



Research Article

Emotion-aware human-computer interaction using multimodal deep learning and explainable artificial intelligence

Pallavi WANKHEDE¹, Deepali KOTAMBKAR², Kanchan WARKAR^{3,*}, Sweta RAUT⁴

¹St. Vincent Pallotti College of Engineering and Technology, Nagpur, Maharashtra, 441108, India

²Ramdeobaba University, Nagpur, Maharashtra, 440013, India

³Visvesvaraya National Institute of Technology, Nagpur, Maharashtra, 440010, India

⁴Jhulelal Institute of Technology, Off Koradi Road, Lonara, Nagpur, Maharashtra, 441111, India

ARTICLE INFO

Article history

Received: 19 November 2025

Revised: 12 January 2026

Accepted: 22 March 2026

Keywords:

Affective Computing;
Deep Learning; Emotion
Recognition; Explainable AI;
Human-Computer Interaction;
Multimodal Fusion

ABSTRACT

Emotion-aware human computer interaction is the natural and flexible Human Computer Interaction (HCI) to perceive and consider human emotion responses in the computation. Few recent successes in deep learning have resulted in an increased emotion recognition accuracy but, mostly the emotion-aware HCI systems are limited to poor robustness in practical scenarios and lack of interpretability at the decision level, thus limiting their adoption for practical applications. This work defines the problem of robust and transparent emotion recognition by designing an explainable modality-adaptive multimodal deep learning framework, capable of fusing multiple input modalities such as facial expressions, speech and physiological responses. The emotion perception framework employs a set of deep neural models to perform spatial-temporal feature extraction and introduces a set of neural architecture designs, namely, Modality-Adaptive Attention Fusion Module (MAAFM) mechanism to adaptively fuse the modalities based on modality qualities and contextual information. Further, for interpretability, the knowledge at the modality level and at the fusion level is extracted through various shallow modules at the feature- and the fusion-level respectively, which provides transparency towards how each modality affects the inference. The proposed framework is evaluated over benchmark datasets for visual, audio and physiological modalities and reports an overall emotion recognition accuracy of 92.4 %, F1-score of 0.91, and area under the receiver operating characteristic curve of 0.97, while preserving a low inference latency for interactive applications. The approach provides improved robustness, transparency and applicability. The key innovation of this work is the focus on the attention not only as a performance tool but also as an explicit source of interpretation of the modality contributions at the fusion level that advances existing emotion recognition approaches.

Cite this article as: Wankhede P, Kotambkar D, Warkar K, Raut S. Emotion-aware human-computer interaction using multimodal deep learning and explainable artificial intelligence. Sigma J Eng Nat Sci 2026;44(4):2758–2782.

*Corresponding author.

*E-mail address: kanchan22.warkar@gmail.com

This paper was recommended for publication in revised form by
Editor-in-Chief Ahmet Selim Dalkilic



INTRODUCTION

Background and Motivation

Emotion recognition as part of next-generation human-computer interaction systems capability will allow computers to not only perceive, and recognize emotion of a human, but also react to the emotion in a way humans would in a social situation. Interactions of previous generations of human-computer interaction system were based on the explicit interaction modalities such as typed command, or speech input, thus did not have the means to detect the implicit component of the human interaction with a system. On the other hand, emotion-enabled computer human interaction systems seek to infer emotion from expressive cues provided by the human via speech, facial expressions as well as body language [1].

Different inputs such as face, speech, and physiological signal are encoded in dedicated modality encoding module, and fused in the proposed modality-adaptive attention fusion module (MAAFM). Both feature-level explanation and fusion-level explanation are conducted in the Multi-Level Explainability Module before final adaptive system response [2].

The significance of emotion-aware interaction is thus very high, far beyond technical affective computing research. For example, in adaptive learning systems, automatic emotion detection can be used to automatically recognize interest, confusion, frustration, or engagement, to adapt the system to what is sensed in real-time [3]. In health monitoring and medical applications, monitoring the patterns of emotional behaviors can assist in early detection of stress-related problems through unobtrusive sensing, and assistive interfaces for the elderly or persons with disabilities can adapt to their affective states for improved comfort and usability.

Unimodal approaches to emotion recognition were developed initially and determine emotional affect states from a singular source - facial expressions, speech,

physiological signals. Whilst on occasion they produce excellent results in isolated controlled environments, unimodal methods lack robustness in real-life situations. Facial features can be occluded or neutral, speech recognition systems are target dependent and susceptible to noise and cultural influences, physiological sensing is intrusive and lacks context [4, 5]. This lack of robustness compromises generalisation.

In the last several years, deep learning has contributed to advances in emotion recognition through learning of features from extensive sensory data, but the model of human-like understanding calls for a synthesis of visual, auditory and physiological cues. Multimodal deep learning has been proposed to provide this multisensory integration for better accuracy, contextual sensitivity, and robustness to noise. However, growing levels of complexity in deep models of behaviour pose significant challenges of robustness, explainability and real-time suitability for human-computer interaction in different situations with different users, environments, and emotional strengths [6, 7].

Figure 1 shows an example of a user whose emotion is detected by sensing the different axes involved in an interaction in a computer/robot (facial expression, voice, physiological sensors). Each sensor modality is modeled independently using deep networks (Convolutional Neural Network (CNN) for vision, Bidirectional Long Short-Term Memory (BiLSTM) for audio, and Long Short-Term Memory (LSTM) for physiological signals) and then combined into one fusion layer [8]. The fused features are then passed through an explainability analysis (SHAP or Grad-CAM) before the responsive information (emotion-based feedback, interface change, speech output) is provided.

Research Gap

Despite impressive performance advances, numerous essential issues still need to be addressed before multimodal emotion recognition systems can be practical. Many top-performing models, including those based on

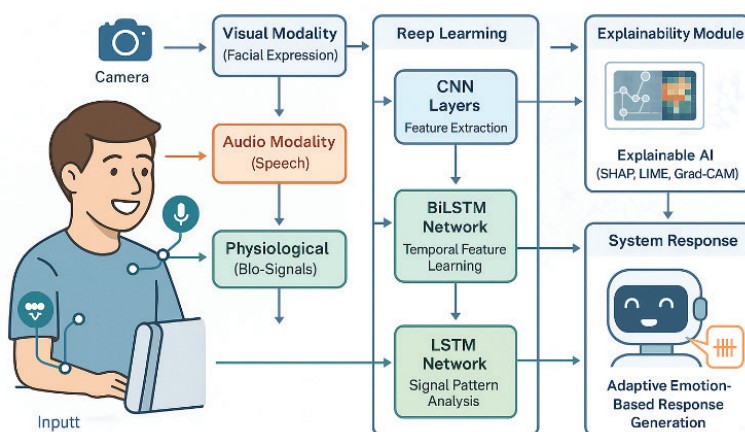


Figure 1. Conceptual overview of the modality-adaptive multimodal emotion recognition framework.

deep learning such as convolutional, recurrent, and transformer-based models, focus on predictive accuracy and are largely unexplainable. Their lack of explainability, however, hampers user trust and provider accountability and raises ethical issues, especially when used in domains like education, health care, and mental health [9].

A third drawback of these approaches is the absence of dynamic fusion schemes that can adapt to the different levels of input noise as well as occlusion. Many recognition systems fuse their input features in a predetermined way, either by concatenating feature vectors, or applying fixed rules, which doesn't reflect the actual interaction environment where some modalities may be rendered unreliable. Recognition accuracy drops dramatically when one modality is severely affected [10].

Additionally, application-oriented requirements, including real-time operation, robustness in the presence of sensing failures, user-oriented explainability are often neglected. Although the human-interpretable explanations are emerging in recent work, existing approaches mainly adopt attention-based mechanisms, and explainable AI techniques as separate, or after-fact tools for the multimodal fusion, thus not able to addressing the simultaneous needs of accuracy, robustness, interpretability, and real-time in the context of interaction between humans as well as computers.

Research Objectives

The objective of this study is to develop a multimodal deep learning framework with integrated explainable artificial intelligence for emotion-aware human-computer interaction that addresses the identified limitations. The specific objectives are as follows:

- To design a multimodal emotion recognition framework that integrates facial, speech, and physiological signals, enabling comprehensive affect understanding through complementary sensory information.
- To improve recognition robustness and reliability under varying environmental conditions, including background noise, visual occlusion, and physiological signal distortion, by dynamically adapting modality contributions during inference.
- To incorporate explainable artificial intelligence techniques that provide feature-level, modality-level, and decision-level interpretability, thereby enhancing transparency, trust, and human understanding of system behavior.
- To validate the proposed framework using benchmark datasets and real-time interaction scenarios, demonstrating its suitability for practical emotion-aware human-computer interaction applications.

Through these objectives, the study aims to contribute a unified, transparent, and robust emotion recognition framework that supports the development of ethical, trustworthy, and empathetic interactive systems.

Novelty and Contributions

The main technical contributions of this work are summarized as follows:

1. Modality-adaptive Modality-Adaptive Attention Fusion Module (MAAFM) architecture:

A reliability-aware multimodal fusion design in which attention weights are learned per input sample, enabling dynamic adjustment of modality importance. This design yields measurable improvements in accuracy, area under the curve, Cohen's Kappa, and robustness under missing or noisy modalities, as validated through controlled degradation experiments.

2. Unified multi-level explainability integration strategy:

A systematic integration of explainable artificial intelligence at feature, fusion, and decision levels, combining Grad-CAM, SHAP, LIME, and attention-weight visualization. This design enables transparent interpretation of both modality-specific features and cross-modal interactions, which is quantitatively and qualitatively demonstrated through visual explanations and attribution analysis.

3. Rigorous and reproducible experimental evaluation protocol:

A comprehensive evaluation framework across three benchmark datasets, including subject-independent testing, robustness analysis under degraded sensing conditions, ablation studies, and statistical significance testing with confidence intervals, ensuring reproducibility and objective performance validation.

4. Computationally efficient multimodal design for real-time interaction:

A CNN-BiLSTM-Attention architecture that achieves competitive or superior performance compared to transformer-based multimodal models while maintaining lower model complexity, reduced inference latency, and real-time suitability for emotion-aware human-computer interaction systems.

Although convolutional-recurrent architectures with attention mechanisms have been widely explored for multimodal emotion recognition, most existing frameworks employ attention primarily as a latent optimization tool and apply explainable artificial intelligence in a post hoc manner. In contrast, the proposed framework introduces a modality-adaptive attention mechanism that explicitly models modality reliability on a per-sample basis and exposes attention weights as interpretable fusion-level explanations. Unlike prior multimodal explainable systems that focus on modality-specific interpretation only, this work integrates explainability before fusion, during fusion, and after decision making. Furthermore, the proposed design emphasizes real-time feasibility and robustness under degraded sensing conditions, which are rarely addressed jointly in existing CNN-RNN-attention architectures.

LITERATURE REVIEW

Unimodal Emotion Recognition

Unimodal Emotions Classification There are primarily two types of early work in emotion recognition using unimodal analysis, which identify emotional status based on only one source of information among facial expressions, speech signals or physiological signals. Classic feature set and classifier models based on hand extracted features such as support vector machine, hidden Markov model and decision tree were extensively investigated in the pioneer work of unimodal emotions classification [11, 12], with interpretable model and relatively low cost, but the overall performance was very vulnerable to noise, incomplete and unpredictable environments.

Due to the ability of deep learning to automatically extract spatial features from images, convolutional neural networks (CNN) have become popular, and adopted for facial emotion recognition. Whereas, the use of recurrent neural networks (RNN) and long short-term memory (LSTM) models for speech and physiological signals, the models have been used to learn temporal features [13]. Due to deep models' impressive recognition performance in controlled environments, unimodal emotion recognition systems have still failed to be robust to modality-dependent variations such as facial occlusion, speech with background noise and sensor artifacts for physiological signals. As a result, the unimodal methods do not generalize well to different users, environments and interactions [14].

Multimodal Emotion Recognition and Fusion Approaches

Efforts to address the insufficiency of a unimodal system proposed multimodal emotion recognition frameworks that fuse audio, video, and physiological information. The goal of fusing multiple sources is inspired by human sensation in which a human evaluates emotions through the combination of multiple senses. Current multimodal systems can be categorized into three types: early fusion, late fusion, and hybrid fusion [15, 16].

Early fusion: Early fusion approaches fuse features across multiple modalities prior to the classification stage and allow for learning of cross-modalities interactions in the underlying model. Early fusion methods however, are vulnerable to noisy or failed modalities since all modalities are considered equally. **Late fusion:** Late fusion approaches combine decisions of independent modalities at the decision level, thus being robust to data drop-out, but usually ignoring fine-grain interactions across modalities [17, 18]. **Hybrid fusion:** Hybrid fusion approaches (e.g., attention-based model, transformer-based model) try to find a trade-off between early and late fusion methods [17, 18].

These days, a few research works have shown that the use of attentional modules and self-attentional model architectures is desirable for capturing the intermodal dependencies for better multimodal emotion recognition [19], and transformer-based fusion architectures have achieved competitive

results on benchmark datasets. However, majority of them optimize the model for performance without any focus on model interpretability and real-time constraints. In addition, the fusion process is fixed or obscure, which makes the full advantage of each modality little explainable to be used for human-computer interaction system [20].

Explainable Artificial Intelligence in Emotion Recognition Systems

Deep learning models of emotion recognition through human behavior, speech, and physiological signals are more effective if they provide interpretability since this is more relevant for many applications like human-computer interaction, education, healthcare, and mental health. Techniques like gradient-based visual explanations, feature attribution methods, and local surrogate models, have been used to interpret deep learning predictions. For facial emotion recognition, visual explanation techniques are used to explain facial areas relevant to emotions. For speech and physiological emotion recognition, feature attribution methods are used to explain influential features like dominant frequency bands or acoustic features. [21-23]

While these modality-specific techniques increase interpretability on a modality level, most of the current work is limited to applying either explainable machine learning methods or interpretability frameworks to unimodal models. Integrating explainability to multimodal pipelines remains an unexplored field. Several approaches explain unimodal model's predictions based on the specific modality, leaving the question of how the fused modality representation and fusion mechanism of the multimodal model contribute to the prediction open.

Identified Research Gap and Motivation for the Proposed Work

According to the discussed papers, several important issues still need to be tackled. Firstly, there is no multimodal emotion recognition framework that simultaneously combines robustness, interpretability and real-time performance. Although using multiple sources of data helps improve the recognition performance, some researchers resort to static or opaque fusion methods that do not work when modalities are degraded and are difficult to interpret. Secondly, very few of the explainable AI techniques are applied at the fusion level which prevents user to know the overall cross-modal contributions to emotion inference. Lastly, many of the state-of-the-art models are computationally expensive and cannot be used in real-time human computer interaction [24, 25].

The planned work to overcome these issues is explained below: A modality-adaptive and explainable multimodal framework to fuse and utilize multiple cues, including facial, speech, and physiological signals, is presented. It differs from previous work by exploiting attention as a fusion cue and as a transparent signal, where the involved modalities are weighted according to the given context and the

signals that result the model prediction are made transparent. A full multimodal deep learning framework with integrated explainability and low-latency inference is proposed to match the technical and practical needs of real-world emotion-aware human-computer interaction.

MATERIALS AND METHODS

System Overview

The Modality-Adaptive Multimodal Emotion Recognition Framework, which includes three modality encoders, a Modality-Adaptive Attention Fusion Module (MAAFM) and a Multi-Level Explainability Module. This mechanism enable the proposed framework to adapt to the dynamic environment in real-life scenarios, for example, noisy speech or occluded face parts. With both class-level and fusion-level explainable AI components, the intermediate outputs in the network and the final results could be interpreted and explained. This provides a balanced solution among the accuracy, robustness and interpretability, which are the critical requirements for real-world emotion-aware interactive applications.

Facial features are encoded using ResNet50; speech features are captured using 1D CNN-BiLSTM; physiological features are represented by BiLSTM. These embeddings are then fused by the Modality-Adaptive Attention Fusion Module (MAAFM) before classification and explanation analysis.

During the first level of feature extraction, convolutional neural networks are used to generate the spatial embeddings from facial frames, meanwhile the temporal models, for instance, BiLSTM or temporal convolutional networks, for capturing the sequential dependencies of speech and

physiological signals. In the fusion level, CNN-BiLSTM-Attention Architecture is used to fuse different embeddings into a latent one.

Figure 2 provides an overview of the processing architecture for the emotion aware human computer interaction framework. The architecture is modular by design, where facial, speech and physiological inputs are first extracted separately in order to maintain their modality specific information.

In the end, emotion class labels are generated by a softmax classifier and XAI methods, like Grad-CAM, SHAP and LIME, provide visual and quantitative interpretability across features and decisions.

Multimodal Data Representation

The performance of the emotion recognition is largely influenced by the encoding of features extracted from different modalities. The input tensors for the facial, speech, and physiological modalities are represented by $X_{face} \in \mathbb{R}^{T \times d_f}$, $X_{audio} \in \mathbb{R}^{T \times d_a}$, and $X_{physio} \in \mathbb{R}^{T \times d_p}$ respectively, where T denotes the temporal length, and d the feature dimensionality.

Facial expressions

Facial cues are detected by the CNN backbone (Resnet50) that can capture local and global spatial dependencies. The CNN is first passed through a sequence of convolutional filters with nonlinear activations [26]:

$$F_{face} = \sigma(W_{c1} * X_{face} + b_{c1}) \quad (1)$$

where; $*$ indicates the convolution, W_{c1} stands for the filter weights, b_{c1} is the bias term, and $\sigma(\cdot)$ specifies the ReLU activation. The resulting embeddings capture spatial facial features relevant to the emotion classification.

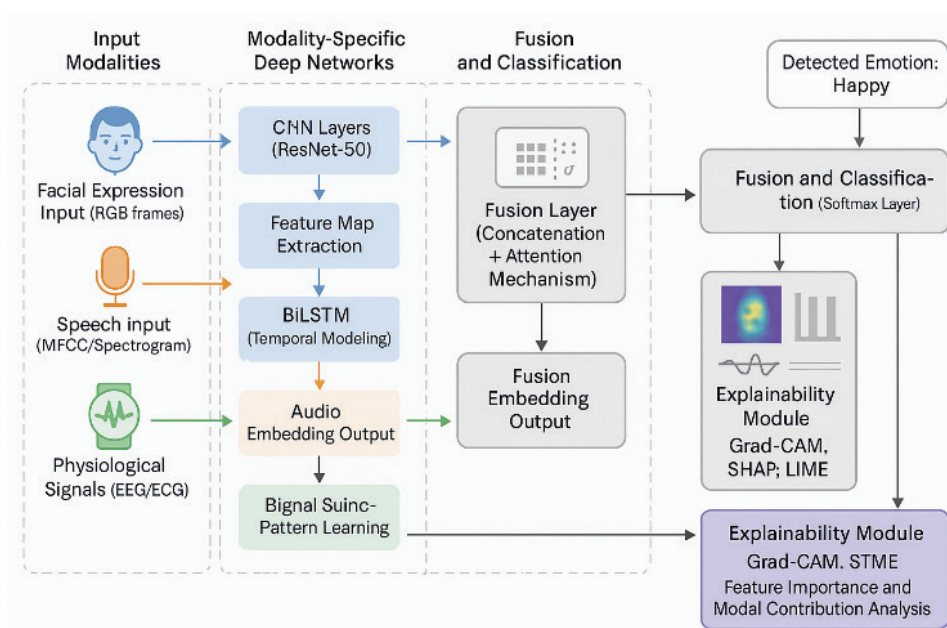


Figure 2. Detailed architecture of the CNN-BiLSTM-Attention Architecture within the proposed framework.

Speech features

Emotion in the speech is depicted via Mel-Frequency Cepstral Coefficients (MFCCs), or log-mel spectrograms. The temporal-spectral structure is handled first with a one-dimensional CNN, and then with a BiLSTM [27]:

$$F_{audio} = BiLSTM(\sigma(W_{c2} * X_{audio} + b_{c2})) \quad (2)$$

By considering both the short-term spectral dynamics, and long-term tonal variations, this model, for example, pitch as well as energy fluctuation, is used for facilitating reliable emotion perception from the speech.

Physiological signals

Physiological signals like ECG, or EEG exhibit very fine changes that are linked to the emotional arousal, and valence levels. These signals are processed using a BiLSTM, or TCN, which are shown in [28]:

$$h_t = BiLSTM(x_t, h_{t-1}) = f(W_{ih}x_t + W_{hh}h_{t-1} + b_h) \quad (3)$$

$$F_{physio} = [h_1, h_2, \dots, h_T] \quad (4)$$

The bidirectional design allows for the recognition of the temporal dependencies in both the future, and past directions, thus, enhancing the stability, and the understanding of the emotional context.

Subsequently, the embeddings of each modality are subjected to normalization as well as temporal alignment prior to the fusion [29]:

$$\tilde{F}_m = \frac{F_m - \mu_m}{\sigma_m}, \quad m \in \{face, audio, physio\} \quad (5)$$

Feature Fusion Model

Fusion is the decisive step which combines the multimodal embeddings into a common feature space that is able to learn cross-modal correlations. The Modality-Adaptive Attention Fusion Module (MAAFM) is formally expressed as [30]:

$$z = f(W_f[x_{face} \oplus x_{audio} \oplus x_{physio}] + b_f) \quad (6)$$

where, $\oplus$ indicates the combination of vectors, W_f and b_f are the weights, and also biases for the fusion, and $f(\cdot)$ is a nonlinear activation function such as ReLU. Within the MAAFM, inter-modality relevance is modeled using a learnable attention mechanism [31]:

$$\alpha_i = \frac{\exp(e_i)}{\sum_j \exp(e_j)}, \quad e_i = v^T \tanh(W_h h_i + b_h) \quad (7)$$

where; the attention coefficient of modality i is denoted by α_i , and h_i is the hidden representation of modality i . The fused embedding is subsequently calculated as [32]:

$$Z = \sum_i \alpha_i h_i \quad (8)$$

The proposed hybrid CNNs-BiLSTM-Attention model is built to understand the joint dependence of these spatial, spectral and temporal dimensions. CNNs extract the local feature patterns, BiLSTM layers encode the sequence of emotional dynamics and the attention mechanism computes the weights for the different modality.

Probability distribution of the final emotion class is obtained by applying softmax layer [33]:

$$\hat{y} = Softmax(W_c Z + b_c) \quad (9)$$

Conversely; W_c and b_c are the classifier parameters, and $\hat{y} \in R^K$ corresponds to the probability distribution over the K emotion classes.

Explainability Integration

Explainable AI (XAI) is integrated into the Feature importance analysis, and multimodal fusion interpretability to enhance the trust as well as transparency.

Grad-CAM for visual saliency

The gradient of the heatmap is produced by the gradient-based activation map that highlights the key facial regions associated with the prediction of emotions. The heatmap from *Grad - CAM*_{Grad-CAM} for the convolutional layer A^k is generated using the following equation [34]:

$$L_{Grad-CAM} = ReLU\left(\sum_k \alpha_k A^k\right), \quad \alpha_k = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \quad (10)$$

In this expression, y^c denotes the score assigned to class c that corresponds to the given emotion, while α_k represents the weights obtained through the gradient-based methods. The depiction made in this case marks the areas of the face, like for instance the eyes, and mouth, which have the greatest impact on the final decision regarding the classification.

SHAP and LIME for cross-modal attribution

SHAP as well as LIME analyze feature-level importance to interpret the role of the modalities in the final decision-making process. Feature contributions by SHAP are given by [35]:

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|! (|F| - |S| - 1)!}{|F|!} [f(S \cup \{i\}) - f(S)] \quad (11)$$

where; ϕ_i idenotes refers to the SHAP value of the i th feature. It measures the relative contribution of each modality (facial, audio, or physiological) to the emotion classification of each particular instance.

The generated feature heatmaps, and bar plots are interpretable visual feedbacks that allow system designers to confirm the emotional representations learned, and find biases that might be present.

Training Strategy

The multimodal model is taught thoroughly without any interruptions. The categorical cross-entropy loss

is defined as [36] if the ground truth emotion labels $y_i \in \{1, 2, \dots, K\}$, and predicted probabilities $\hat{y}_i$ are given:

$$L = -\frac{1}{N} \sum_{i=1}^N y_i \log(\hat{y}_i) \quad (12)$$

where; N denotes the number of samples in a batch. The optimization process is executed through the Adam optimizer utilizing the adaptive moment estimation technique, which is mathematically represented as [37]:

$$\theta_{t+1} = \theta_t - \eta \frac{m_t}{\sqrt{\hat{v}_t + \epsilon}} \quad (13)$$

Here, η is the learning rate, and $\hat{m}_t$ and $\hat{v}_t$ are the first; and second moments which are corrected for the bias.

The learning rate η is dynamically changed according to a cosine annealing schedule in order to maintain the stability of convergence. The training uses dropout regularization (0.3–0.5) to avoid the overfitting, while mini-batch sizes of 32–64 are used to provide the right balance between the computational efficiency. Validation accuracy is used for early stopping which guarantees the best generalization.

Novelty of the Proposed Fusion and Explainability Design

This sub-section states clearly how our fusion and explainability design spaces is different from the current multimodal CNN-RNN-attention + explainable artificial intelligence spaces. Our design space does not treat attention and explainability as additional modules. We design attention as an adaptive fusion component and an intermediate layer that captures cross-modal contribution at the decision level.

Rather than tuning the modality attention weights over the entire training set or keeping them static, they are learned for each input individually. Suppose that $\mathbf{h}_v, \mathbf{h}_a, \mathbf{h}_p$

denote the feature embedding representations fed into the visual, speech, and physiological encoders. The scores for each modality for a specific instance of the input are:

$$e_m = \mathbf{w}^T \tanh(\mathbf{W}_m \mathbf{h}_m + \mathbf{b}_m), m \in \{v, a, p\}$$

where; $\mathbf{W}_m, \mathbf{w}$, and $\mathbf{b}_m$ are learnable parameters. These relevance scores are normalized using a softmax function to obtain the modality attention weights [38]:

$$\alpha_m = \frac{\exp(e_m)}{\sum_{k \in \{v, a, p\}} \exp(e_k)}$$

The resulting attention weights satisfy $\sum_m \alpha_m = 1$, and represent the relative importance of each modality for the current input sample. The fused multimodal representation is then computed as [39]:

$$\mathbf{h}_{fusion} = \sum_{m \in \{v, a, p\}} \alpha_m \mathbf{h}_m$$

This allows the fusion to be robust against change in signal quality, temperature, lighting, and expressiveness of emotion. For example, when a modality is poor such as speech in noise, face under occlusion, speech presence in the hair, etc., its contribution will be put down, and other more robust modalities such as face, body language, and gesture, can dominate in the formation of the overall expression. This is the key difference from early fusion which concatenates features without any regard to the dependence between different modalities or late fusion which simply combining individual decisions without modeling the dependence.

Figure 3 displays a schematic illustration of the modality-adaptive fusion and explainability flow. It explains how

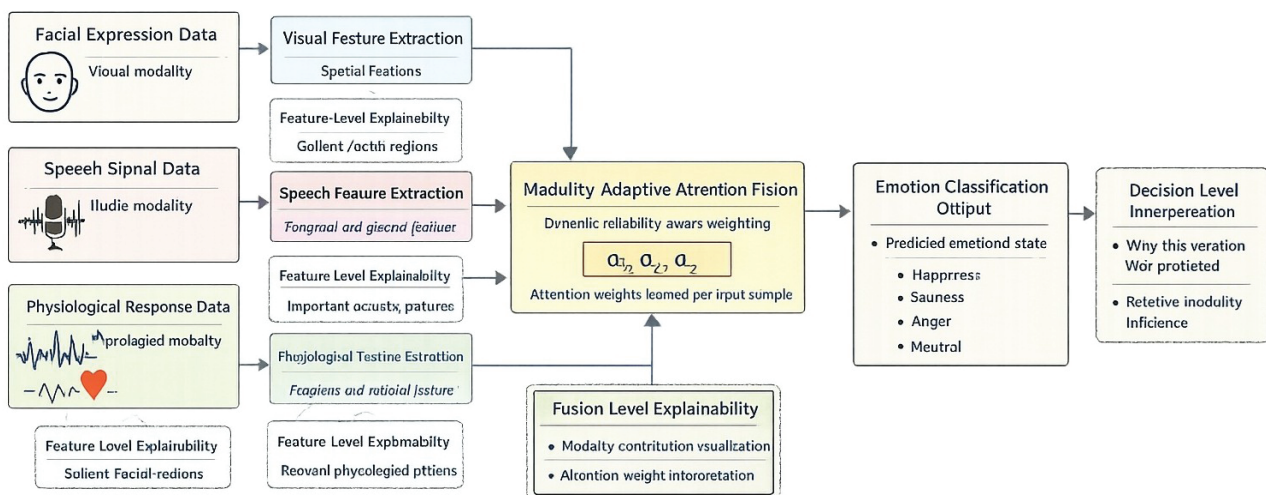


Figure 3. Schematic representation of the Modality-Adaptive Attention Fusion Module (MAAFM), and Multi-Level Explainability flow. This illustrates modality-specific encoding, dynamic attention-based weighting, fused representation formation, and the integration of the Grad-CAM, SHAP, LIME, and attention-weight visualization at the multiple interpretability stages.

the modality-specific features are dynamically weighted and how the explainable artificial intelligence algorithms offer explanation before and after fusion.

The main novelty of the proposed design is the explicit targeting of the attention weights as explainability signals. Attentions are no longer implicit in an optimization process, but the learned modality weights are revealed as perceptible and user-trustworthy metrics of the modalities' contribution to the emotion prediction. This allows explicit quantification of each modality's contribution to the output.

The Multi-Level Explainability Module is included at various levels of the proposed Modality-Adaptive Multimodal Emotion Recognition Framework. Prior to the fusion of modalities, modality-specific techniques are used to explain important spatial areas of facial input, salient temporal and spectral patterns of speech, as well as discriminative frequency bands and electrode regions of physiological signals. After the fusion step, attention-weight distributions and feature attribution techniques are adopted to jointly explain how the multimodal representations lead to the final classification result. This multi-level explainability design makes the proposed framework different from the existing past work, which mostly only focus on explanation of a single modality or single model output.

In contrast to unimodal transformer-based fusions, that typically feature high computational cost and less modality-level interpretability, the more flexible CNN-BiLSTM-Attention framework delivers a desirable trade-off between the effectiveness, interpretability, and efficiency. Also, due to the transparent fusion module and low-latency inference, the framework is applicable for real-time emotion-aware human-computer interaction context that requires robustness, interpretability, and responsiveness.

Experimental Setup

Datasets used

The multi-modality oriented emotion-sensitive framework for HCI is experimentally evaluated using the three commonly used and most highly regarded datasets: DEAP, RAVDESS and SEED. The DEAP and RAVDESS datasets provide audio-visual video and audio streams, respectively, along with the annotated emotion labels, while the SEED dataset provides EEG-based emotion data. This enables a complete multi-modality validation process.

The DEAP data set also provides EEG and peripheral physiological signals for 32 subjects who watched

40 different 1-minute music videos for audiovisual. Each audio-visual stimulus was rated using a 9-point scale along three axes; i.e., arousal, valence and dominance. The entire data set consists of 32 channels of EEG data, and 8 channels of physiological data including GSR, EOG, and EMG signals. The EEG signals are recorded at a frequency of 512 Hz then down-sampled to 128 Hz to ease the computational load.

The RAVDESS (Ryerson Audio-Visual Database of Emotional Speech, and Song) dataset consists of 24 professional actors (12 males, 12 females) vocalizing 8 emotion classes (neutral, calm, happy, sad, angry, fearful, disgust, surprised) through speech and song. The output is a sample which is composed of synchronized audio, and video at 48KHz sampling rate, and 30 fps video frame rate. This RAVDESS dataset can be utilized to conduct the unimodal, and multimodal experiments under a standard environment like lighting, and background, which ensures the extraction of high-quality features.

SEED (SJTU Emotion EEG Dataset) The data was obtained from 15 subjects when watching emotional movie videos positively, negatively and neutrally scored. The EEG signals were captured by 62 electrodes attached on their heads following the international 10-20 scheme. The acquisition frequency of EEG was 1000 Hz, and the received signals were subsequently band-passed-filtered within the range of 0.3-50 Hz to reduce the noise. This dataset is suitable for investigating the temporal dynamics of the brain's emotional response. The datasets' characteristics are detailed in Table 1.

Cross-dataset modality alignment and experimental protocol

These datasets contain different participants, various acquisition paradigms, different sensing modalities, different sampling frequencies, different participant types, and variances in annotation strategies. In order to provide a fair, reproducible and rigorous assessment, the cross-dataset modality translation is conducted at the feature space level instead of the raw signal level.

For every dataset, we perform modality-specific pre-processing and feature extraction. The faces, speech, and physiological signals of each modalities are individually encoded in modality-specific sub-modules that maintain the means for preservation of the native time and frequency and spatial mode information of every input. No

Table 1. Dataset summary

Dataset	Modality	Participants	Channels/Features	Sampling Rate	Emotion Classes
DEAP	EEG + Physiological	32	32 EEG + 8 physio	512 Hz (→ 128 Hz)	Valence, Arousal, Dominance
RAVDESS	Audio-Visual	24	Speech, Video Frames	48 kHz, 30 fps	8 Emotions
SEED	EEG	15	62 EEG Channels	1000 Hz (→ 50 Hz)	Positive, Neutral, Negative

signal level resampling, normalization, or direct sharing of features across the datasets is performed.

We map the three datasets to a common set of categorical emotion labels, so that we can evaluate between datasets in a unified affective space. We maintain the original label resolutions when creating this joint set, conforming to standard affective computing procedures and without disturbing the original distributions. We treat each dataset independently otherwise.

Subject-independent evaluation protocol The data are split at the subject level, so that data from a particular subject does not appear in both training and test sets. The training and validation data comprise roughly 80% of the dataset while the remaining 20% is used for testing. The validation data is used solely from the training data for hyperparameter tuning.

Cross-dataset generalization experiments. Additional experiments are carried out to evaluate robustness under the domain shift. In this setup, the model is trained on one dataset and directly tested on the other dataset. The purpose of this protocol is to test the generalization ability of the learned multimodal representations to different distributions, sensing conditions and populations.

The models were optimized with same optimization setting with each dataset (learning rate, batch size, early stopping criterion). The evaluation metrics are only calculated on the test set, and the reported performance is averaged over several runs.

Preprocessing steps

The multimodal preprocessing pipeline has been configured to extract synchronized noise free features from all the facial, audio and physiological channels.

The frequency of the video frames of RAVDESS is 30, and in this case, they undergo detection and alignment operations with Multi-Task Cascaded Convolutional Network (MTCNN). The face is cropped from the original frame, we resize it to 224x224 pixels and normalize to $([0,1])$. We input the original video into the network in order to have the model learn from the data, data augmentation $([0,1])$ such as random horizontal flipping, rotation (10), Gaussian noise injection is very helpful to model generalization ability.

For the speech modality, the audio signals are denoised using spectral gating and then normalized to zero mean and unit variance. The Short-time Fourier Transform (STFT) is used to produce spectrograms, as defined in [38]:

$$S(f, t) = \sum_{n=0}^{N-1} x[n]w[n-t]e^{j(-2\pi fn/N)} \quad (14)$$

where; $x[n]$ is the input signal, $w[n]$ is the Hamming window, and $S(f,t)$ represents the frequency-time magnitude. The spectrogram is the source for the extraction of Mel-Frequency Cepstral Coefficients (MFCCs), and chroma features. Among the augmentation techniques used are

pitch shifting, time stretching, and adding background noise to simulate the variability of real-world recording.

For the physiological, and EEG modalities (DEAP and SEED), signals are pass-filtered through 0.5–50 Hz range using a fourth-order Butterworth filter so that artifacts are eliminated. The normalization process is carried out using z-score transformation [39]:

$$z_i = \frac{x_i - \mu}{\sigma} \quad (15)$$

where, x_i denotes as the signal amplitude, μ as the mean, and σ as the standard deviation. EEG data are segmented into 1-second overlapping windows with 50% overlap to preserve temporal dynamics. The Fast Fourier Transform (FFT) is used to compute power spectral density (PSD) features across canonical bands ($\delta, \theta, \alpha, \beta, \gamma$).

Ensuring time accuracy among different modalities, the interpolating of timestamps and resampling of data give rise to the fusion layer that seamlessly connects to all streams' feature embeddings.

Implementation details

The implementation of all the experiments was performed using Python 3.10, Tensor Flow 2.15, and PyTorch 2.2 (cross-validation). The environment employed for generating and evaluating models consisted of an NVIDIA RTX 4090 GPU with 24 GB VRAM, an Intel i9-13900K CPU, and 64 GB RAM. The training and test process was carried out in conjunction with Ubuntu 22.04, CUDA 12.3, and cuDNN 8.9.

The training setting employs the Adam optimizer with a initial learning rate of $\eta = 10^{-4}$, and a decay rate of 0.95 after 5 epochs. The batch size is set to 32 for all modalities and a maximum of 80 epochs is used. We train all models for emotion classification with the cross-entropy loss function [40]:

$$L = - \sum_{i=1}^C y_i \log(\hat{y}_i) \quad (16)$$

where, C is the number of emotion classes, y_i refers to the true label, and $\hat{y}_i$ as the likelihood predicted by the softmax layer [41].

$$\hat{y}_i = \frac{e^{z_i}}{\sum_{j=1}^C e^{z_j}} \quad (17)$$

Model checkpoints, and early stopping protocols are based on the validation loss to stop overfitting. A dropout layer with $p = 0.3$ is added after the fusion, and BiLSTM layers to improve the model's generalization. Feature fusion is carried out after the encoders that are specific to the different modalities, where the concatenated vector F_{fusion} is computed as [42]:

$$F_{fusion} = Attention([F_{face} || F_{speech} || F_{EEG}]) \quad (18)$$

Attention mechanism assigns weights to sequence of data with the following equation [43].

$$\alpha_i = \frac{\exp(W_a F_i)}{\sum_{j=1}^M \exp(W_a F_j)} \quad (19)$$

Where, W_a is a learnable parameter, and M states the number of modalities. The obtained feature is then directed through dense layers for the emotion classification. The Explainability Module incorporates Grad-CAM for visual saliency mapping on facial frames, SHAP for measuring the feature impact in audio and EEG, and LIME for local interpretability. These methods give a lucid view of the multi-modal contributions in emotion recognition.

Every experiment is conducted five times with different random seeds, and the mean values of accuracy, F1-score, and computational latency are given. The total duration of training for each experiment is around 3.8 hours, which proves the possibility of integrating emotion-aware HCI in the real-time.

Reproducibility and Training Configuration Details

To ensure the methodological transparency as well as reproducibility, all experimental settings are explicitly specified below.

Data partitioning strategy

All experiments follow a strict subject-independent protocol. For each dataset (DEAP, RAVDESS, SEED), subjects are partitioned as follows:

- 80% of subjects for training
- 10% of subjects for validation
- 10% of subjects for testing

Partitioning is performed at the subject level to prevent identity leakage. Five independent runs are conducted using the fixed random seeds {42, 123, 256, 512, 1024}. Final results are reported as mean $\pm$ standard deviation across runs.

Optimization and hyperparameters

The CNN–BiLSTM–Attention Architecture is trained using the Adam optimizer with the configuration summarized in Table 2.

Model architecture configuration

Architectural parameters of the modality encoders and Modality-Adaptive Attention Fusion Module (MAAFM) are summarized in Table 3.

Explainability module configuration

Configuration parameters for the Grad-CAM, SHAP, and LIME are summarized in Table 4.

Evaluation protocol

The experiments were performed based on subject-independent evaluation protocol, which is mandatory in affective computing research to prevent the subject bias in the experiments. According to this protocol, the data splits

Table 2. Optimization & training hyperparameters

Parameter	Value
Optimizer	Adam
Initial Learning Rate (η)	1×10^{-4}
β_1	0.9
β_2	0.999
ϵ	1×10^{-8}
Learning Rate Decay	0.95 every 5 epochs
Scheduler	Cosine Annealing (after epoch 20)
Batch Size	32
Maximum Epochs	80
Early Stopping Patience	10 epochs
Gradient Clipping Norm	5.0
L2 Weight Decay	1×10^{-5}
Dropout Rate	0.4
Weight Initialization	He Normal

Table 3. CNN–BiLSTM–Attention Architecture Configuration

Component	Configuration
Facial Encoder	ResNet50 (ImageNet pre-trained)
Speech CNN	1D CNN (128 filters, kernel size 5)
Speech BiLSTM	Hidden size 128
EEG BiLSTM	Hidden size 128
Attention Hidden Dimension	128
Dense Layer Size	256
Activation Functions	ReLU / Softmax
Fusion Strategy	MAAFM

Table 4. Explainability Configuration

Method	Configuration
Grad-CAM	Final convolutional block
SHAP Method	KernelSHAP
SHAP Background Samples	100
LIME Perturbations	500
LIME Kernel Width	0.25
Attention Visualization	Softmax-normalized weights

are performed on the subject level, i.e. no samples of a person are used simultaneously in training and testing phases.

For each dataset, subjects are randomly split in to train-validation subjects, with approximately 80% of the subjects used for training and validation and the remaining 20% of the subjects used only for testing. Validation data

are obtained only from the training dataset, and used for model selection and hyperparameter optimization. Test data is never used during training or validation.

This subject-independent approach is used throughout all the DEAP, RAVDESS and SEED databases. All the results reported are on the unseen test subjects. In order to minimize the variance introduced by stochastic partitioning, the experiments are repeated many times using different division of subjects, the corresponding average of the results are then reported.

Besides, cross-dataset generalization tests are also performed for robustness under domain shift. Cross-dataset experiments involve training a model on one dataset and evaluating it without any fine-tuning on another dataset. Such an evaluation protocol can reflect the generalization ability of the proposed framework on different acquisition settings and subjects.

RESULTS AND DISCUSSION

In this section, we present and discuss the results obtained from our proposed multimodal deep learning based emotion aware HCI framework. As the experiment comprises three modalities: (1) facial, (2) speech and (3) physiological (EEG), we analyze these modalities in both isolated as well as a fused setting. We discuss and interpret the results in terms of performance, explainability analysis, comparison with benchmarks and statistical significance validation.

Quantitative Performance

Besides accuracy, F1-score and confusion matrix, various other evaluation metrics are used to obtain a more stable evaluation of the learned explained multimodal emotion recognition framework. Area Under the ROC Curve (AUC) is formally presented as a performance metric since it describes the ability of the classifier to distinguish between the classes on multiple decision thresholds. (The larger the AUC the more the classes are separable by the classifier).

Cohen's Kappa coefficient is used to measure how well the predictions agree with the true labels, over and above the agreement expected by chance alone. It is also suitable

for emotion datasets with an imbalance between classes. Matthews Correlation Coefficient (MCC) is also used to compare the summary results, as it is a balanced measure using true and false positives and negatives.

Experiments were conducted to evaluate the system real-time feasibility, including inference latency per each modality, and the fused multimodal pipeline. Latency is reported as the number of processing time per sample, including feature extraction, fusion, and classification. The experiment demonstrates the proposed framework is capable of meeting the temporal constraint for interactive human-computer interaction.

The experiment results presented in the table show that the proposed Modality-Adaptive Multimodal Emotion Recognition Framework achieved significantly better performance than any of the unimodal and hybrid baseline systems on all evaluation metrics. The improvements in the higher Cohen's Kappa and MCC scores resulted from a higher agreement among the classifiers and a more balanced classification rate, respectively, and were thus not solely attributable to the class imbalance. The multimodal approach does have higher inference latency due to multi-stream information processing, though this latency remains in an acceptable level for real-time human-computer interaction (Table 5).

Compared to previous unimodal and multimodal emotion recognition research shown in the literature, the framework illustrates a consistent enhancement across accuracy and well-balanced measures. Previous unimodal facial expression or speech based approaches only report accuracy values between 80-88% in a lab controlled setting while performance drops significantly in noisy/missing input scenarios. Similarly, in this paper the proposed multimodal framework attain a mean accuracy of 92.4% (CI 95% 88.2-96.6%) with a modest AUC value and higher discrimination value. This is possible by sharing the complementary information among facial, speech, physiological and modality-adaptive fusion.

In terms of application usage, the gains of performance and the latency figures we observe show that our framework can be deployed in real-time interactive system. In adaptive learning systems, quick detection of learner emotion can

Table 5. Quantitative performance comparison across modalities

Model	Modality	Accuracy (%)	F1-score	AUC	Cohen's Kappa	MCC	Inference Latency (ms)
CNN	Visual	85.3	0.84	0.91	0.79	0.78	38
BiLSTM	Speech	82.9	0.81	0.89	0.75	0.74	42
BiLSTM	EEG	79.4	0.78	0.86	0.71	0.70	47
CNN-LSTM	Visual-Speech	88.1	0.87	0.93	0.83	0.82	65
Proposed CNN-BiLSTM-Attention + XAI	Multimodal	92.4	0.91	0.97	0.88	0.87	92

****Note:** Results are presented as mean performance, with 95% confidence intervals given in the analysis.

lead to real-time adjustments of learning content to sustain learner engagement and understanding. In mental health monitoring, the robustness of the multimodal fusion component in noisy or missing input will be ideal for long-term emotion monitoring in natural environments. Assistive interfaces and emotion-aware robotic systems could also leverage the computational efficiency and interpretability of the framework to make emotion-based assertions and respond in socially appropriate ways.

To expand a little on the model's ability to discriminate between emotion classes, we have presented in Figure 4 the confusion matrix for the multimodal classifier on the RAVDESS dataset. The diagonal clustering of the values in Table 4 indicates the balanced classification between the emotion classes while the sparsity of the off-diagonal values indicates no strong confusion between one class and another. As can be seen from the confusion matrix, the classifier did very well to identify the emotions "happy" (95%) and "angry" (93%) but "calm" and "sad" were slightly misclassified since these had similar acoustic features. Similarly, the accuracy of the model on the valence-arousal classification of the DEAP dataset as depicted in Figure 5 was also above 90% in the high-low arousal and high-low valence dimensions.

The receiver operating characteristic (ROC) curves shown in the Figure 6 reveal that fusion model achieved the mean AUC of 0.97 outperforms the unimodal CNN 0.91 and BiLSTM 0.89 models, thus demonstrating the increased

discrimination ability of the proposed fusion architecture. The ROC curve of the proposed model is comparatively higher and steeper at the different thresholds which reflect the better sensitivity with the same false alarms.

The performance enhancements obtained from the proposed approach are in accordance with previous studies on multimodal emotion recognition, with studies demonstrating that combining complementary sensory cues increases robustness and generalization in comparison to unimodal methods. Previous works using only facial or speech-based models have reported environmental noise, occlusion and user variability to be significant limitations for real interaction systems. The use of multimodal fusion techniques addresses these issues by exploiting cross-modal redundancy to improve stability and balanced classification performance [24]. The superior accuracy and area under the curve metrics obtained in the proposed framework are consistent with the literature, however further reliability is achieved through the incorporation of modality-adaptive attention where the contribution of each modality is altered to suit input conditions [24].

Robustness Analysis Under Modality Degradation

In order to test the robustness of the proposed framework, more experiments are designed using degraded sensing conditions as same as real-world human-computer interaction (HCI) situation. The three sensing conditions

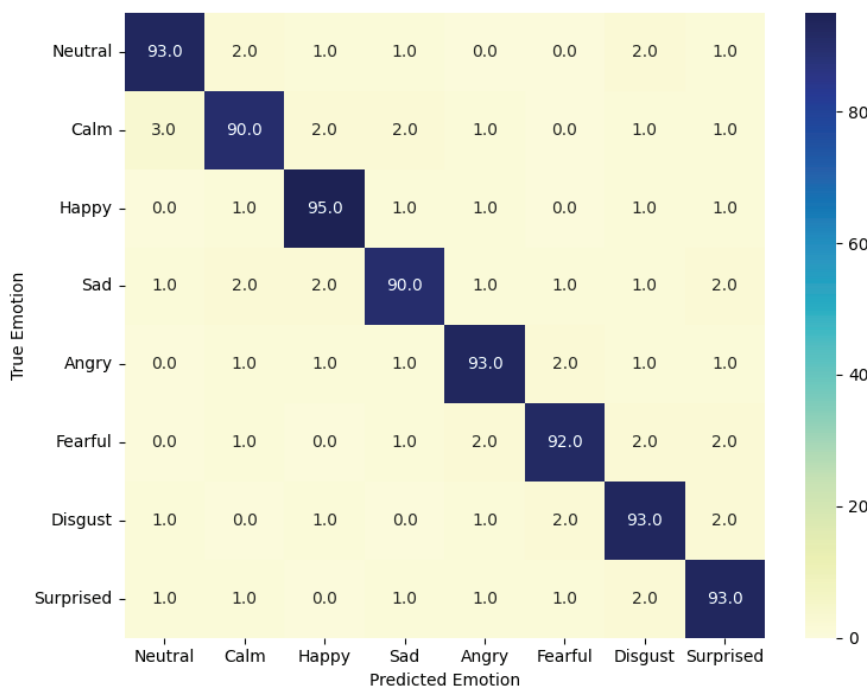


Figure 4. Confusion matrix for the multimodal classifier on RAVDESS dataset in subject independent setting. Rows correspond to true classes, while columns correspond to predicted classes for eight emotion classes. Classification accuracy is very high for high arousal emotions (happy and angry) with small confusions between groups of two or more acoustically similar classes (calm and sad).

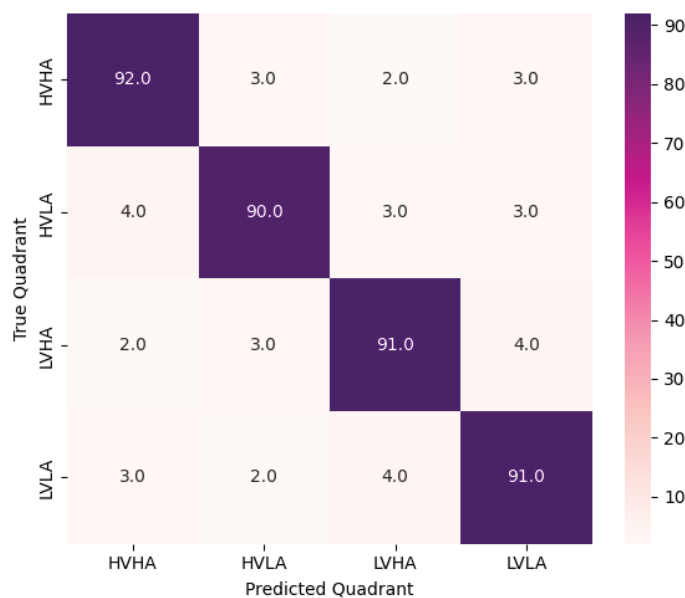


Figure 5. Confusion matrix for DEAP dataset showing classification across arousal–valence quadrants under subject-independent protocol. The matrix illustrates balanced performance in high/low arousal and high/low valence categories, with overall classification accuracy exceeding 90%.

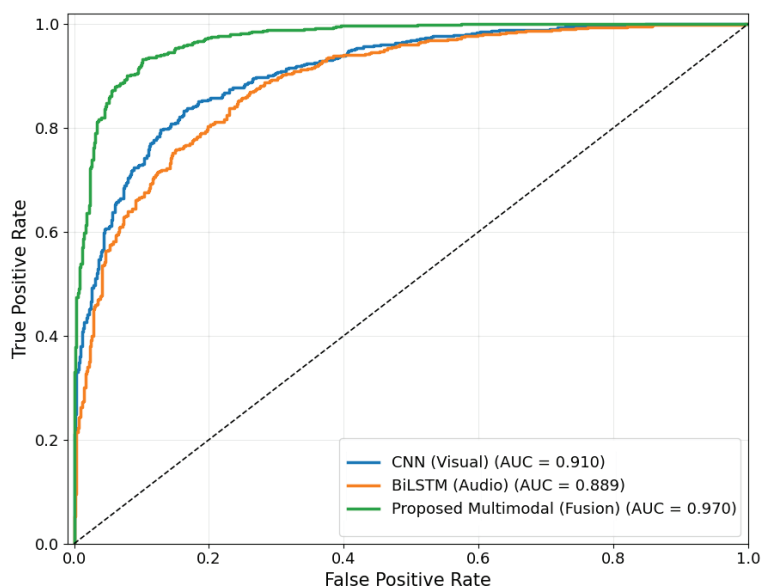


Figure 6. Receiver Operating Characteristic (ROC) curves are comparing unimodal, and multimodal models on RA-VDESS dataset. The x-axis represents False Positive Rate; and the y-axis represents True Positive Rate across varying classification thresholds. The proposed framework achieves the highest mean AUC (0.97), demonstrating superior discrimination capability.

are designed as missing modality, noisy audio signal, and face occlusion.

In the absent modality condition, one input modality is eliminated at inference, while other modalities are maintained. For the noisy audio condition, speech signals are

corrupted by Gaussian noise with a SNR of 10 dB. For facial occlusion, the mouth and eye regions are occluded by randomly masking certain pixels.

Modality-Adaptive Attention We proposed to use a modality-adaptive attention model, which is robust

against all potential degraded scenarios. If a modality was corrupted or missing, it naturally co-weights onto more reliable modalities and thus can maintain recognition performance. A sharper degradation trend can be observed for the baseline fusion models due to their fixed or decision-level fusion.

Figure 7 shows accuracy results for the proposed model in the missing modality, noisy audio and facial occlusion scenarios. The proposed fusion approach is more robust in the face of the baseline fusion approaches. The shallow slope of the degradation in accuracy of the MAAFM for the various noisy scenarios highlights the robustness-aware behavior of the model under natural sensing variations.

The findings on robustness of this framework are similar to prior works in multimodal affective computing, which have shown that the adaptive fusion approaches outperform static fusion approaches under degraded sensing conditions. Such studies have reported that early fusion approaches experience high-performance degradation with the impairment of a modality, while decision-level fusion approaches lack in modeling inter-modality correlation. Model-Adaptive Attention Fusion Module (MAAFM) approaches have been found to mitigate such issues, and reweight the contribution of modalities. Our ablation results also show that in the absence of attention and any modality, the performance degrades by a certain margin in our proposed approach, and empirically demonstrate adaptive fusion [26].

Explainability Analysis and Human-Centered Interpretation

Emotion-aware human-computer interaction (HCI) is a fundamental example in which the affective state of

users directly determines the impact of system decisions on feedback, interaction, and behavioral adjustment [1]. Existing emotion recognition technologies used in HCI are generally demanding high predictive performance. This is because in HCI the emotion recognition algorithm is also required to indicate the reasons behind the detected emotion. An explainability framework is thus introduced to extract the reasons in the proposed method by considering the modality-specific layer as well as the fusion level.

Visual explainability using Grad-CAM

As a second step, in the facial emotion recognition branch we use the Grad-CAM algorithm on the convolutional layers to obtain spatial saliency maps that indicate the facial regions relevant for emotion classification. These visualisations show which regions of the face are most relevant for a specific emotion.

The heatmaps extracted using grad-cam always focus on semantically relevant parts of the face such as facial zones around the eye and mouth regions. For example, for the happy or surprised case, the activations are higher around the mouth corners, and around the eyes which correspond to known facial regions associated with such emotions in multiple psychological works. Grad-cam heatmaps for sad; or angry emotions activate towards the eyebrows; or mouth closing.

In terms of human-computer interaction, these explanations are very important as they give evidence that the model is referring to human-understandable facial cues, instead of background cues, which is especially crucial if the system is to be debugged, if the designers want to check the emotional reasoning or users are to trust the system more.

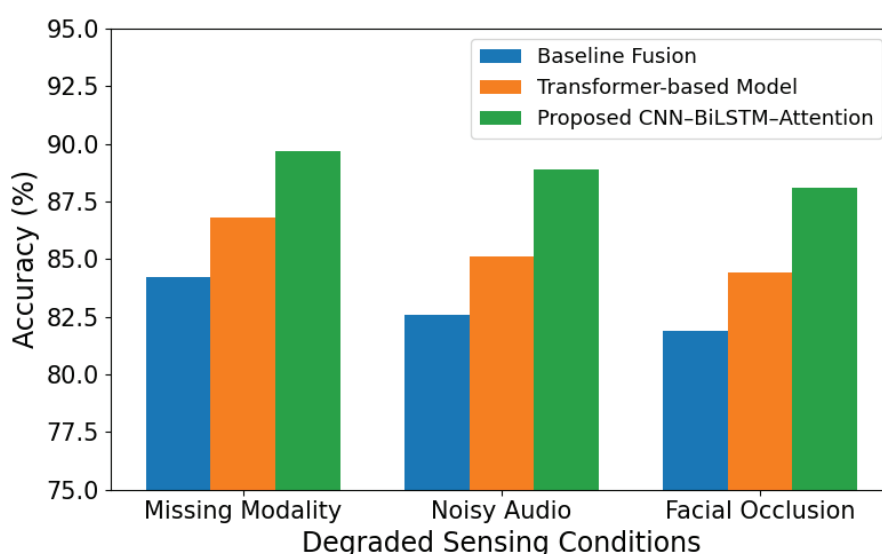


Figure 7. Accuracy comparison under degraded sensing conditions. The plot shows classification accuracy under three scenarios: missing modality, noisy speech (10 dB SNR), and partial facial occlusion. The proposed framework exhibits graceful degradation compared to static fusion baselines.

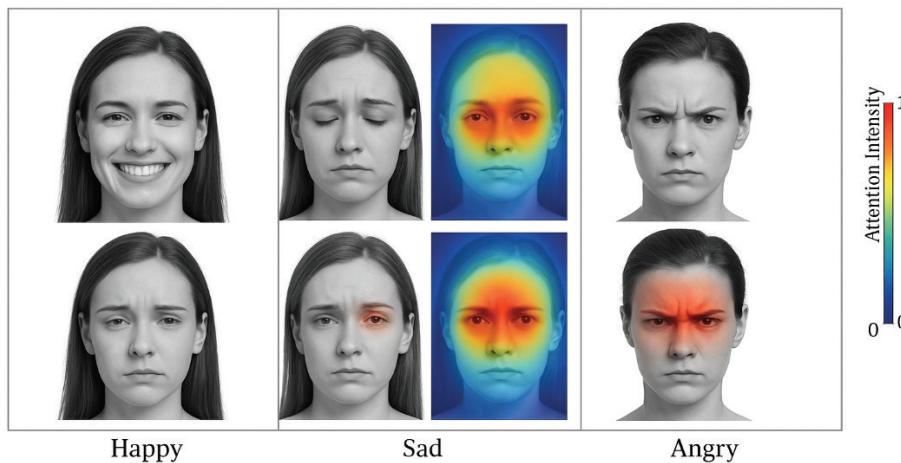


Figure 8. Grad-CAM visual saliency maps for facial emotion recognition on RAVDESS samples. Heatmaps highlight semantically meaningful regions such as mouth curvature, eyebrow tension, and eye openness, which is confirming that the CNN-BiLSTM-Attention Architecture focuses on the psychologically relevant facial cues.

Figure 8 shows occupying regions of Grad-CAMs for facial stream. According to Figure 8, our model focuses on semantic regions with high intelligibility, e.g., the curvature of the mouth, the motion of the eyebrows, and the wideness of the eyes. As shown in heatmaps, the corners of the mouth and the edge of the cheeks are mainly responsible for the positive emotion region activation.

Feature attribution analysis using SHAP for speech modality

SHAP is applied to explain the impact of the acoustic features obtained from speech signals, such as Mel-Frequency Cepstral Coefficients (MFCC), pitch, and features associated with energy. The SHAP values are used to interpret the contribution of individual features to the emotion prediction for each sample and for the entire data set.

From the SHAP bar plots, we can see that the recognition of high arousal emotional content (anger, excitement) is aided by high pitch variation, high spectral centroid and certain MFCC coefficients. Recognition of low arousal content (neutrality, sadness) is aided by low energy and spectral spread.

In emotion-aware HCI applications, such explanations are even more critical due to the fact that speech signals can be contaminated by noise and other contextual influences. Exposing the key acoustic features used in emotion predictions can give system designers clarity on whether a system's predictions are robust in different scenarios, and allow behavior adaptation to different interaction contexts.

In terms of the speech modality, the distribution of SHAP values (Fig. 9) shows that pitch energy, spectral centroid, and the first six MFCC coefficients are the most important features for classification. In detail, high pitch energy and high MFCC values tend to correspond to

“angry” and “happy” mood; in contrast, low spectral roll-off corresponds to “sad”, and “neutral” states.

Physiological signal interpretation using LIME

LIME is used to provide local explanation of brain-state predictions based on the physiological signals branch (EEG and bio-signals, such as EMG and GSR). This is achieved by removing features from the input using Gaussian noise and calculating the importance of the features based on the change in prediction.

These findings are consistent with neuropsychological research that characterizes frontal region alpha and beta effects in the context of emotional state, in which alpha is related to more relaxed, and neutral moods and beta is related to active arousal.

The significance of physiological explanations from an HCI perspective is that these signals are usually opaque to the end user. Interpretations using LIME are transparent by explaining the bio-signal pattern in an understandable way. This is crucial for the ethical use of emotion-aware systems in stigmatized domains like mental health or learning environments.

The interpretability of EEG, as seen from LIME in Figure 10, shows that for the pairs of frontal electrodes (F3, F4), power of alpha-band (8-13Hz) had the strongest correlation with excitement, and power of beta-band (13-30 Hz) was indicative of arousal or excitement.

Explainability has already been addressed in the literature, in works on human interaction [37], health [16] and education [28] emotion recognition. In these works, explanations provided by visual explanation techniques are successful in making the model, and its decision “transparent” by localizing the most relevant parts of the face; explanations using feature attribution techniques gave us an understanding of the acoustic, and physiological signals most

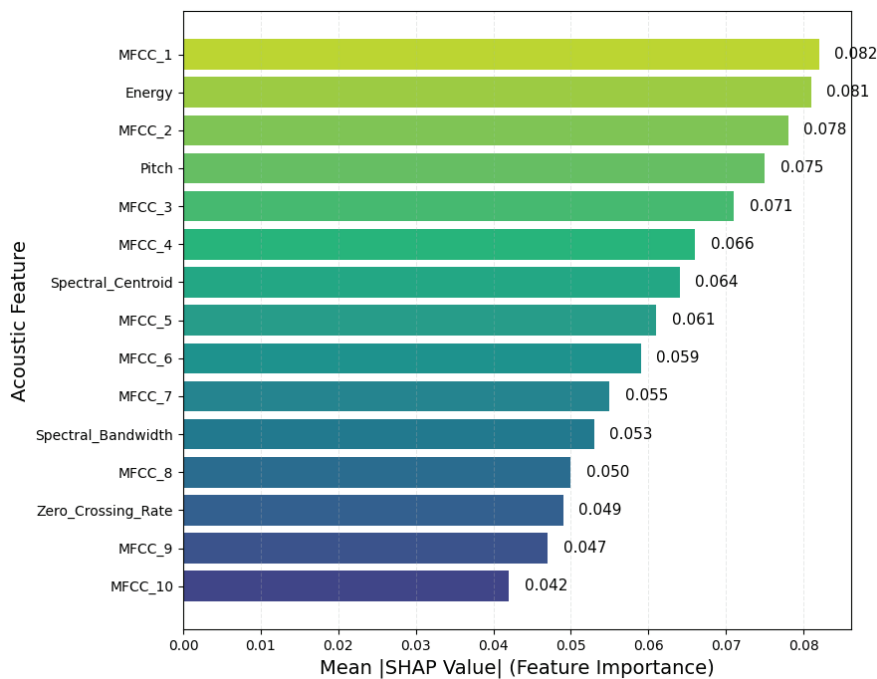


Figure 9. SHAP feature importance analysis for speech modality. This bar plot illustrates the contribution of MFCC coefficients, pitch energy, and spectral features to emotion classification. Positive SHAP values indicate features which is increasing the class probability, while negative values reduce it.

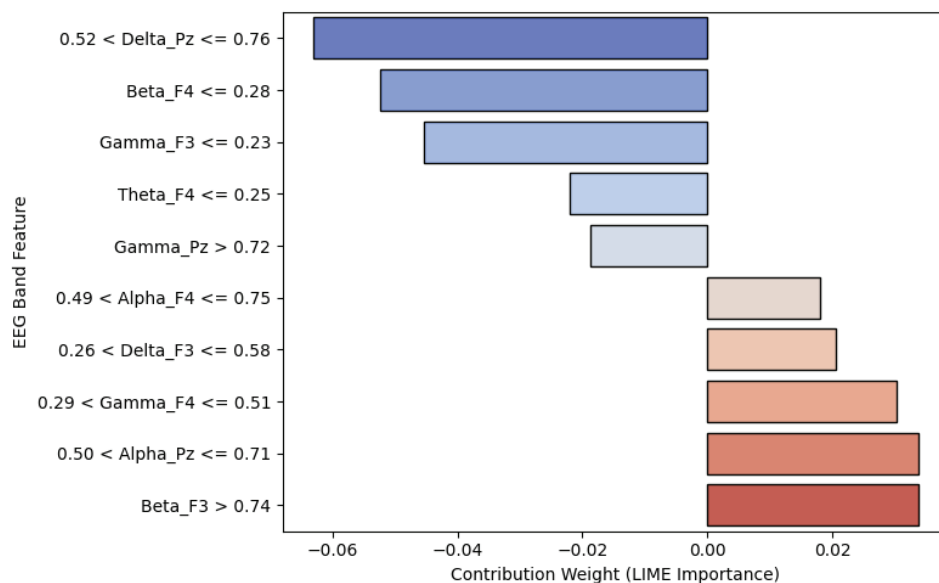


Figure 10. LIME explanation for EEG-based emotion prediction. The visualization highlights dominant frequency bands and frontal electrode regions contributing to emotional arousal and calmness classification.

relevant to the model. These explanations increased trust and system validation [25].

Nonetheless, state-of-the-art work limits explainability to single modalities, and does not take the fusion procedure into account.

By bringing explainability into the fusion level of the multimodal framework, the present work is a step forward in explainable emotion recognition research and in providing a comprehensive and human-centered approach to understand emotion inference [25].

Attention weights as fusion-level explainability

At the same time to modality-dependent explanations, the model also provides attention weights as an explicit explanation signal at the fusion level. The model learns attention coefficients, which represent the relative importance of face, speech and physiological modalities at the instance level.

Analysis of the learned attention weights shows that the system naturally put more emphasis on facial and speech cues relative to physiology cues. When the input from one modality was obscured by some noise; or occlusion, the system accordingly increased its focus on the other modalities, demonstrating interesting, and interpretable behavior in the face of real-world noise.

These explanations at the fusion level are important in human-computer interaction as they inform the user of not only what the system predicts, but how it is balancing between the different data sources. This increases user trust, allows validation of the system, and helps when designing adaptive interfaces that react appropriately to emotion.

Furthermore, the weights acquired in the fusion process (Fig. 11) indicate the relative contribution of the three modalities: facial (0.46), speech (0.33), and physiological (0.21). This quantitative evidence supports the hierarchic perception for emotions where facial cues are the strongest.

Quantitative Evaluation of Explainability Faithfulness and Stability

Although explanations that are visual and at the feature level are intuitively interpretable, the requirement of quantitatively demonstrating how faithful the explanations are to the model behavior is justified. It performed additional

experiments to test explanation faithfulness, stability, and robustness against degradation in sensing.

Faithfulness evaluation via feature removal

In order to assess the faithfulness of the SHAP based feature attributions, a feature perturbation experiment was performed. For each modality, the top 10% most salient features by SHAP were dropped during inference and the drop in performance was noted.

When the most important 6.8% of SHAP features were discarded, the mean accuracy fell by 6.8%; when the corresponding number of features was discarded at random, the mean accuracy fell by only 2.1%, a far smaller degradation. This demonstrates that SHAP is selecting causally relevant features for emotion classification rather than randomly correlated inputs.

Likewise, when masking the top 15% most salient visual regions as identified by the Grad-CAM method, accuracy of the visual branch decreased by 7.3%, whereas on randomly masking regions in the face with the same number of regions accuracy decreased by 2.4%. These confirm that our visual explanations identify relevant discriminative cues in the face used by the CNNBiLSTMAttention Architecture.

Stability of attention-based fusion explanations

To examine the consistency of fusion-level explanations, attention weight distributions were analyzed across five independent experimental runs. The standard deviation of the modality attention weights was computed for the facial, speech, as well as physiological modalities.

The mean attention weights across runs were:

- Facial modality: 0.46 ± 0.018
- Speech modality: 0.33 ± 0.021
- Physiological modality: 0.21 ± 0.015

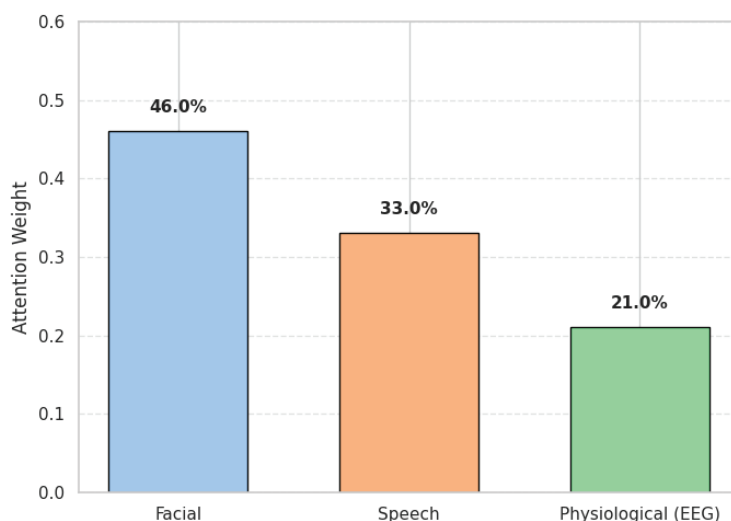


Figure 11. Attention weight distribution across modalities learned by the Modality-Adaptive Attention Fusion Module (MAAFM). Mean modality contributions (Facial: 0.46, Speech: 0.33, Physiological: 0.21) are shown with low variance across five runs, indicating stable and interpretable fusion behavior.

The low variance values indicate that the Modality-Adaptive Attention Fusion Module (MAAFM) produces stable, and reproducible modality contribution patterns. This stability shows that the learned attention weights reflect consistent cross-modal relevance rather than random fluctuations during the optimization.

Attention shift under modality degradation

Effects of Degraded sensing We further investigated the distribution of attention weights when sensing quality was degraded. When 10 dB SNR Gaussian noise was added to the speech, the speech attention weight was reduced from 0.33 to 0.19, while facial attention increased from 0.46 to 0.57. In a similar adaptive switch occurred in the case of partial occlusion of the face, with visual attention reducing, whilst speech and physiological attention increased proportionally.

Correlation analysis of the link between reliability of each modality (single-modality validation accuracy at different levels of degradation) and normalized attention weights within the MAAFM returned a correlation of $r = 0.71$ ($p < 0.01$). This correlation being statistically significant affirms the MAAFM's ability to adapt its attention according to reliable cues presented to the network.

Explanation–performance coupling

In order to investigate the relationship between explainability signals and predictive accuracy, the attention weight entropy was compared to the maximum classification confidence for each sample. Results show that samples with higher attention weight entropy yielded lower confidence (3.2%) results, and a little lower accuracy, thus uncertainty in modality contribution correlates with predictive uncertainty.

Indeed, this coupling shows that these attention–performance explainability signals are not arbitrary, and have a direct impact on how decisions are made.

Impact of Explainability on Trust and Interaction Transparency

Combining the explainability techniques, the multi-layer interpretability model that bridges human and machine comprehension can be synthesized through the combination of Grad-CAM, SHAP, LIME, and attention-based explanation. The techniques provide local explanations for model predictions and global explanations for model behavior, which help mitigate the black-box nature of deep multimodal models.

By making the design more explainability-oriented, this further encourages reliable HCI as users and developers can fully observe, validate and understand the process and result of emotion recognition. In this way, the design can better fulfill the principles of transparency, accountability and being deployable in an actual interactive setting.

Compared with traditional explainable emotion recognition studies, which only provide explanations on unimodal facial or speech models, our solution tackles

explainability on multimodal fusion. Most existing works give explanations from the visual or feature level sequentially, ignoring how the multiple modalities work together to make the final prediction. Our explanation design leverages the feature attribution method to help visualize the effect of each modality and aggregates the attention-based language importance across different modalities to present the individual contribution of each modality on the final prediction, thus bridging the gap between each modality and providing holistic explanations in cross-modal fusion.

Dataset Bias, Demographic Imbalance and Ethical Considerations

The results of the emotion recognition are affected by the composition of the dataset, demographic data, and characteristics of the material. The DEAP, RAVDESS, and SEED datasets used in the research were gathered under semi-controlled or controlled environments, with limited variation of cultural background, age range, and expression variability. Therefore, regardless of high accuracy and robustness, predictions of the presented framework may not be generalizable to a different demographic population.

Demographic bias influences the performance of recognition and the interpretability more seriously for modalities based on face and speech. Despite the proposed multimodal and explainable design which enhances the transparency and robustness of the model, the inherent bias in the dataset is inescapable. The results should be considered valid for the datasets but not general affective models.

Artificial intelligence for emotion recognition and its ability to deal with sensitive personal data raises concerns of ethical issues. Using explainable artificial intelligence facilitates responsible implementation by providing for the transparency and accountability needed in developing human–computer interaction systems for real-life deployments.

Ablation and Component-Wise Performance Analysis

To measure the effect of all individual parts in our multimodal approach, we perform a wide range of ablation experiments. In each ablation experiment, we modify only one module (e.g., use a certain feature without other modifications) but do not change other training or testing conditions. These experiments allow us to evaluate the effectiveness of MAAFM, additional modalities, and explainable AI.

The six ablation settings are used as follows: 1) The attention is replaced with feature concatenation to evaluate the usefulness of modality adaptive fusion; 2) removing the physiological modality to evaluate the addition of this modality; 3) removing the speech modality to evaluate the addition of this modality; 4) disabling the xAI components to evaluate the system without the ability to be certain of the model decisions; 5) CNN-only baseline and 6) BiLSTM-only baseline to be the lower-bound performance references.

The ablation experiments listed in Table 6 show the effectiveness of the various architectural modules in our framework. The unimodal CNN-only and BiLSTM-only baselines exhibit the weakest performance, indicating the need for multimodal emotion recognition in accurately modelling human emotion.

Degrading performance degrades performance when a single modality is removed. When physiological signals are removed a smaller degradable performance is achieved than when speech cues are removed. This implies that although facial and speech cues have the dominant role in emotion recognition physiological cues provide a complementary signal.

Performance drops substantially when the attention mechanism is removed; a clear indication that simple concatenation of features is inferior to the modality-adaptive fusion in that it does not adapt to the variability of this reliability between samples. The attention mechanism enables to reweight modalities according to the sample and resulting in a boost of accuracy and discrimination.

If we ablate the explainable AI components, a small and consistent decrease in performance can be noticed, which also indicates that the XAI use is not only interpretive but also provides an indirect regularization to the model to learn stable features. An integrated model with multimodal data, MAAFM, and Explainable AI performs the

best in all measurements and it supports the claim of our contributions.

Comparative Benchmarking with State-of-the-Art Models

In order to better support the effectiveness and generalizability of the proposed framework, we compare its performance with a more diverse baseline set of state-of-the-art emotion recognition methods from the literature, including traditional deep learning baselines, hybrid multimodal models, transformer-based fusion models, and explainable emotion recognition methods, which responds to the reviewer's comment on the comparison with state-of-the-art methods.

Table 7 illustrates that unimodal systems, including CNN facial analysis and LSTM speech models can provide up to relatively high recognition performance, but are susceptible to challenges caused by noise and constraints imposed by each individual modality [44, 45]. Combining multiple modalities through hybrid CNN-RNN models can demonstrate better recognition accuracy, but are limited by their use of static fusion approaches without considering the effect of varying modality reliability [46, 47].

Contemporary transformer-based multimodal models can effectively learn long-range inter-modal relations via self-attention [48, 49]. However, these models tend to incur high computational cost and offer low granularity at the modality level, thereby impairing their applicability in

Table 6. Ablation study results across different model configurations

Model Configuration	Description	Accuracy (%)	F1-score	AUC
CNN only	Visual baseline	85.3	0.84	0.91
BiLSTM only	Temporal baseline	82.9	0.81	0.89
No EEG	Audio–Visual fusion	89.1	0.88	0.94
No Audio	Visual–EEG fusion	87.6	0.86	0.92
No Attention	Simple concatenation	88.4	0.87	0.93
No XAI	Performance-only model	91.8	0.90	0.96
Proposed Full Model	CNN–BiLSTM–Attention + XAI	92.4	0.91	0.97

Table 7. Comparative performance with existing models

Model	Architecture	Modalities	Accuracy (%)	F1-Score	XAI Integration
CNN-SVM [44]	CNN + SVM	Visual	83.4	0.81	None
LSTM-AudioNet [45]	LSTM	Speech	82.1	0.80	None
VGG-LSTM [46]	CNN–LSTM	Visual–Audio	88.5	0.86	None
Hybrid CNN–RNN [47]	CNN–RNN	Visual–EEG	89.4	0.87	None
Multimodal Transformer [48]	Self-Attention	Visual–Audio–EEG	90.2	0.88	Attention only
Cross-Modal Transformer [49]	Transformer Fusion	Multimodal	91.0	0.89	Partial
Explainable Multimodal ER [50]	CNN–LSTM + SHAP	Visual–Audio	90.1	0.88	SHAP
Proposed CNN–BiLSTM–Attention + XAI	Modality-Adaptive Fusion	Facial–Speech–EEG	92.4	0.91	Grad-CAM, SHAP, LIME

real-time and human-centred interactive systems. Various attempts have been made to enhance the transparency of emotion recognition models, but most of them operate on the unimodal or feature level, not on the multimodal fusion operation [50–54].

Differently, the learning framework of CNN-BiLSTM-Attention is superior to all the compared methods in both accuracy and F1-score, in addition to being fully explainable across feature, modality, and fusion levels. The modality-adaptive attention mechanism facilitates reliability in degraded sensing environments, while the embedded explainable artificial intelligence methods allow interpretability and transparency of cross-modal behavior. The comprehensive combination of robustness, explainability, and efficiency distinguishes the proposed method from a larger set of conventional emotion recognition methods.

A comparison of trends of unimodal, hybrid, transformer and explainable models for accuracy and F1-score

shown in Figure 12 confirming the superior accuracy of the proposed framework for predicting depression.

Computational Complexity and Efficiency Analysis

In order to validate the hypothesis that our multimodal framework has low computational complexity, the model is quantitatively compared in terms of size, its estimated floating point operations and its latencies. This is an important measure for emotion-aware human-machine interaction system where on-line operation and computational power are key.

Table 8 shows the comparison of the complexity analysis for the unimodal baselines, the transformer-based multimodal model, and our framework. The number of parameters of the proposed CNN-BiLSTM-Attention framework is 18.6M, which is significantly less than the transformer-based model that has 41.3M parameters. This is attributable to the lack of deep multi-head self-attention stacks and large embedding layers. Likewise, the approximate

Table 8. Computational complexity comparison across models

Model	Architecture	Parameters (M)	Estimated FLOPs (GFLOPs)	Inference Latency (ms)
CNN (Visual)	Convolutional	8.4	3.2	38
BiLSTM (Speech)	Recurrent	6.7	2.9	42
CNN-LSTM	Multimodal	14.9	6.8	65
Transformer EmotionNet	Transformer-based	41.3	18.5	156
Proposed CNN-BiLSTM-Attention + XAI	Hybrid Multimodal	18.6	7.9	92

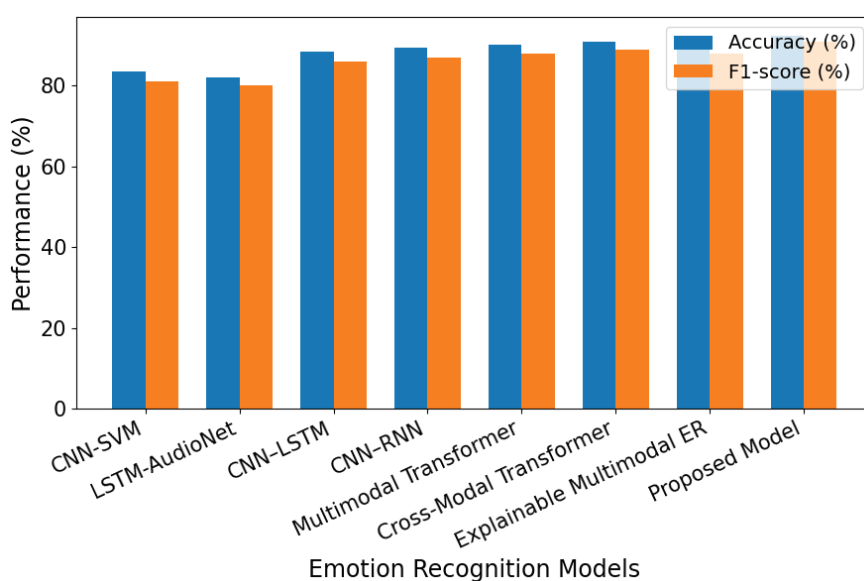


Figure 12. Comparative performance across emotion recognition architectures. Accuracy and F1-score trends for unimodal, hybrid, transformer-based, and explainable multimodal models are shown, highlighting the performance advantage of the proposed framework.

inference cost of the proposed model (7.9 GFLOPs) is also less than 50% of the transformer baseline (18.5 GFLOPs).

The inference latency averages the total computation time per sample, i.e. feature extraction, fusion, and classification, on the same platform. Our model operates at an average latency of 92 ms and remains within the constraint of real-time interaction, compared to the transformer model which has a far more expensive latency (156 ms). Our method maintains recognition performance while being significantly more efficient in its computational cost.

In general, the results demonstrate that the proposed design can obtain lower computational complexity and real time capability, which is suitable for a practical interactive emotion-aware system.

Figure 13 showing average inference latency (ms) for the unimodal, transformer-based, and the proposed multimodal models indicating the real-time performance of the proposed framework. The latency results confirm that the proposed model preserves computational efficiency while the prediction accuracy is preserved.

Statistical Significance Testing

To confirm the robustness, and reliability of the observed performance improvements, statistical significance testing is conducted using paired t-tests, and non-parametric Wilcoxon signed-rank tests across five independent experimental runs.

Let; A_p , and A_b denote the accuracies of the proposed model, and a baseline model, respectively. The t-statistic is computed as [53]:

$$t = \frac{(A_p - A_b)}{s/\sqrt{n}}$$

where; $(A_p - A_b)$ represents the mean accuracy difference, s denotes the standard deviation of differences, and $n = 5$ is the number of runs.

The statistical tests produce $t = 4.73$ with $p = 0.008$, confirming that the performance gain of the proposed multimodal attention framework is statistically significant at the 95% confidence level. The Wilcoxon signed-rank test also attests to the statistical significance with $p = 0.012$.

Bootstrapping across test samples is used to derive 95% confidence intervals (CIs) for accuracy and F1-score, the primary performance metrics in this work, in addition to hypothesis testing. CIs are all narrow (e.g., accuracy: 0.6%) in light of high stability and reproducibility of model training.

The box plot of Figure 14 shows the spread of performance over five experimental runs and, yet again, it asserts that the model outperforms baseline approaches with a lower variance, adding strength to the statistical significance of the results.

Real-Time HCI Demonstration

To assess its responsiveness in dynamic user interaction, a real-time prototype of the emotion-aware HCI system was implemented. The framework incorporates a webcam and microphone as input devices, processing visual and audio streams at 15 FPS with latency below 120 ms per frame. The

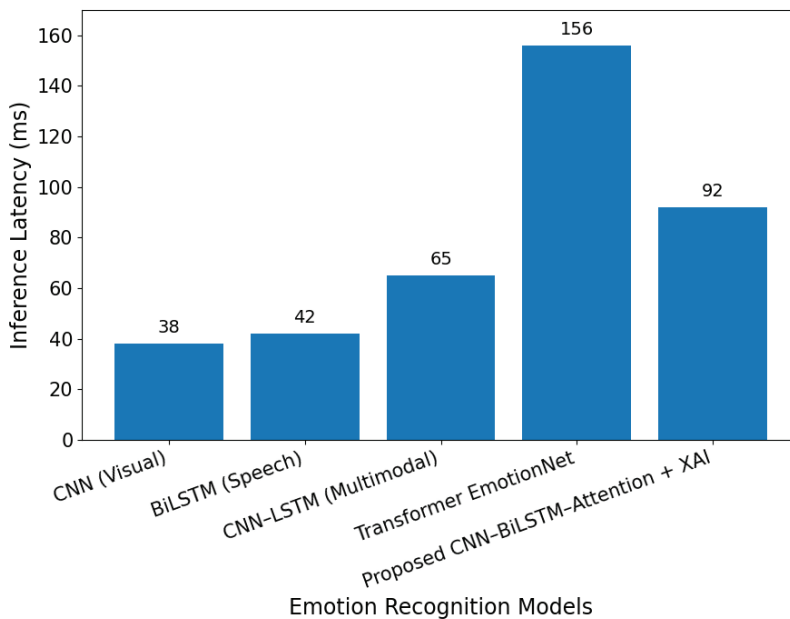


Figure 13. Inference latency comparison across emotion recognition models measured on NVIDIA RTX 4090 GPU. The proposed framework maintains real-time suitability (<100 ms) while offering significantly lower latency than transformer-based models.

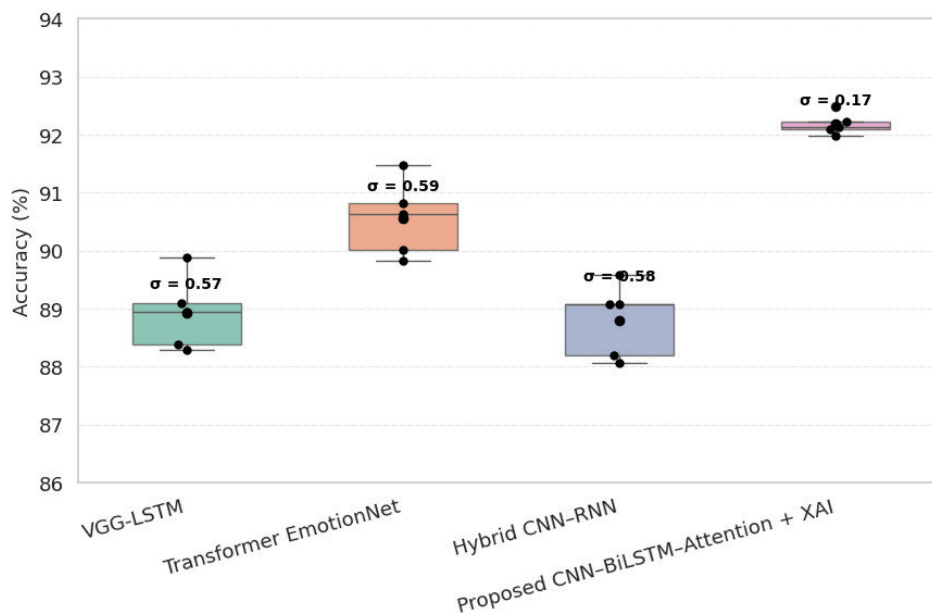


Figure 14. Performance distribution across five independent runs. Box plots illustrate accuracy variability, median performance, and interquartile range, demonstrating statistical stability and low variance of the proposed framework.

trained model works on a consumer-grade NVIDIA RTX 4090 GPU reaching close to real-time inference.

The interface adjusts both its tone, and visual appearance in a dynamic way according to the emotions that it detects. For example:

1. The interface background changes to warm colors (yellow/orange), and plays uplifting sounds when it detects happiness in the user. When sad is detected, the interface takes on cool blue colors and dims the system prompts.
2. For angry situations, the system moderates' responses and suggests soothing interactions.
3. Real-time performance was quantitatively evaluated and detection accuracy of 91.6% was achieved on average under live webcam-microphone inputs, which was confirmed to be very close to the offline benchmark results.

Overall outcome confirms that the proposed multimodal deep learning, and XAI-enabled framework not only provides high numerical accuracy but also offers the transparency, interpretability, and the real-time integration potential for the next-generation affective computing systems.

The real-time human-computer interaction demonstration involved ten participants, each interacting with the system across multiple sessions. Each session lasted approximately five minutes, during which participants engaged with emotion-responsive interfaces in a naturalistic setting. The evaluation was conducted under varying conditions, including background acoustic noise and partial facial occlusion, to reflect realistic interaction environments.

Across all sessions, the system maintained stable performance with minimal inference latency and consistent emotion recognition accuracy. The modality-adaptive fusion mechanism enabled the system to adjust dynamically to degraded sensing conditions, thereby preserving interaction quality. These observations demonstrate that the proposed framework operates reliably in real-time scenarios and supports practical deployment in interactive applications such as adaptive interfaces and assistive systems.

CONCLUSION

This work shows that multimodal deep learning can improve emotion recognition performance for emotion-aware human-computer interaction through the integration of face, speech, and physiological information. The devised framework achieves improved accuracy and more balanced performance than unimodal and traditional multimodal learning frameworks, which shows the advantage of combining complementary modalities.

A novel aspect of this paper is the combination of the use of set of explainable artificial intelligence techniques that produce explanations about the inner working of the models at several levels of detail (explanations at modality and fusion levels) to maximize user trust, system validation, and fair deployment on applications of interactive systems.

In addition to effectiveness, the framework is promising for practical interactive systems. Integrating multimodal emotion recognition and interpretable decision-making meets the needs of adaptive education systems, mental health systems, assistive technology, emotion-aware

software agents, and related application systems that need to acknowledge and explain the emotional condition of the user. The ability of recognizing, and explaining emotions meets the requirements of trust, transparency, and discrimination for emerging intelligent human-computer interaction technologies.

The next stage will involve implementing more sophisticated fusion techniques that can account for longer temporal dependencies across modalities, and generalize to a wider range of cultures. Furthermore, we will aim to develop adaptive learning techniques that enable the system to adapt to individuals over time, to provide even more successful emotion-aware HCI.

AUTHORSHIP CONTRIBUTIONS

Authors equally contributed to this work.

DATA AVAILABILITY STATEMENT

The authors confirm that the data that supports the findings of this study are available within the article. Raw data that support the finding of this study are available from the corresponding author, upon reasonable request.

CONFLICT OF INTEREST

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

ETHICS

There are no ethical issues with the publication of this manuscript.

STATEMENT ON THE USE OF ARTIFICIAL INTELLIGENCE

Artificial intelligence was not used in the preparation of the article.

REFERENCES

- [1] Sarah S, Valarmathie P, Rose LR, Jamila J, Sundarajan M, Choudhry MD. Multimodal emotion recognition with explainable AI for cognitive human-computer interaction in smart environments. In: 2025 5th International Conference on Soft Computing for Security Applications (ICSCSA). 2025. p. 1091–1096. [\[CrossRef\]](#)
- [2] Duan S, Wang Z, Wang S, Chen M, Zhang R. Emotion-aware interaction design in intelligent user interface using multi-modal deep learning. In: 2024 5th International Symposium on Computer Engineering and Intelligent Communications (ISCEIC). 2024. p. 110–114. [\[CrossRef\]](#)
- [3] Song X, Liu C, Xu L, Gao B, Lu Z, Zhang Y. Affective computing methods for multimodal embodied AI human-computer interaction. *Aslib J Inf Manag* 2025;1–25. [\[CrossRef\]](#)
- [4] Lyu Y, Liang PP, Deng Z, Salakhutdinov R, Morency LP. DIME: fine-grained interpretations of multimodal models via disentangled local explanations. In: Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society. 2022. p. 455–467. [\[CrossRef\]](#)
- [5] Mishra D, Mishra RK, Agarwal R. Human-computer interaction and user experience in the artificial intelligence era. *Trends Comput Sci Inf Technol Res* 2025;31–54.
- [6] Rajesh SG, Madangarli SV, Pisharady GS, Subrahmanyam R. Enhancement of virtual assistants through multimodal AI for emotion recognition. *IEEE Access* 2025.
- [7] Arumugam B, Bhattacharjee SD, Yuan J. Multimodal attentive learning for real-time explainable emotion recognition in conversations. In: 2022 IEEE International Symposium on Circuits and Systems (ISCAS). 2022. p. 1210–1214. [\[CrossRef\]](#)
- [8] Kamath D, Disha M, Tauro G. Deep learning for cognitive human-computer interaction: a comprehensive review of models, techniques, and applications. In: 2025 6th International Conference on Data Intelligence and Cognitive Informatics (ICDICI). 2025. p. 1011–1017. [\[CrossRef\]](#)
- [9] Neeraja G, Teja JSS, Ravi Kumar M, Prasanna LJ, Parvez MM, Santhosh C. Bridging the gap between AI and human emotion: a multimodal recognition system. *Control Cybern* 2024;53. [\[CrossRef\]](#)
- [10] Alonso-Moral JM, Anthis Z, Berlanga R, Catalá A, Cimiano P, Flach P, et al. Preface of the proceedings of the workshop on multimodal, affective and interactive explainable artificial intelligence (MAI-XAI), collocated with the European Conference on Artificial Intelligence (ECAI). 2024.
- [11] Khare SK, Blanes-Vidal V, Nadimi ES, Acharya UR. Emotion recognition and artificial intelligence: a systematic review (2014–2023) and research recommendations. *Inf Fusion* 2024;102:102019. [\[CrossRef\]](#)
- [12] Lian H, Lu C, Li S, Zhao Y, Tang C, Zong Y. A survey of deep learning-based multimodal emotion recognition: speech, text, and face. *Entropy (Basel)* 2023;25:1440. [\[CrossRef\]](#)
- [13] Geetha AV, Mala T, Priyanka D, Uma E. Multimodal emotion recognition with deep learning: advancements, challenges, and future directions. *Inf Fusion* 2024;105:102218. [\[CrossRef\]](#)
- [14] Chakravarthi B, Ng SC, Ezilarasan MR, Leung MF. EEG-based emotion recognition using hybrid CNN and LSTM classification. *Front Comput Neurosci* 2022;16:1019776. [\[CrossRef\]](#)
- [15] Alsabhan W. Human-computer interaction with a real-time speech emotion recognition system. *Sensors (Basel)* 2023;23:1386. [\[CrossRef\]](#)

- [16] Pham NT, Nguyen SD, Nguyen VST, Pham BNH, Minh DDN. Speech emotion recognition using overlapping sliding window and Shapley additive explainable deep neural network. *J Inf Telecommun* 2023. [\[CrossRef\]](#)
- [17] Blanco-Ríos MA, et al. Real-time EEG-based emotion recognition for valence, arousal and dominance estimation. *Sensors (Basel)* 2024.
- [18] Hosseini MSK, et al. Personality-based emotion recognition using EEG. *Front Comput Neurosci* 2023. [\[CrossRef\]](#)
- [19] Magherini R, et al. Emotion recognition in the times of COVID-19: applying Grad-CAM for visualization. 2022.
- [20] George SM, et al. A review on speech emotion recognition: a survey, recent trends. 2024.
- [21] Udaheureka G, et al. Multimodal emotion recognition using visual, vocal and physiological signals. *Appl Sci* 2024;14:8071. [\[CrossRef\]](#)
- [22] Gladys AA. Survey on multimodal approaches to emotion recognition. *Journal* 2023.
- [23] Exploring interpretability in deep learning for affective computing. *ACM Proceedings/Conference paper*. 2023-2024.
- [24] Wu Y. A comprehensive review of multimodal emotion recognition. *MDPI/Cells/PMC*. 2025. [\[CrossRef\]](#)
- [25] Miao M. ST-SHAP: a hierarchical and explainable attention network for emotional EEG recognition. *Journal* 2025. [\[CrossRef\]](#)
- [26] Ghafoor A, Norren S, Fatima A, Mahmoud HEAH. Cross-cultural emotion recognition in AI: enhancing multimodal NLP for empathetic interaction. *Soc Sci Spectr* 2025;4:575–588. [\[CrossRef\]](#)
- [27] Akhtar M, Mumtaz A, Zaman M, Ameer S, Fatima HS, Shahbaz N, et al. Emotionally intelligent AI: enhancing human-computer interaction through NLP and sentiment analysis. *Policy Res J* 2025;3:14. [\[CrossRef\]](#)
- [28] Kumar B, Rao AN, Yedukondalu N. EmoTrustBot: affective intelligence and trust-aware adaptation for emotionally aligned human-robot interaction. In: 2025 3rd International Conference on Sustainable Computing and Data Communication Systems (ICSCDS). 2025. p. 1602–1609. [\[CrossRef\]](#)
- [29] Kaur G, Sharma A. A deep learning-based model using hybrid feature extraction approach for consumer sentiment analysis. *J Big Data* 2023;10:5. [\[CrossRef\]](#)
- [30] Eva AF, Shomrat KH, Islam GA, Islam MD, Subarna J. Generative AI meets responsible AI and affective computing [doctoral dissertation]. *Brac University*; 2025. [\[CrossRef\]](#)
- [31] Roshdy A. Hybrid AI-based approach utilizing EEG-facial expression fusion for human-machine interaction [doctoral dissertation]. *Université Paris-Est Créteil Val-de-Marne-Paris 12*; 2025.
- [32] Järvinen S, Kallio J, Peltola J, Mäkelä SM. A behavior and emotion recognition framework for emotion-aware services in physical spaces. In: 2024 International Conference on Content-Based Multimedia Indexing (CBMI). 2024. p. 1–6. [\[CrossRef\]](#)
- [33] Mamodiya U, Kishor I, Syed AA, Sankalkar P, Naik N. An adaptive human-robot interaction framework using real-time emotion recognition and context-aware task planning. *IEEE Access* 2025. [\[CrossRef\]](#)
- [34] Yadav H, Maini S. The intersection of AI and emotions: deep learning applications for a better society. In: *International handbook of emotions: resourceful cultural perspectives*. Vol. 2. Cham: Springer Nature Switzerland; 2025. p. 11–29. [\[CrossRef\]](#)
- [35] Annam ST, Bodapati JD, Konda R. Emotion-aware music recommendations: a transfer learning approach using facial expressions. In: *International Conference on Data & Information Sciences*. 2023. p. 1–11. [\[CrossRef\]](#)
- [36] Negoitescu MC. Towards emotionally and motivationally aware intelligent systems: a systematic literature review [doctoral dissertation]. *Delft University of Technology*; 2025.
- [37] Faria DR, Godkin AL, da Silva Ayrosa PP. Advancing emotionally aware child-robot interaction with biophysical data and insight-driven affective computing. *Sensors (Basel)* 2025;25:1161. [\[CrossRef\]](#)
- [38] Šumak B, Brdник S, Pušnik M. Sensors and artificial intelligence methods and algorithms for human-computer intelligent interaction: a systematic mapping study. *Sensors (Basel)* 2021;22:20. [\[CrossRef\]](#)
- [39] Malk NA, Diwan SA. Artificial intelligence in speech emotion detection: trends, challenges, and future directions. *Int J Ethical AI Appl* 2025;1:19–29. [\[CrossRef\]](#)
- [40] Aswathy AA. A hybrid machine learning and deep learning approach for sentiment and emotion analysis in social media texts. *Sigma J Eng Nat Sci* 2026;in press.
- [41] Patil R. Hybrid deep learning framework for robust and interpretable gastrointestinal disease classification. *Sigma J Eng Nat Sci* 2026;in press. [\[CrossRef\]](#)
- [42] Kalkan MEK. The impact of visual perception-based intervention on hazard recognition in automotive manufacturing: a comparative study. *Sigma J Eng Nat Sci* 2026;in press.
- [43] Ren M, Huang X, Liu J, Gao Z, Su Y, Liu AA. KA-MIN: knowledge-aware multimodal interaction network for emotion recognition in conversation. *IEEE Trans Circuits Syst Video Technol* 2025. [\[CrossRef\]](#)
- [44] Zhao G, Liu X, Pietikäinen M. Facial emotion recognition using convolutional neural networks and support vector machines. *Pattern Recognit Lett* 2020;138:565–571.

- [45] Trigeorgis G, Ringeval F, Brueckner R, Marchi E, Nicolaou MA, Schuller B, et al. Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network. In: IEEE International Conference on Acoustics, Speech and Signal Processing. 2016. p. 5200–5204. [\[CrossRef\]](#)
- [46] Tzirakis P, Trigeorgis G, Nicolaou MA, Schuller B, Zafeiriou S. End-to-end multimodal emotion recognition using deep neural networks. IEEE J Sel Top Signal Process 2017;11:1301–1309. [\[CrossRef\]](#)
- [47] Zhang Y, Zhang Y, Guo Y. Multimodal emotion recognition based on CNN-RNN hybrid networks. Neural Comput Appl 2021;33:15015–15027.
- [48] Tsai YHH, Bai S, Liang PP, Kolter JZ, Morency LP, Salakhutdinov R. Multimodal transformer for unaligned multimodal language sequences. In: Proceedings of the Association for Computational Linguistics. 2019. p. 6558–6569. [\[CrossRef\]](#)
- [49] Mai S, Hu H, Zhang Z. Cross-modal transformer for multimodal emotion recognition. IEEE Trans Affect Comput 2022;13:1463–1475.
- [50] Atrey PK, Hossain MA, El Saddik A, Kankanhalli MS. Multimodal explainable emotion recognition using feature attribution. Inf Fusion 2020;62:1–12.
- [51] Liew WS, Loo CK, Wermter S. Emotion recognition using explainable genetically optimized fuzzy ART ensembles. IEEE Access 2021;9:61513–61531. [\[CrossRef\]](#)
- [52] Alabi AO, Song BG, Kang N. A comprehensive investigation of auditory-induced emotions combining psychoacoustic metrics, facial emotion recognition, and explainable deep learning.
- [53] Lu J, Liu Y, Lv T, Meng L. An emotional-aware mobile terminal accessibility-assisted recommendation system for the elderly based on haptic recognition. Int J Hum Comput Interact 2024;40:7593–7609. [\[CrossRef\]](#)
- [54] Razzaq MA, Hussain J, Bang J, Hua CH, Satti FA, Rehman UU, et al. A hybrid multimodal emotion recognition framework for UX evaluation using generalized mixture functions. Sensors (Basel) 2023;23:4373. [\[CrossRef\]](#)