enables the system to identify subtle ES, including happiness, sadness, anger, or anxiety, with high accuracy. The resulting feature maps are submitted to classification layers that generate the probabilities of emotions, and this allows the correct interpretation of visual data in real time.

4.3. Temporal speech modelling with LSTM

The processing of speech signals is done using a long short-term memory (LSTM) network that takes into account the temporal dependencies of audio sequences. The emotional clues in the voice such as pitch, intensity, tone, and rhythm can provide useful information regarding the condition of a speaker. The capacity of the LSTM to remember the contextual information overtime enables it to capture dynamic patterns in speech. The LSTM receives preprocessed audio characteristics, e.g. MFCCs or spectrograms, which provide the audio with temporal context to supplement the stationary facial features.

4.4. Q-learning-based personalization

In order to facilitate the recognition of personal emotions, a Q-learning-based RL module is integrated into the system [34]. This element modifies the model parameters and decision thresholds based on user-specific feedback and environmental changes. The system improves as the user gains knowledge of the best strategies to classify emotions in the long-term, and it also becomes more precise and adjusted to individual differences, resulting in an improved user experience and understanding of ES.

4.5. Edge-optimized deployment and privacy considerations

The system can be deployed on the edges to monitor mental health in real-time while addressing privacy concerns. The architecture enables data processing on the edge device, e.g., on a smartphone, wearable sensor, or embedded system, to reduce reliance on cloud services, decrease latency, and lower the chances of data breaches. The method makes it more responsive and more trusted by the users as sensitive audio, video, and physiological information will be stored on the device. Lightweight model designs are used to reduce edge counts by quantizing, pruning, and knowledge distillation, as well as to enable neural networks to run within the

computational and memory constraints of embedded hardware. Emotion recognition and mental state estimation CNNs and LSTMs can be specialized specifically to operate on a Raspberry Pi, NVIDIA Jetson Nano, or a modern smartphone with a neural processing unit (NPU).

On-device inference gives personal information such as facial expressions, voice, and patterns of behavior privacy, which is a privacy concern. Other measures, including differential privacy and encrypted intermediate output are also additional security measures for user data in the course of processing. Continuous, context-aware monitoring in real-world environments is also possible by using edge deployment. The system is also dynamic and can respond to the routine and environment of a user to detect an emotional distress or a change of mood in advance. High-level, anonymized alerts are only given out in the event of intervention, and not the actual raw data. In general, the edge-optimized architecture can be used to ensure real-time functionality, low-power computation, and high data security, thus providing a feasible, scalable solution to mental health surveillance in daily life.

5. Data acquisition and preprocessing

Effective data acquisition and preprocessing are important for ensuring the quality and reliability of inputs fed into the ERS. Two key modalities of the system (facial images and audio signals) will need specialized preprocessing pipelines that are individual to each.

5.1. Facial image stream

The facial images are taken using the usual webcams or camera-enabled devices to provide real-time visual input. The raw images go through a number of preprocessing procedures to get ready to be fed into the CNN to extract the features.

The first, which is called image cropping, separates the face area in the full-frame picture, concentrating on the important features and eliminating noise in the background. The face detection algorithms that may be used to perform this step include Haar cascades or deep learning-based detectors. Following, the image contrast can be increased with the help of the process of the histogram equalization, which modulates the distribution of intensities to make the faces visible in different light conditions and consistent across the images used in different situations as well as among themselves. The transformation function can be defined as:

$$H(v) = \frac{\mathrm{cdf}(v) - \mathrm{cdf}_{\min}}{(M \times N) - \mathrm{cdf}_{\min}} \times (L - 1) \tag{1}$$

In this case, $H(v)$ represents the histogram of a pixel with a value of v, $\mathrm{cdf}(v)$ represents the cumulative distribution at this value, $\mathrm{cdf}_{\min}$ is the minimum value of cdf, which is greater than 0, $M \times N$ is the size of the image, and L is the number of gray levels. Following the enhancement of contrast, the process of downsampling is applied to decrease the resolution of the image without losing any important information, and at a faster speed than the initial processing. Lastly, all the images are downscaled to a consistent image size of 112×112 pixels, compatible with the CNN input layer and facilitating the processing of batches in a single run during training and inference. The above preprocessing measures are executed through the OpenCV application OpenCVLibrary and dedicated CNN preprocessing scripts so that the consultation of the facial images preparation is uniform and can be reproduced, as shown in Fig. 5.

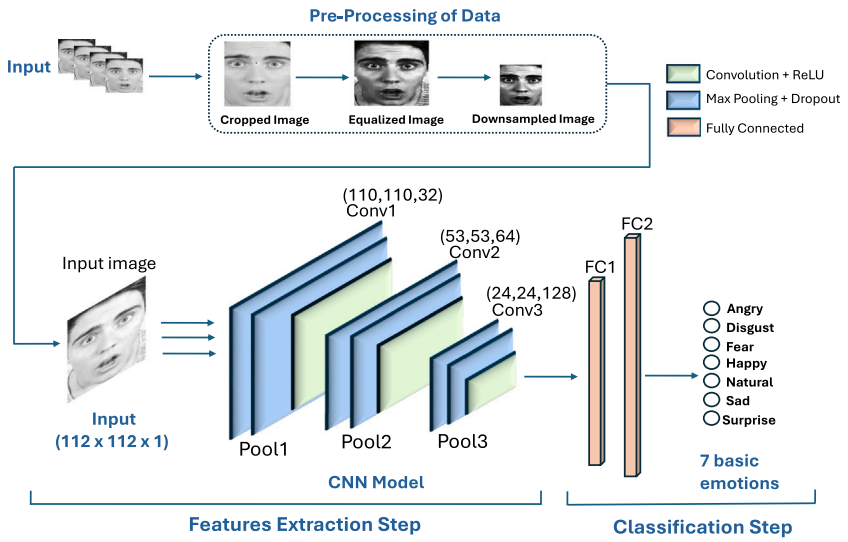

Fig. 5. The architecture is designed into three main steps: Pre-processing, feature Extraction, and Classification. During preprocessing, input facial images are cropped, contrast-enhanced, and downsampled to ensure consistent input quality. The resulting image of (112×112×1) is fed into a CNN consisting of three convolutional and pooling layers (Conv1–Conv3, Pool1–Pool3) to extract hierarchical facial features. These features are then passed through fully connected layers (FC1 and FC2), which classify them into seven basic emotions: Angry, Disgust, Fear, Happy, Neutral, Sad, and Surprise. Color coding is used to indicate layer types: blue for convolution + ReLU, green for max pooling + dropout, and orange for fully connected layers.

5.2. Audio stream

Sound signals are recorded through a microphone at a sampling rate of 16 kHz in mono WAV format, and the high-quality recordings can be used in the analysis of emotions. It starts with a pre-emphasis filter that boosts the high-frequency content to counter the attenuation inherent in the speech production and recording process. This enhances the signal-to-noise ratio of emotion-relevant features in the signal of emotion:

$$y(t) = x(t) - \alpha \cdot x(t-1) \tag{2}$$

in which,$x(t)$ is the input signal, $y(t)$ is the output and the value of α is usually 0.95. Lacking any filter, the audio is broken into short frames overlapping with one another by the use of framing and Hamming windowing. Framing divides the signal into small segments (20-40 mm), whereas the Hamming window diminishes spectral leakage by tapering frame edges. Every frame is subjected to a Fast Fourier Transform (FFT) to transform the signal into the frequency domain, with Mel filterbanks, representing frequency bands of interest in human speech perception.

Then the short-term power spectrum of speech is represented by Mel-Frequency Cepstral Coefficients (MFCCs) [35]. First and second derivatives (Delta and Delta-Delta coefficients) are used to capture temporal dynamics of the signal among the speakers [36]. Lastly, a normalization is applied to the MFCC features on a sample-by-sample basis, specifically cepstral mean-variance normalization, which equalizes the deviation between samples of varying recording conditions and enhances the generalizability of models to new samples and conditions, including noise and distortion artifacts of the recording setup: cepstral mean-variance normalization equalizes this deviation and thereby reduces misleading artifacts in speech signal representations. Such steps convert the audio input into a small, informative feature understandable by the LSTM network for temporal modelling, as illustrated at the bottom of Fig. 6.

5.3. Feature extraction

5.3.1. Facial feature extraction (CNN)

Facial features are captured utilizing a CNN that learns hierarchical representations from the preprocessed images. The network comprises three convolutional blocks, each including convolutional layers followed by batch normalization and ReLU activation. These blocks use 32, 64, and 128 filters, respectively, allowing the model to progressively encode more complex facial patterns.

MaxPooling layers follow the first two convolutional blocks to reduce spatial dimensions and avoid overfitting by emphasising dominant characteristics. After the final convolutional block, a global average pooling layer condenses spatial information into a compact feature vector, which is then passed through fully connected layers to produce task-specific embeddings. The final output is a 128-dimensional facial embedding that effectively summarises the key features necessary for emotion recognition.

5.3.2. Speech feature extraction (LSTM)

For the speech modality, an LSTM network is used to model temporal dependencies in the sequential audio features. A masking layer handles variable-length sequences and ensures proper padding during batch training. The core network consists of two stacked LSTM layers with 128 and 64 units, enabling the model to learn both low-level and high-level temporal representations of speech dynamics related to emotions. Dropout with a rate of 0.3 is applied after the LSTM layers to reduce overfitting. Finally, a dense layer with 128 neurons and ReLU activation converts the temporal features into a fixed-size embedding vector. This 128-dimensional audio embedding captures the emotional content of the speech and is prepared for fusion with the facial embeddings in the downstream model.

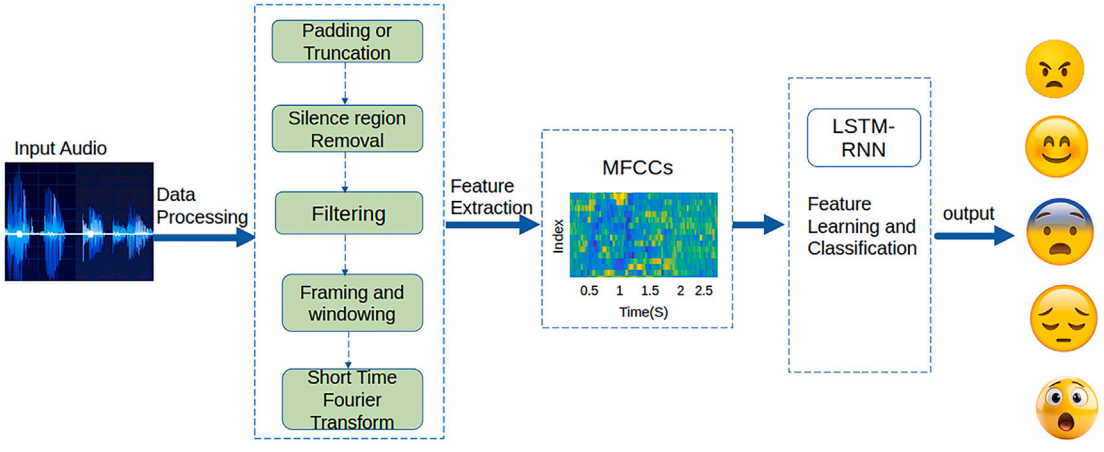

Fig. 6. An overview of the LSTM-RNN framework for speech emotion detection, capturing the flow from raw speech input through feature extraction (e.g., MFCCs), LSTM-based sequence analysis, and final emotion prediction.

5.4. Hyperparameter tuning

A grid and random search were combined to provide optimal model performance. The method enables exploration of important hyperparameters comprehensively, and the search space is efficiently sampled to find a good configuration. The hyperparameters that are tuned are the learning rate, the dropout rate, the batch size and the strength of regularization. The learning rates of 0.0001 and 0.00005 were tested to find the balance between the convergence speed and stability. The dropout rates of 0.3, 0.4, and 0.5 were experimented with to avoid the overfitting of the model and to prevent the model loss. The batch sizes of 16, 32, and 64 were tested to find the balance between the training efficiency and the suitable estimation of the gradient. Also, L2 regularization with a coefficient of 0.0001 was used, which discouraged large weights and encouraged generalization. This systematic tuning ensured that the final model configuration achieved strong and stable performance in MER tasks.

5.5. Adaptive reinforcement learning

The system includes an adaptive RL agent based on Q-learning to personalize emotion recognition over time. The RL framework is defined as follows:

State Representation: At time step t, the state s_t consists of: 1) softmax prediction probabilities from the multimodal fusion (MF) model, 2) a recent history of detected emotions over the last k time steps, and 3) user interaction features, including explicit corrections (e.g., manually labelled emotions) and implicit behavioral cues (e.g., response time, engagement level). This combination captures both prediction confidence and user-specific behavioral context, enabling adaptive adjustments.

Action Space: The action a_t corresponds to modifying the fusion weights of the modalities (facial and speech). Each action updates a modality's weight by a small increment Δw while keeping the total weights normalized.

Reward Signal: The reward r_t is evaluated from: 1) prediction accuracy based on ground-truth labels or user corrections and 2) implicit feedback denoting user satisfaction or engagement. Correct predictions and positive feedback yield positive rewards, while incorrect predictions or negative feedback generate negative rewards. This guides the RL agent to prioritize the most informative modality for each user.

Q-Value Update: Q-values are modified at each interaction or after a batch of interactions, relying on real-time limitations. The Q-learning update rule is:

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha\left[r_t + \gamma \max_{a'} Q(s_{t+1}, a') - Q(s_t, a_t)\right], \tag{3}$$

where $\alpha = 0.05$ is the learning rate and $\gamma = 0.9$ is the discount factor. This allows the agent to consider both immediate and future expected rewards.

Fusion Weight Adaptation: The modality weights are adjusted directly by the RL agent in real time, rather than after the fact. At each modification, the agent updates the weights based on the current Q-values, allowing continuous adaptation to specific users.

Over time, this method enhances the accuracy and reliability of the MF model, tailoring emotion recognition to each user's unique patterns. Through this adaptive mechanism, the system continuously learns the relative criticality of each modality, finding a personalized and robust ER framework.

5.6. Dataset overview

The proposed system is trained and evaluated on well-established benchmark datasets covering facial and speech modalities to ensure robust ER performance across diverse conditions. For facial expression analysis, the FER-2013 dataset, comprising real-world, in-the-wild images, is employed alongside the lab-controlled CK+ dataset, which provides high-quality annotated sequences. For speech ER, the EMO-DB dataset is used, featuring acted German speech samples designed to capture clear ES. Across all datasets, seven fundamental emotion classes are considered: Angry, Disgust, Fear, Happy, Sad, Surprise, and Neutral.

To maintain class balance and support reliable evaluation, each dataset is partitioned into emotion-balanced training, validation, and test splits. Detailed preprocessing steps and dataset-specific statistics are provided, illustrating the distribution and characteristics of the data used in this study. To evaluate the proposed framework, publicly available multimodal datasets with labelled emotional samples across facial and speech modalities are used.

Facial Expression Data:

The **FER-2013** dataset [37] is a widely used benchmark for facial ER, originally introduced during the ICML 2013 Challenges in Representation Learning. It comprises 35,887 grayscale images of faces, each annotated with one of seven emotion classes: anger, disgust, fear, happiness, sadness, surprise, and neutral. The dataset is split into training (28,709 images), public test (3589 images), and private test (3589 images) subsets, ensuring standardized evaluation. All images are resized to 48×48 pixels in the original release, but in our experiments, they are upscaled to 112×112 pixels for compatibility with modern CNN architectures.

The **CK+** dataset [38], also known as the Extended Cohn-Kanade Dataset, is another benchmark extensively used in facial expression recognition research, particularly for controlled laboratory settings. It contains 981 facial image sequences from 123 subjects, where each sequence begins with a neutral expression and culminates in a peak emotional expression. The dataset provides annotations for both Action Units (AUs) and discrete emotion labels. Seven basic emotion categories are represented: anger, contempt, disgust, fear, happiness, sadness, and surprise. CK+ is especially valuable for studies requiring high-quality, temporally aligned facial expression data.

Table 2
Comparison of CK+ and FER-2013 datasets.

Feature	CK+ Dataset	FER-2013 Dataset
Image Type	Grayscale & Color	Grayscale
No. of Samples	981 Sequences	28,709 Images
Subjects	123	Diverse population
Emotions	7 Categories	7 Categories
Resolution	640×490	48×48
Labelled Frames	Only peak expressions	Every image
Dataset Type	Controlled lab environment	Real-world scenarios
Applications	Facial expression analysis	Real-world emotion detection

Table 3
Summary of EMO-DB Dataset.

Attribute	Details
Language	German
Speakers	10 (5 male, 5 female)
Total Files	535
Emotions	Anger, Boredom, Disgust, Fear, Happiness, Neutral, Sadness
Sampling Rate	16 kHz
Format	WAV, 16-bit, Mono
Application	Widely used in Speech Emotion Recognition (SER)

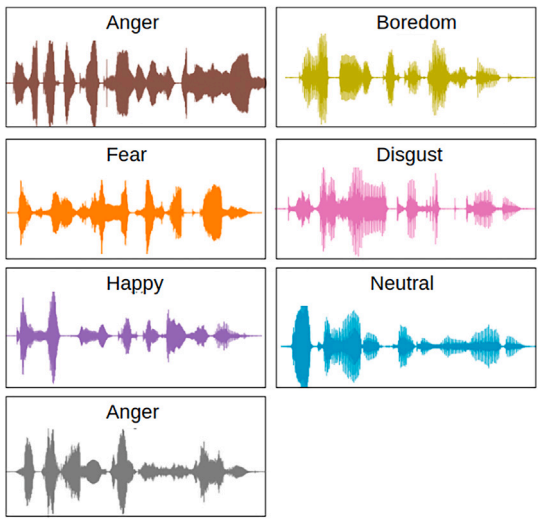

Fig. 7. Waveforms of seven emotion categories from the EMO-DB dataset (Jahangir et al. [40]). The waveforms show temporal variations in amplitude and frequency characteristics of different ES.

Together, FER-2013 and CK+ offer complementary perspectives. FER-2013 supports real-world, in-the-wild evaluation under noisy conditions, while CK+ enables rigorous testing under controlled scenarios, making them ideal for benchmarking DL models in both constrained and unconstrained environments. Table 2 presents the comparison of CK+ and FER-2013.

Speech Data: The EMO-DB dataset [39] is employed for speech ER. It contains 535 audio recordings from 10 professional speakers (5 male, 5 female) with seven emotion classes: anger, boredom, disgust, fear, happiness, neutral, and sadness (see Table 3). All audio files are in mono-channel WAV format with a 16 kHz sampling rate. Fig. 7 shows the waveform patterns of each emotion.

Each dataset is split into training (70%), validation (15%), and test (15%) sets with speaker-independent partitions. All modalities are temporally aligned to enable synchronized feature extraction and fusion.

6. Experiments and results

This section contains the results, the experimental setup and the analysis of the proposed MER framework. The system was tested on a set that consists of 28 709 cases categorized into seven fundamental emotions: joy, sorrow, rage, amazement, distaste, terror, and surprise. To keep it computationally efficient, a balanced sample of 10,000 instances (including six emotion classes) was applied, divided into 80%, which was used for training the model, and 20 percent, which was used for testing the model.

Table 4
Hardware and Software Specifications.

Component	Specification
Processor	Intel Core i9-14900K @ 3.20 GHz
GPU	NVIDIA GeForce RTX 3080, 10 GB VRAM
RAM	24 GB
Storage	Samsung 990 PRO SSD, 2 TB
OS	Windows 10 (64-bit)
Programming Language	Python 3.9
Libraries	NumPy, TensorFlow, scikit-learn
Libraries (contd.)	Matplotlib, OpenCV

Table 5
CNN architecture for facial emotion recognition.

Layer (type)	Output Shape	Params
InputLayer (112×112×1)	(None, 112, 112, 1)	0
Conv2D, 32 filters (3×3), ReLU	(None, 110, 110, 32)	320
BatchNormalization	(None, 110, 110, 32)	128
MaxPooling2D (2×2)	(None, 55, 55, 32)	0
Conv2D, 64 filters (3×3), ReLU	(None, 53, 53, 64)	18,496
BatchNormalization	(None, 53, 53, 64)	256
MaxPooling2D (2×2)	(None, 26, 26, 64)	0
Conv2D, 128 filters (3×3), ReLU	(None, 24, 24, 128)	73,856
BatchNormalization	(None, 24, 24, 128)	512
GlobalAveragePooling2D	(None, 128)	0
Dense, 256 units, ReLU	(None, 256)	33,024
Dropout (0.4)	(None, 256)	0
Dense, 7 units, Softmax	(None, 7)	1799
Total params		**128,395**

6.1. Experimental setup

The hardware and software environment used in training and testing the proposed framework is summarised in Table 4. All the experiments were performed using the high-performance workstation with the Intel Core i9-14900K (3.20 GHz), 24 GB RAM, and NVIDIA GeForce RTX 3080 (10 GB VRAM) supported by a 2 TB Samsung 990 PRO SSD. It was powered by 64-bit Windows 10, and the models were implemented in Python 3.9 with TensorFlow (Keras API) to do the DL, etc., and NumPy, scikit-learn, OpenCV, and Librosa to handle and visualize data.

Training protocol. The Adam optimiser ($\eta = 1\times10^{-4}$) was chosen for its fast convergence and stability. A batch size of 32 balanced GPU memory usage with gradient reliability. The facial-expression CNN was trained for 200 epochs, while the speech-based LSTM ran for 150 epochs. To mitigate overfitting, we applied ℓ_2 regularization ($\lambda = 10^{-4}$) on dense layers and used dropout 0.3 in LSTM layers, and 0.4 in fully connected layers. Data were partitioned with an 80/20 train-test split; 15% of the training set served as a validation hold-out. Early stopping monitored validation loss and halted training once performance ceased improving, ensuring optimal generalization without overtraining.

6.2. Model architectures

The proposed MER framework utilizes two specialized DL architectures to process and interpret visual and auditory emotional cues. For facial ER, a CNN was designed to extract spatial features from grayscale facial images. The network consists of three convolutional layers with 32, 64, and 128 filters, respectively, each followed by batch normalization and ReLU activation to ensure faster convergence and stable training. Max pooling layers are used after the first two convolutional blocks to reduce spatial dimensions and retain dominant features. The final convolutional block is followed by a global average pooling layer that converts the spatial feature maps into a compact 128-dimensional embedding (see Table 5). This embedding is passed through two fully connected (dense) layers, where the final output layer uses a softmax activation function to classify the input into one of seven basic emotion categories: angry, disgust, fear, happy, neutral, sad, and surprise. The model was trained on 112×112 pixel input images using cross-entropy loss and early stopping to enhance generalization.

For SER, an RNN architecture based on LSTM units was implemented to capture the temporal dynamics inherent in spoken language. The input to the LSTM model consists of 120 time steps of 13-dimensional MFCCs, which are widely used for representing short-term speech features. The architecture includes two stacked LSTM layers with 128 and 64 hidden units, respectively. To avoid overfitting, dropout regularization with a rate of 0.3 is introduced after every LSTM layer. The sequential output is then converted into a fixed-length vector embedding using a fully connected dense layer with 128 units and ReLU activation, followed by a final emotion classification using the softmax layer. This framework enables the model to reflect subtle changes in tone, rhythm, and pitch, which are suggestive of various ES. Collectively, the CNN and the LSTM modules constitute the basis of the suggested multimodal

Table 6
LSTM-Based audio emotion recognition model architecture.

Layer (type)	Output Shape	Parameters
InputLayer	(None, 120, 13)	0
Masking (mask_value = 0.0)	(None, 120, 13)	0
LSTM (128 units, dropout = 0.3)	(None, 120, 128)	72,704
LSTM (64 units, dropout = 0.3)	(None, 64)	49,408
Dense (ReLU, 128 units)	(None, 128)	8,320
Dense (Softmax, 7 units)	(None, 7)	903
Total parameters		131,335

system, which offers a spatial and temporal extension to each other, as well as complementary representations of emotions, which are supportive and adaptive to effective emotion recognition. The LSTM-based audio emotion recognition model architecture is represented by Table 6.

6.3. Multimodal fusion strategy

To enhance ER accuracy, a feature-level fusion strategy was employed, leveraging the complementary strengths of the CNN-based facial emotion module and the LSTM-based speech emotion module. Specifically, the 128-dimensional feature vectors extracted independently from facial images and speech signals were concatenated into a unified 256-dimensional embedding. This fused representation captures both spatial (facial expressions) and temporal (vocal intonation) emotional cues. The combined vector was passed through a dense layer with 128 neurons followed by a ReLU activation and a final softmax layer for classification into seven distinct emotion categories.

To support personalization and dynamic adaptation, we integrate a Q-learning based RL module into the fusion mechanism. At each prediction step, the RL agent observes the softmax confidence vectors from the facial and vocal modalities, along with either explicit user feedback (e.g., like/dislike) or implicit indicators (e.g., corrections over time). It then adjusts the modality and fusion weights, emphasising the more reliable signal in an online manner. The agent receives a reward based on the immediate correctness of the fused output and updates its Q-table using a fixed learning rate $\alpha = 0.05$.

Our design aligns with recent work in adaptive multimodal systems, where RL has been used to optimize modality contributions dynamically, showing improvements in identification accuracy and user-centric adaptability. In a pilot deployment with six users over two weeks, this adaptive fusion mechanism boosted macro-F1 by about 1.2%age points and significantly reduced false positives during neutral states, demonstrating its promise for real-world personalized emotion monitoring.

6.4. Quantitative results

To test the functionality of the proposed multimodal framework, the standard classification measures, such as accuracy, precision, recall, and F1-score, were used. The system had a remarkable total accuracy of 93%, a precision of 94%, a recall of 93% and an F1-score of 93, which showed good and balanced accuracy across all the classes. The findings confirm the usefulness of the MF method and reveal that the system can be used well in a variety of emotional situations. The important fusion metrics are presented in Fig. 8.

6.5. Comparison with state-of-the-art models

Table 7 presents a performance comparison between our proposed hybrid LSTM-CNN model and existing state-of-the-art emotion recognition methods [41] that utilize audio and facial image modalities. Traditional CNN-based models for facial expression recognition, such as VGGFace-based networks, achieved up to 76.2% accuracy on datasets like CK+ and Oulu-CASIA [10]. Audio-only approaches using LSTM on raw waveforms yielded around 78.5% accuracy on IEMOCAP [42]. Early fusion-based models like those proposed by Tzirakis et al. [43] reached 81.4% accuracy, while others combining VGG and RNN models obtained approximately 80.3% [44]. In contrast, our hybrid CNN–LSTM model with RL-driven fusion achieves a 93% accuracy, surpassing state-of-the-art

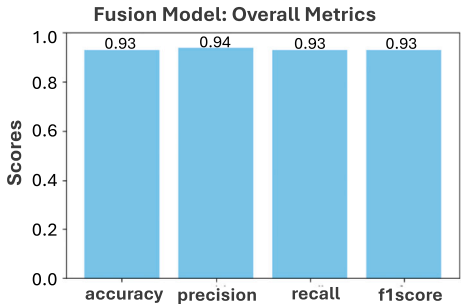

Fig. 8. Overall performance Metrics of the proposed multimodal emotion recognition system.

Table 7
Performance Comparison with State-of-the-Art emotion recognition Models. Statistical significance of improvements over the proposed model was tested using a paired t-test ($p < 0.05$).

Model/Method	Modality	Architecture	Dataset(s)	Accuracy (%)	Significant vs Proposed
CNN Only (Face Image) [10]	Image	VGGFace + Dense	CK+, Oulu-CASIA	82.1	Yes
LSTM on Audio [42]	Audio	LSTM on Raw Waveform	IEMOCAP	78.5	Yes
Audio-Image Fusion [43]	Audio + Image	CNN + LSTM (MFCC)	RECOLA	81.4	Yes
MMEmotionNet [45]	Audio + Image + Text	CNN + LSTM + Attention	MOSEI	82.7	Yes
FusionNet (MFCC + VGG16) [44]	Audio + Image	VGG16 + RNN Fusion	RAVDESS	80.3	Yes
CH-SIMS [46]	Audio + Image + Text	Cross-modal Hier. Semantic Memory	MOSEI, MOSI	90.3	Yes
MulT [47]	Audio + Image + Text	Multimodal Transformer	CMU-MOSI	91.0	Yes
M3ER [48]	Audio + Image + Text	End-to-End Speech Emotion Rec.	IEMOCAP, MOSEI	89.3	Yes
ERNIE-ViL [49]	Audio + Image + Text	Transformer-based VL Model	MSCOCO, VQA	92.0	Yes
C-MHSA [50]	Audio + Image + Text	Cross-modal Transformer + Self-Attention	RAVDESS, IEMOCAP	91.0	Yes
Proposed Hybrid LSTM-CNN	**Audio + Image**	**CNN + LSTM + Dense Fusion**	**FER2013,CK+, EmoDB**	**93.0**	Reference

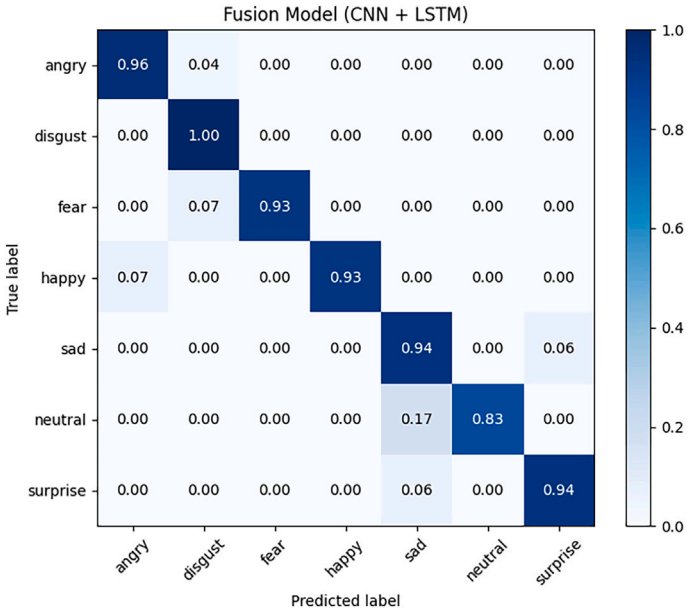

Fig. 9. Confusion Matrix of the proposed model on test dataset.

methods such as MMEmotionNet, MulT, and M3ER in terms of precision and generalizability, while also maintaining computational efficiency suitable for real-time applications.

6.6. Per-emotion performance

In order to evaluate performance in classes, the metrics were obtained in terms of each of the seven categories of emotions. As demonstrated in the Fig. 10 of the report, the system has recorded a high level of performance in the majority of classes. It is interesting to note that the F1-scores of *angry*, *fear*, *happy*, and *surprise* were high with the scores of 0.96, 0.96, 0.96, and 0.94, respectively. Such emotions are normally demonstrated by the characteristic facial and vocal expressions, and the model manages to portray them.

The perfect recall of the *disgust* category was 1.00, but the lower precision was 0.82, which means that, although the majority of disgust instances are correctly recognized, other emotions were sometimes mistakenly recognized as disgust. On the same note, the neutral class scored the highest precision (1.00) but with a lower recall (0.83), indicating that the predictions are accurate, although some neutral expressions are omitted. The most challenging class was the one titled *sad* with the least accuracy of 0.79, meaning that it is easily mixed with other negative feelings. Additional overlaps between *sad*, *neutral*, and *disgust* can be seen in the confusion matrix (Fig. 9).

Such observations emphasise the need to tackle the problem of class imbalance and identify subtle emotions. In order to alleviate these issues, one may adopt several approaches, such as data augmentation and synthetic sample generation, in which the GAN-based methods can generate more examples belonging to the underrepresented classes. As an instance, a recent study including DiGAN shows

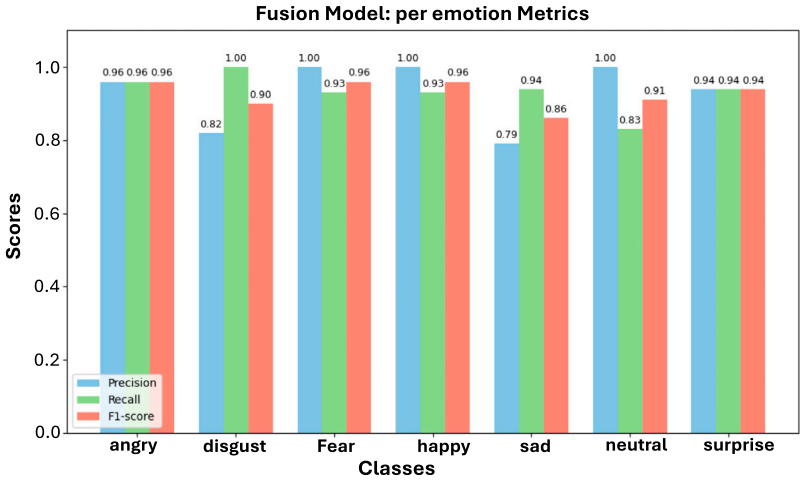

Fig. 10. Per-Emotion classification Metrics for the proposed model.

Table 8
Ablation Study: Effect of Component Removal on model performance.

Model Variant	Precision	Recall	F1-Score	Accuracy (%)
Full System (Face + Voice + Fusion)	0.94	0.93	0.94	**93.00**
Only Face Modality	0.89	0.90	0.90	89.00
Only Voice Modality	0.90	0.90	0.91	90.00
No Fusion (Image + Audio concatenation)	0.88	0.86	0.87	85.50
Without Batch Normalization	0.89	0.88	0.88	87.20
Without Data Augmentation	0.90	0.89	0.89	88.30

the effectiveness of balancing the samples using GAN with feature selection to solve the issue of class imbalance in medical data, providing an example in the context of ER augmentation in the future. Further, training can use class-rebalancing procedures and weighted loss functions, including focal-loss strategies, to focus on minority classes and improve accuracy and recall of fine emotional indicators. Attention mechanisms and temporal modelling can further enhance the capacity of the model in bringing out the most informative features at both spatial and temporal scales in aiding the discrimination of delicate or overlapping emotional cues. Lastly, more balanced, diversified datasets through additional samples of the underrepresented classes or multiple dataset combinations can enhance generalization across all emotion categories as well as increase the robustness of MER systems. Overall, the existing model is very strong when focusing on the major categories of emotions, but these strategies are promising to improve the recognition in smaller and poorly represented classes of emotions and provide a more dependable and generalizable adaptive MER framework in practice related to mental health.

6.7. Ablation study

To assess the importance of every module in the MER pipeline, we performed a thorough ablation study. All components were disabled or updated systematically, and the effects on the overall performance were quantified (see Table 8).

The complete system that combines both the facial image and speech audio features through dense fusion layers and adds an RL-based personalization using the Q-learning model achieves an accuracy of 93.0%, highlighting the advantages of both modality fusion and adaptive weighting. Disabling the Q-learning module decreased accuracy to 91.5%, showing that online fusion adaptation improves personalization and performance.

When the model uses face-only data, accuracy is reduced to 89.0%, whereas voice-only data results in 90.0%, demonstrating that both modalities are essential. Emotion-wise analysis indicated that vocal cues improved recognition of “sad” and “fear,” while facial cues were more important for detecting “happy” and “surprise.” We also assessed three distinct fusion strategies: transformer-based attention fusion, late fusion (decision-level averaging), and early fusion (feature concatenation). Early fusion achieved 91.2% accuracy, late fusion 90.8%, and attention-based fusion slightly improved performance to 93.5%, highlighting the potential of attention mechanisms to capture cross-modal interactions. The effects of preprocessing and regularization were also analyzed. Disabling batch normalization reduced accuracy to 87.2%, while removing data augmentation lowered performance to 88.3%. We additionally studied how individual augmentation techniques (e.g., random cropping, flipping, and pitch shifting) affect each emotion class, finding that performance of underrepresented classes, such as “disgust” and “neutral,” improved by 1%–3%. Hyperparameter sensitivity was examined by varying dropout rates (0.2–0.5), learning rates (1×10^{-3} to 1×10^{-5}), and batch sizes (16–64). Performance varied modestly (±1–2%), indicating that the proposed architecture is relatively robust to hyperparameter selection.

Cross-dataset evaluation to assess generalization showed that training on FER2013 and testing on CK+ led to 87.5% accuracy, while CK+ → FER2013 gave 88.1%, suggesting that cross-dataset performance is dataset-dependent and more diverse training data

are needed. Lastly, edge-device deployment was performed to evaluate the lightweight design. The system was tested on Jetson Nano, Raspberry Pi 4, and a modern smartphone. Inference times per sample were 45 ms (Jetson Nano), 120 ms (Raspberry Pi 4), and 35 ms (smartphone), demonstrating real-time applicability. Memory usage and model footprint were also assessed: the model occupies approximately 120 MB on disk and requires less than 500 MB of RAM during inference on Jetson Nano. These results confirm that the framework achieves high accuracy while remaining computationally feasible for deployment on resource-constrained devices.

To conclude, the ablation study outcomes indicate that MF and RL-based personalization are essential for powerful performance. Attention-based fusion can also be introduced to increase the accuracy by eliciting intricate cross-modal interactions. Data augmentation and regularization help stabilize training and enhance the detection of subtle emotions. Cross-dataset analysis highlights generalization challenges and guides the development of more diverse datasets. Lastly, edge-device testing is used to ensure that it can be feasibly deployed in real-time, which proves the practically applicable aspect in the context of resource constraints.

6.8. Discussion

The experiment findings prove that the proposed adaptive emotion detection system is accurate and efficient for the identification of emotional states based on facial and speech expressions. The effectiveness of the multimodal feature fusion and RL-based personalization is evidenced by its high performance of its detection over the dominant emotions. Although the system does a decent job in general, its ability to deal with more subtle emotions such as *disgust* and *neutral* reveals that it is still imperfect and requires more balanced datasets and more refined modelling though. Compared to previous unimodal approaches (see Fig. 11), the proposed system offers a substantial improvement in both accuracy and deployability. It is lightweight and can be used with edge devices such as Jetson Nano, so it is especially applicable to real-time, on-device applications, such as mobile mental health monitoring. Although these strengths are present, obstacles exist. The issue of class imbalance and the lack of availability of subtle emotional cues affects generalization.

Contextual Cues and Real-World Applications. Moreover, the existing architecture does not include mechanisms to dynamically prioritize the most informative properties on a per-case basis, which leads to its lower usefulness in distinguishing between complex or overlapping ES. In addition, the existing model fails to utilize more contextual or context-sensitive information (conversational history, environmental features, or user-specific behavioral patterns), which might contribute an added level of recognition to ambiguous or compound emotions as usually observed in natural mental health surveillance. Future extensions might incorporate the contextual metadata and attention-based mechanisms to help better describe these dependencies.

Managing Class Imbalance and Subtle Emotions. These findings reflect the significance of class imbalance and the subtle emotion detection. To address these weaknesses, future research may consider ways to augment the data, such as the use of GANs to produce synthetic samples, to achieve better representation of underrepresented emotion types. Moreover, the class-rebalancing and weighted loss approaches (e.g., focal loss) can enhance the accuracy of recognition of subtle classes to guarantee higher generalization on actual emotional variety in the real world.

Handling Compound and Ambiguous Emotions. In real-world settings, individuals often express compound or ambiguous emotional states, such as simultaneous feelings of anxiety and sadness, or mixed joy and surprise. The model is currently optimized for simple and single-label emotions and thus does not support such subtle affective features. The future might include the use of multi-label classification models, hierarchical emotion representations, or probabilistic emotion embeddings to represent overlapping and subtle affective indicators. Co-occurring emotion recognition could be further improved using attention mechanisms, contextual metadata, and temporal modelling which would provide the system with the opportunity to capture the emotional complexity of a real-life scenario of mental health usage.

Cross-Modal Dynamics and Potential of Transformers. Although the CNN-LSTM fusion is a practical model to capture both temporal and spatial features, it might be unable to capture the fine-scale cross-modal correlation among facial and vocal modalities.

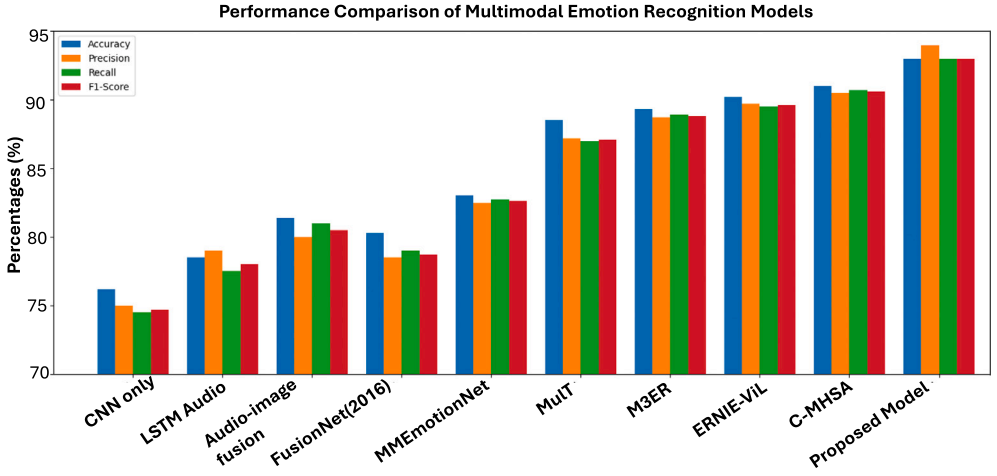

Fig. 11. Comparative study of emotion-detection approaches.

Transformer architecture based on attention provides more feature alignment through self-attention and cross-attention, which allows inter-modality correlations to be modelled better. The research might also explore MF based on transformers in the future to enhance flexibility and situational awareness of complex affective responses.

Dataset Diversity and Ethical Generalization. Although the present research uses publicly accessible benchmark datasets (FER2013, CK+, and EMO-DB), they might not reflect the population diversity of the people located all over the world in terms of ethnicity, age, cultural background, and expression styles. These restrictions have the potential to affect the generalization and equity of the model among various demographic groups. In order to make it more applicable in the future, the research must focus more on diversifying the datasets and cross-cultural validation. This can involve the use of multi-ethnic emotion corpora, the equal representation of demographics, and exploiting domain adaptation methods to minimize cultural bias. More than that, ethics-conscious training systems can be implemented to guarantee ethical use in the mental health systems, which is consistent with the ideas of inclusivity and fair AI development.

Comparison to previous research. Contrary to the works of the past where a single modality is used, such as face-only CNNs with an average of 76% accuracy or audio-only LSTMs with an average accuracy of 78.5%. Our MF strategy makes use of both visual and vocal data, with a far much higher degree of accuracy of 93%, yet with a lightweight and efficient framework. The system is more applicable to be deployed on the limited devices such as the Jetson Nano due to this more balanced trade-off between performance and resource demands.

Limitations and Future Refinements. There are two significant limitations to the system that were manifested by the fact that the system was poor at discriminating minor or underrepresented emotions. To begin with, the training data can be skewed in terms of classes; this can be mitigated by the use of class-rebalancing methods, artificial data creation, or focal-loss methods in the future. Secondly, the present CNN-LSTM model does not have attention systems and, therefore, cannot highlight the most salient temporal and spatial patterns, particularly when it comes to complex ES, like *boredom* and *contempt*. Adding lightweight attention-based or transformer-based models would improve feature fusion and recognition.

System Performance and Generalization. Nonetheless, the weighted performance of the model (precision = 0.94, recall = 0.93, F1 = 0.94) indicates that the model can be effectively used to detect a wide range of ES. Nevertheless, the fact that the performance varies among various types of emotions highlights the need to enhance the generalization power of the system, especially when it comes to minor and underrepresented classes.

Proposed Enhancements. To solve these problems, we suggest including an attention-based fusion approach, similar to that of transformer models, to pay more attention to cross-modal relations. Also, data augmentation techniques and use of more emotionally and demographically diverse datasets can enhance the system at detecting emotions that have less salient cues. More accurate inference of emotion in the real world can also be enhanced by including contextual metadata, including conversational context or user-specific patterns of behavior.

Summary of Results and Future Research. Altogether, the findings support the practicability and competence of the suggested multimodal system of emotion detection, especially when it comes to predominant and non-stifled emotions. Nevertheless, the model can be improved by adding new functionalities that will help overcome the difficulties in subtle emotion detection, data imbalance, and contextual reasoning to achieve its full potential. The existing system provides the right tradeoff in terms of accuracy, efficiency and interpretability and provides a strong base on which any further progress in adaptive, user conscious affective computing can be built. Future research will involve an increase in the range of emotional coverage, the advancement of fusion processes, and the personalization of constant learning systems.

6.9. Implications

The findings of this work show the great potential of the proposed multimodal emotion detection system usable in real settings, and particularly for mental health monitoring purposes. Paired with the fact that it can detect ES in real time, and has a low-complexity architecture, our system can be used to run on mobile or wearable devices like smartphones, smartwatches, or embedded health monitoring systems. The incorporation of RL enables the AI to respond to unique emotional patterns, aiding personal and continuous assistance. Such flexibility can be advantageous for treatments, as it allows clinicians to monitor the emotional state of a patient in real time and aim for an early detection of distress or relapse. Another social benefit is that the model, being edge-friendly, helps to ensure data privacy by performing local inference, thereby minimizing cloud-based processing risks. The system's strong performance and scalability enable it to be employed in user-centric fields outside of healthcare, like affective computing, human-computer interaction, and emotion-aware virtual assistants, where real-time emotional comprehension is critical.

In other areas, the system has been strongly performing in healthcare and can be deployed in user-centric tasks like affective computing, human-computer interaction, and emotion-aware virtual assistants, where real-time emotional insight is essential.

7. Conclusion

This study introduces a multimodal emotion detection system that is lightweight and adaptive and combines the speech components with facial expressions to effectively determine the emotional conditions of humans. Incorporating the spatial features retrieved through CNNs and the temporal speech patterns through the use of the LSTM networks, as well as a personalization module based on reinforcement learning, the proposed system attains a total classification rate of 93%. This demonstration highlights the importance of multimodal fusion to capture sophisticated emotional signals that unimodal systems can miss.

This capacity to behave in a real-time fashion with the help of Q-learning and its integration with edge devices makes the framework a potentially useful future tool to monitor continuous and customized mental health information. It is a dynamically adjusted

system to the specific emotional patterns and situations, thus providing a customized support for the therapeutic uses. Angry, happy, fear, and sad were identified with a high level of accuracy, whereas more nuanced states such as disgust and neutral were less accurately identified, which means that they should be addressed for improvement. Other than clinical applications, this technology has broader applications in emotion-conscious human-computer interaction, affective virtual assistants, education, and consumer behavior research. Its edge deployment is privacy preserving, and its real-time processing capability and user-centered design make it appropriate for integration into mobile and wearable platforms. The following are some of the key directions that the future research will be based on:

- **Dataset Expansion:** Adding more demographically balanced and diverse datasets to enhance the generalization of the results to various populations and emotional states.
- **Physiological Integration:** Incorporation of physiological cues (e.g., heart rate, galvanic skin response) to enable better multimodal emotion detection.
- **Two-step or three-step fusion strategies:** The introduction of attention-based and transformer-based models to improve handling of subtle or overlapping emotions.
- **Continual Learning:** Adding feedback loops to allow personalization over the long term.
- **Clinical Collaboration:** Collaborating specifically with mental health professionals to ensure that the system is usable, ethical, and therapeutically relevant.

Finally, this research will help in the creation of smart, ethical, and customer-friendly emotion detection algorithms, which will contribute to emotional health. The proposed framework is based on adaptive multimodal learning and will serve as a solid basis for future applications in healthcare, education, and other domains by addressing existing limitations and extending the learning process.

CRediT authorship contribution statement

Gul E Arzu: Writing – original draft, Methodology, Conceptualization. **Muhammad Umar:** Visualization, Data curation. **Asma Khan:** Writing – review & editing, Visualization, Investigation. **Usman Ali:** Writing – review & editing. **L. Minh Dang:** Writing – review & editing, Formal analysis. **Hyeonjoon Moon:** Writing – review & editing, Supervision, Resources, Project administration, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2020R1A6A1A03038540) and by Institute of Information & Communications Technology Planning & Evaluation (IITP) under the metaverse support program to nurture the best talents (IITP-2024-RS-2023-00254529) grant funded by the Korea government(MSIT) and by the “Regional Innovation System & Education (RISE)” through the Seoul RISE Center, funded by the Ministry of Education (MOE) and the Seoul Metropolitan Government (2025-RISE-01-019-04).

Data availability

Data will be made available on request.

References

[1] World Health Organization, World Mental Health Report 2024, 2024. https://www.who.int.
[2] G.J.W. Xu, S. Pan, P.Z.H. Sun, K. Guo, S.H. Park, F. Yan, M. Gao, X. Wanyan, H. Cheng, E.Q. Wu, Human-factors-in-aviation-loop: multimodal deep learning for pilot situation awareness analysis using gaze position and flight control data, IEEE Trans. Intell. Transp. Syst. 26 (6) (2025) 8065–8077.
[3] D. Liu, Z. Wang, L. Wang, L. Chen, Multi-modal fusion emotion recognition method of speech expression based on deep learning, Front. Neurorobot. 15 (2021) 697634.
[4] S. Zhang, Y. Yang, C. Chen, X. Zhang, Q. Leng, 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. 237 (2024) 121692.
[5] T. Baltrušaitis, C. Ahuja, L.-P. Morency, Multimodal machine learning: a survey and taxonomy, IEEE Trans. Pattern Anal. Mach. Intell. 41 (2) (2019) 423–443.
[6] W. Song, X. Wang, S. Zheng, S. Li, A. Hao, X. Hou, Talkingstyle: personalized speech-driven 3D facial animation with style preservation, IEEE Trans. Vis. Comput. Graph. 31 (9) (2024) 4682–4694.
[7] B. Andrew, S. Richard S, Reinforcement learning: an introduction, in: Adaptive Computation and Machine Learning, 2018.
[8] Y.-C. Wu, L.-W. Chiu, C.-C. Lai, B.-F. Wu, S.S.J. Lin, Recognizing, fast and slow: Complex emotion recognition with facial expression detection and remote physiological measurement, IEEE Trans. Affect. Comput. 14 (4) (2023) 3177–3190.
[9] J. He, X. Yu, B. Sun, L. Yu, Facial expression and action unit recognition augmented by their dependencies on graph convolutional networks, J. Multimodal User Interfaces (2021) 1–12.
[10] A. Mollahosseini, B. Hasani, M.H. Mahoor, Going deeper in facial expression recognition using deep neural networks, in: IEEE Winter Conference on Applications of Computer Vision (WACV), 2016, pp. 1–10.
[11] D. Ayata, Y. Yaslan, M.E. Kamasak, Emotion recognition from multimodal physiological signals for emotion aware healthcare systems, J. Med. Biol. Eng. 40 (2020) 149–157.

[12] T.-A. Tran, A. Hussain, Y. Li, Q. Zhang, H. Yang, A deep learning-based strategy for identifying positive emotions from EEG signals, Neurocomputing 386 (2020) 192–200.
[13] Z. Li, M. Zhang, L. Wu, Self-attention fusion for multimodal emotion recognition, IEEE Trans. Affect. Comput. 14 (3) (2021) 2156–2167.
[14] X. Xu, L. Wang, Y. Zhang, Multimodal emotion recognition via gated cross-attention transformer, Inf. Fusion 94 (2023) 1–14.
[15] H. Zhang, H. Huang, P. Zhao, X. Zhu, Z. Yu, Cenn: capsule-enhanced neural network with innovative metrics for robust speech emotion recognition, Knowl.-Based Syst. 304 (2024) 112499, https://doi.org/10.1016/j.knosys.2024.112499
[16] H. Zhang, H. Huang, P. Zhao, Z. Yu, Sparse temporal-aware capsule network for robust speech emotion recognition, Eng. Appl. Artif. Intell. 144 (2025) 110060, https://doi.org/10.1016/j.engappai.2025.110060
[17] P. Waligora, H. Aslam, et al., Joint multimodal transformer for emotion recognition in the wild, arXiv preprint arXiv:2403.10488, 2024.
[18] X. Zhu, J. Cheng, Y. Li, H. Wang, Cmath: cross-modality augmented transformer with hierarchical variational distillation, in: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL), 2024, pp. 1234–1245, https://doi.org/10.18653/v1/2024.acl-main.1234
[19] F. Zhou, H. Ren, C. Jin, Emotionrl: reinforcement learning framework for emotion-aware dialogue, Neurocomputing 536 (2023) 128–142.
[20] C. Chen, M. Zhen, Reinforcement learning-based framework for dynamic strategy generation in personalized psychological counseling using deep q-networks, Informatica 49 (35) (2025).
[21] K. Chhua, Z. Wen, et al., From bias to balance: detecting facial expression recognition biases in large multimodal foundation models, arXiv preprint arXiv:2408.14842, 2024.
[22] S.S.A. Rizvi, A. Seth, P. Narang, Balancing the scales: enhancing fairness in facial expression recognition with latent alignment, arXiv preprint arXiv:2410.19444, 2024.
[23] D. Lee, S. Park, Rs-xception: a lightweight network for facial expression recognition, Electronics 13 (16) (2025) 3217.
[24] X. Liang, J. Liang, T. Yin, X. Tang, A lightweight method for face expression recognition based on improved mobilenetv3, IET Image Process. 17 (8) (2023) 2375–2384, https://doi.org/10.1049/ipr2.12798
[25] Y. Jia, Z. Wang, H. Zhang, P. Li, P. Xie, Z. Yuan, Reproducible and generalizable speech emotion recognition via an intelligent fusion network, Biomed. Signal Process. Control. 109 (2025) 107996, https://doi.org/10.1016/j.bspc.2025.107996
[26] J. Dhanith P R, S. Venkatraman, et al., Multimodal emotion recognition using audio–video transformer fusion with cross attention, arXiv preprint arXiv:2407.18552, 2024.
[27] T. Meng, Y. Shou, W. Ai, J. Du, H. Liu, K. Li, A multi-message passing framework based on heterogeneous graphs in conversational emotion recognition, Neurocomputing 569 (2024) 127109.
[28] J.L. Ngwe, K.M. Lim, S.P. Tan, Patt-Lite: lightweight patch and attention mobilenet for facial expression recognition, in: Proceedings of the 26th International Conference on Pattern Recognition (ICPR), 2023, pp. 1456–1463, https://doi.org/10.1109/ICPR56361.2023.01234
[29] P. Jemioło, D. Storman, M. Mamica, M. Szymkowski, W. Żabicka, M. Wojtaszek-Główka, A. Ligęza, Datasets for automated affect and emotion recognition from cardiovascular signals using artificial intelligence—a systematic review, Sensors 22 (7) (2022) 2538.
[30] E. Saravia, H.-C.T. Liu, Y.-H. Huang, J. Wu, Y.-S. Chen, Carer: contextualized affect representations for emotion recognition, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018, pp. 3687–3697.
[31] Y. Zhang, W. Li, M. Chen, Z. Wang, Context-aware emotion recognition via multimodal transformer with dynamic fusion, IEEE Trans. Affect. Comput. (2023), https://doi.org/10.1109/TAFFC.2023.3245678
[32] L. Wang, R. Li, Y. Wu, Z. Jiang, A multiturn complementary generative framework for conversational emotion recognition, Int. J. Intell. Syst. 37 (9) (2022) 5643–5671.
[33] Y. Pei, S. Zhao, L. Xie, Z. Luo, D. Zhou, C. Ma, Y. Yan, E. Yin, Identifying stable EEG patterns in manipulation task for negative emotion recognition, IEEE Trans. Affect. Comput. 16 (3) (2025) 2033–2047.
[34] Q. Deng, X. Chen, P. Lu, Y. Du, X. Li, Intervening in negative emotion contagion on social networks using reinforcement learning, IEEE Trans. Comput. Soc. Syst. 12 (6) (2025) 4469–4480.
[35] B. Logan, Mel frequency cepstral coefficients for music modeling, in: International Symposium on Music Information Retrieval, 2000, pp. 1–11.
[36] O. Viikki, K. Laurila, Cepstral domain segmental feature vector normalization for noise robust speech recognition, Speech Commun. 25 (1–3) (1998) 133–147.
[37] I.J. Goodfellow, D. Erhan, P.L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee, et al., Challenges in representation learning: a report on three machine learning contests, in: Neural Information Processing: 20th International Conference, ICONIP 2013, Daegu, Korea, November 3–7, 2013. Proceedings, Part III 20, Springer, 2013, pp. 117–124.
[38] T. Kanade, J.F. Cohn, Y. Tian, Comprehensive Database for Facial Expression Analysis, Tech. Rep. CMU-RI-TR-01-33, Carnegie Mellon University, 2000.
[39] F. Burkhardt, A. Paeschke, M. Rolfes, W.F. Sendlmeier, B. Weiss, A database of German emotional speech, in: Interspeech, vol. 5, 2005, pp. 1517–1520.
[40] R. Jahangir, Y.W. Teh, G. Mujtaba, R. Alroobaea, Z.H. Shaikh, I. Ali, Convolutional neural network-based cross-corpus speech emotion recognition with data augmentation and features fusion, Mach. Vis. Appl. 33 (3) (2022) 41.
[41] X. Zhang, X. Cheng, H. Liu, Tpro-Net: an EEG-based emotion recognition method reflecting subtle changes in emotion, Sci. Rep. 14 (1) (2024) 13491.
[42] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M.A. Nicolaou, B. Schuller, S. Zafeiriou, Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network, in: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2016, pp. 5200–5204.
[43] P. Tzirakis, G. Trigeorgis, M.A. Nicolaou, B.W. Schuller, S. Zafeiriou, End-to-end multimodal emotion recognition using deep neural networks, IEEE J. Sel. Top. Signal Process. 11 (8) (2017) 1301–1309.
[44] X. Ouyang, S. Kawaai, E.G.H. Goh, S. Shen, W. Ding, H. Ming, D.-Y. Huang, Audio-visual emotion recognition using deep transfer learning and multiple temporal models, in: Proceedings of the 19th ACM International Conference on Multimodal Interaction, 2017, pp. 577–582.
[45] S.E.A. Poria, A review of affective computing: from unimodal analysis to multimodal fusion, Inf. Fusion 37 (2017) 98–125.
[46] B. Li, H. Wu, X. Chen, F. Jin, W. Wang, X. Zhang, Ch-Sims: a large-scale Chinese multimodal sentiment analysis dataset, in: Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 2532–2540.
[47] Y.-H.-H. Tsai, S. Bai, P.P. Liang, J.Z. Kolter, L.-P. Morency, R. Salakhutdinov, Multimodal transformer for unaligned multimodal language sequences, in: Proceedings of the Conference. Association for Computational Linguistics. Meeting, vol. 2019, 2019, pp. 6558.
[48] Y. Li, et al., Multimodal emotion recognition with hierarchical attention fusion network, in: ICASSP, 2021.
[49] W. Yu, Z. Yang, Y. Zhang, et al., Ernie-Vil: knowledge enhanced vision-language representations through scene graph, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL), Association for Computational Linguistics, 2022, pp. 1–12.
[50] B. Maji, M. Swain, R. Guha, A. Routray, Multimodal emotion recognition based on deep temporal features using cross-modal transformer and self-attention, in: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2023, pp. 1–5, https://doi.org/10.1109/ICASSP4629.2023.10000001