Journal of Innovative Image Processing (ISSN: 2582-4252)
https://irojournals.com/iroiip/

Interpretable Deep Multimodal Fusion of ECG and Facial Expressions for Robust Emotion Detection

Sankar Ganesh Karuppasamy[1], Rajeev Kumar M.[2]

[1]Assistant Professor, [2]Professor, Department of Computer Science and Engineering, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Chennai, India.

E-mail: [1]vtd1108@veltech.edu.in, [2]drmrajeevkumar@veltech.edu.in

Orcid ID: [1]0000-0002-4307-5294, [2]0000-0002-6122-2366

Abstract

In healthcare, emotional computing, and dynamic human-computer interaction, emotion recognition is a key component of human-centric Artificial Intelligence (AI). However, existing approaches depend on either physiological or visual signals and often show limited robustness or accuracy in noisy or inadequate data environments. This study proposes a task-specific bidirectional cross-modal fusion framework in which physiological ECG representations and facial-expression representations mutually attend to each other, while an auxiliary embedding-alignment objective constrains the two modalities to a shared latent space. While the facial images are processed using EfficientNet-B0 to extract discriminative visual features, the ECG signals are transformed using a one-dimensional Convolutional Neural Network (1D-CNN) in conjunction with a Bidirectional Long Short-Term Memory (Bi-LSTM) network to extract local morphological features and long-range temporal inference. A cross-attention fusion technique is then used to fuse the trained ECG models with facial expression data. The inter-modal dependencies between physiological and behavioral signals are represented by a cross-attention fusion module, which improves the model's dependability and broadens its capacity to identify a variety of emotional classes, specifically focusing on six fundamental emotions: happy, sad, angry, fearful, surprised, and neutral. The acquired representation is classified using a fully connected attention-based decoder. Model interpretability is attained through SHAP analysis of ECG embeddings and Grad-CAM visualization of the facial areas. When compared to current and baseline fusion techniques, the evaluation conducted using the MAHNOB-HCI and AMIGOS datasets demonstrated that the proposed model achieved higher performance, including an accuracy of 97%, precision of 96%, recall of 97%, F1-score of 96%, and improved adaptability.

Keywords: Emotion Recognition, Electrocardiogram (ECG), Facial Expression Analysis, Cross-Attention Mechanism, Multimodal Fusion, SHAP, Grad-CAM.

1. Introduction

Emotion recognition is the foundation of human-centric intelligent machines, enabling computational systems to detect, understand, and efficiently deal with human emotional classes, such as happy, sad, anger, fear, surprise, and neutral, which are the focus of this study

DOI: https://doi.org/10.36548/jiip.2026.4.010
Received: 10.08.2026, received in revised form: 13.09.2026, accepted: 05.10.2026, published: 17.10.2026
© 2026 Inventive Research Organization. This is an open access article under the Creative Commons Attribution-NonCommercial International (CC BY-NC 4.0) License

[1]. Most existing emotion identification algorithms rely on unimodal input, including voice, facial movement, and physiological signs [2]. Facial expressions act as an impulsive way to distinguish emotional classes; however, their interpretation is often hindered by differences in lighting, posture, and other obstacles. Physiological signals, such as electrocardiograms (ECG), on the other hand, provide an unbiased perception into how the neurological nervous system responds to emotional incentives; however, they do not directly react to emotion. Identifying the limitations of unimodal systems [3] has been developed as a practical method that combines complementary data from multiple sources to enhance resilience and predictive ability. Recent developments in transformer architectures [4] and cross-attention mechanisms [5] have improved the ability to model complex relationships. However, even with these developments, the interpretability of multimodal emotion recognition systems that identify the functions of both physiological and visual signals remains difficult. In particular, cross-attention mechanisms allow models to concentrate on the relevant areas of one modality based on information from another [6].

Although multimodal emotion-recognition methods increasingly use CNNs, recurrent networks, attention, and Transformers, most fusion approaches simply concatenate modality features or combine independent predictions. They do not explicitly capture the reciprocal interaction between physiological and facial cues. NeuroSym-CRXNet addresses this limitation through bidirectional cross-modal interaction, where ECG and facial features alternately serve as queries and key–value representations, allowing each modality to dynamically condition the other. A joint cross-modal alignment objective further maps both modalities into a compatible latent space. Thus, the key contribution is the proposed reciprocal cross-modal interaction and alignment strategy, rather than a simple integration of existing networks.

The novelty of the proposed framework is not attributed to the individual backbone networks or post-hoc explainability algorithms, which are established techniques. Instead, the contribution lies in their task-specific organization around a bidirectional ECG–facial interaction mechanism and a joint modality-alignment objective. Unlike conventional multimodal fusion in which modality representations are concatenated or independently classified, the proposed architecture enables reciprocal cross-modal conditioning, allowing physiological information to influence visual feature selection and visual information to influence physiological feature representation. The alignment objective further constrains the modality projections toward a shared latent representation, thereby complementing the cross-attention mechanism during optimization.

The main contributions of this study are as follows:

1. A bidirectional ECG–facial cross-attention mechanism in which ECG and facial representations serve as query and key–value representations, enabling reciprocal information exchange rather than conventional feature concatenation or independent decision fusion.
2. A cross-modal alignment objective incorporated with the classification objective to reduce the representation discrepancy between ECG and facial embeddings and encourage a compatible shared latent representation.
3. The modality-specific encoders and temporal attention modules are organized around the proposed reciprocal interaction mechanism, allowing physiological and visual

information to dynamically influence one another before classification.

4. The fusion strategy is validated through comparisons with unimodal models, feature concatenation, late fusion, and ablated versions of the proposed cross-attention and alignment mechanisms.
5. SHAP and Grad-CAM provide physiological and facial evidence contributing to the final prediction.

2. Related Work

2.1 ECG-based Emotion Recognition

Heart-Rate Variability (HRV) features derived from datasets like WESAD, DREAMER, and ASCERTAIN are used in recent advances in ECG-based emotion detection, including deep learning techniques like a deep CNN incorporating spatial attention for multi-emotion classification. Concurrently, systems like the noncontagious ECG-based system have proven more traditional machine learning techniques [7], such as SVM and Random Forest classifiers. Although ECG-based methods have demonstrated promise [8], they frequently encounter issues such as individual discrepancy, comparatively small labelled datasets, and a shortage of behavioural signals. Due to the limitations of depending solely on ECG-based techniques, it is vital to integrate additional models. Because of its close connection to emotional expression, Facial Expression Recognition (FER) stands out among the other behavioral signals as one of the most advanced and extensively researched models for emotion detection.

2.2 Facial Expression-based Emotion Recognition

Current investigations have shown the development of deep learning architectures applied to FER and identified major issues such as brightness, posture distinction, individuality bias, and overfitting from limited data [9]. Recently, [10] reviewed deep-learning FER systems and their challenges. The study in [11] shows that while facial signals provide eloquence, they are delicate under visual conditions and may be deficient in physiological fixation.

2.3 Multimodal Fusion for Emotion Recognition

To overcome the constraints of unimodal approaches, such as physiological or behavioural alone, researchers have progressively accepted combining signals such as ECG/EEG, facial features, speech, and body signals. A real-time embedded system for emotion identification was demonstrated that combined face vision with EEG, ECG, and EMG inputs [12]. Prior research has carried out multimodal emotion identification using deep learning and fusion techniques [12]. A recent end-to-end model combined facial expressions with remote PPG signals via a transformer-based cross-modal attention framework, reporting improved accuracy for valence/arousal classification [13]. These studies show that the fusion of behavioural and physiological models can improve performance and strength. However, many existing fusion methods still depend on relatively lightweight concatenation or simple late fusion and deficit deep modelling of inter-modal interaction or joint learning with explanations.

2.4 Explainable Models in Emotion Recognition

Recent studies on ECG-based emotion recognition have included attention modules to highlight relevant ECG channels and temporal segments [14]. In this domain, few studies have deeply integrated interpretable mechanisms for multimodal emotion recognition [15]. In the

proposed work, the combination of SHAP (for physiological input) and Grad-CAM (for visual input) aims to bridge this gap by providing method interpretability in a deep fusion architecture.

Existing emotion-recognition studies differ substantially in their choice of modality, feature representation, fusion strategy, and degree of cross-modal interaction. ECG-only approaches can capture physiological responses and temporal variations but may be affected by inter-subject variability and the limited direct relationship between cardiac activity and externally expressed emotion. Facial-expression approaches provide more direct behavioural evidence but are sensitive to illumination, pose, occlusion, and subject-specific appearance. Multimodal approaches reduce some of these limitations by combining complementary information; however, several existing methods rely primarily on feature concatenation or late decision fusion, which may not explicitly model the temporal and semantic relationships between modalities. Transformer-based cross-modal approaches provide stronger interaction modelling, but their additional complexity can increase computational requirements, while multimodal interpretability remains comparatively limited. Therefore, the distinction of the proposed framework is not the use of individual CNN or recurrent backbones, but the reciprocal ECG–facial cross-modal interaction combined with an alignment objective that encourages compatible multimodal representations. This comparison motivates the need for evaluating whether bidirectional interaction provides measurable benefits over unimodal, concatenation-based, and late-fusion alternatives.

3. Dataset

Two publicly accessible datasets, MAHNOB-HCI [16] and AMIGOS [17], were used to assess the performance of the proposed model. The model considered both front-facing face recordings and ECG recordings. A variety of human emotions, such as happiness, sadness, fear, surprise, rage, and neutral emotions, are recorded in the MAHNOB-HCI dataset. The ECG signals available in the datasets are captured at an acquisition rate of 256 Hz using the Bio-Semi Active Two acquisition system, which provides recordings suitable for HRV analysis and ECG waveform analysis. In parallel, the front face videos were recorded at 25 frames per second, which helps to perform comprehensive analysis of facial features during each trial. The labelling for the emotions from every participant was captured using the Self-Assessment Manikin (SAM), a self-assessment sheet completed by each participant once the experiment was over. The AMIGOS dataset [17] includes multimodal recordings obtained during individual and collective-based experimental sessions. 50 participants viewed emotionally stimulating video clips individually and participated in longer group-viewing sessions designed to evoke diverse emotional responses. Wearable sensors attached to the chest were used to collect ECG signals at a sample frequency of 60 Hz. Table 1 shows the distribution of the emotion samples utilized in this investigation.

Table 1. Dataset distribution across emotion

Emotion	Members	ECG Recordings	Facial Video Segments	Total Samples
Neutral	50	150	150	300
Happy	48	144	144	288
Sad	47	141	141	282
Anger	45	135	135	270
Fear	46	138	138	276
Surprise	44	132	132	264

Total	50	840	840	1680

As shown in Table 2, the emotion labels from MAHNOB-HCI and AMIGOS were harmonized using the valence and arousal ratings obtained through the Self-Assessment Manikin (SAM). The original valence and arousal scores were normalized to a common scale and categorized into low, medium, and high levels using predefined threshold intervals. Scores ≤ 3 were considered low, scores between 4 and 6 were considered medium, and scores ≥ 7 were considered high. Each sample was then assigned to one of six emotion classes according to its corresponding valence–arousal category. High-valence/high-arousal samples were assigned to Happy, high-valence/medium-to-high-arousal samples to Surprise, medium-valence/medium-arousal samples to Neutral, low-valence/low-arousal samples to Sad, low-valence/high-arousal samples to Fear, and high- or stimulus-specific-valence/high-arousal samples to Anger. The same label-harmonization procedure was applied consistently to both datasets before temporal synchronization, segmentation, and model training.

Table 2. Explicit valence–arousal criteria for six-class emotion label harmonization

Valence range	Arousal range	Emotion	Decision rule
High	High	Happy	Positive, high-arousal response
High	Medium–High	Surprise	Distinctive high-arousal surprise response
Medium	Medium	Neutral	Moderate valence and arousal
Low	Low	Sad	Negative, low-arousal response
Low	High	Fear	Negative, high-arousal response
High/Stimulus-specific	High	Anger	High-arousal anger-related response

ECG signals from MAHNOB-HCI (256 Hz) and AMIGOS (60 Hz) were resampled to a common sampling rate after signal-quality preprocessing. ECG and facial streams were synchronized using trial timestamps and boundaries, retaining only their common temporal intervals and generating 5–10 s paired windows. Band-pass filtering (0.5–40 Hz) was applied to ECG data, Pan–Tompkins was used to identify R-peaks, and participant-wise z-score normalization was used. RGB conversion, MTCNN detection and alignment, cropping, resizing to 224 × 224 pixels, and quality filtering were applied to facial frames.

4. Methodology

4.1 ECG Preprocessing

The ECG signals from MAHNOB-HCI (256 Hz) and AMIGOS (60 Hz) were independently pre-processed and resampled to a common sampling frequency before segmentation. After reducing baseline drift, motion artifacts, and high-frequency noise with a 0.5–40 Hz Butterworth band-pass filter, the Pan–Tompkins method was used to detect R-peaks. The ECG recordings were synchronized with the corresponding facial videos using trial timestamps, and only temporally overlapping intervals were retained. The synchronized signals were divided into 5–10 s windows, with facial frames from the corresponding intervals paired with each ECG segment. Each ECG window was standardized using z-score normalization to reduce inter-subject physiological variability.

4.2 Deep ECG Encoder

Let $x(t) \in \mathbb{R}^L$denote a preprocessed ECG sequence of length L. The encoder begins with a stack of 1D-CNN layers, which are responsible for extracting hierarchical

representations from the input signal. While preserving the chronological order of the ECG signal, convolutional layers progressively collect local waveform features, such as the P-wave, QRS complex, and T-wave. Equation (1) provides a description of a typical convolution operation.

$$h_k(t) = (x * w_k)(t) = \sum_{\tau=0}^{K-1} x(t-\tau)\, w_k(\tau) \tag{1}$$

where $h_k(t)$ is the feature that follows, and $w_k\ (t)$. τ is the position/index within the convolution kernel (k^{th} filter). The convolutional output is forwarded to a BiLSTM [18]. The LSTM uses equation (2) to calculate for each time step t**.**

$$\begin{aligned} i_t &= \sigma(W_i x_t + U_i h_{t-1} + b_i), \\ f_t &= \sigma(W_f x_t + U_f h_{t-1} + b_f), \\ o_t &= \sigma(W_o x_t + U_o h_{t-1} + b_o), \\ \tilde{c}_t &= \tanh(W_c x_t + U_c h_{t-1} + b_c), \\ c_t &= f_t \odot c_{t-1} + i_t \odot \tilde{c}_t, \\ h_t &= o_t \odot \tanh(c_t), \end{aligned} \tag{2}$$

where x_t is the input feature vector at time step t, c_t and h_t stand the memory cell and hidden state, respectively, and i_t, f_t, o_t indicate the input, forget, and output gates, respectively, h_{t-1}represents the previous hidden state. $b_i,\ b_f, b_o, b_c$ are the learnable bias vectors. $W_i,\ W_f, W_o, W_c$ are the learnable input-weight matrices. σ is the sigmoid activation function. tanh is the Hyperbolic tangent activation. t is the temporal index The encoder combines the forward and backward hidden states as $H_t = [h_t^{\rightarrow}; h_t^{\leftarrow}]$, using a bidirectional structure providing an improved temporal illustration that integrates both past and future viewpoints within each ECG window.

Given hidden states H_t, in a W learnable attention projection matrix, with a learnable attention vector v, the attention weight $\alpha_t$is computed as shown in equation (3). Here, j is the index over temporal positions and $\top$is the total number of temporal positions.

$$\alpha_t = \frac{\exp\left(v^{\top} \tanh(W H_t)\right)}{\sum_j \exp(v^{\top} \tanh(W H_j))} \tag{3}$$

The ECG embedding is reached as the weighted sum of $z_{\text{ECG}} = \sum_t \alpha_t H_t$. This learned scheme $z_{\text{ECG}} \in \mathbb{R}^d$attains both short-term wavelength morphology and long-term heart-rate factors, making it compatible for facial expression features.

4.3 Facial Preprocessing

All the images (video frames) are converted to RGB colour space and later modified to a uniform resolution across all participants. Facial regions are captured by the Multi-task Cascaded Convolutional Neural Network (MTCNN) [19], which pinpoints the facial regions and extracts them from each frame. To adhere to the input criteria of the extraction network, the normalized face area was clipped to preserve the most important regions associated with emotional manifestations and reduced to 224×224 pixels. Pixel intensities were standardized using z-score normalization, which transforms each pixel value according to equation (4).

$$I'(x,y) = \frac{I(x,y) - \mu}{\sigma} \tag{4}$$

where and σ are and standard deviation measured throughout the frame, and $I'(x, y)$ is the pixel value at position $I(x, y)$.

Histogram equalization [22] is applied for datasets with lighting inconsistency. Frames showing blurred facial areas, motion artifacts, or shrinkages are identified using a sharpness metric based on Laplacian variance [23] and are omitted if they fall below a boundary. For temporal multimodal modelling, the preprocessed facial frames were synchronized with the corresponding ECG windows using the recording timestamps and trial boundaries. For each ECG window, facial frames falling within the same temporal interval were selected, thereby ensuring that the ECG and facial modalities represented the same portion of the emotional stimulus.

4.4 Deep Facial Encoder

The preprocessed facial frames are processed using a pretrained EfficientNet-B0 backbone to extract spatial facial features from each frame. Each video segment is represented as an ordered sequence of aligned facial frames. EfficientNet-B0 captures spatial characteristics associated with facial expressions, including eye movements, mouth movements, and other facial muscle changes. The final global average pooling layer generates a compact frame-level feature representation. The sequence of frame-level features is subsequently provided to a bidirectional Gated Recurrent Unit (Bi-GRU) to capture temporal dependencies across consecutive facial frames. Bi-GRU hidden states are then subjected to an attention mechanism, which highlights the frames that have the most impact on emotion detection. The weighted temporal representation is used as the final facial embedding, which is subsequently provided to the bidirectional cross-attention fusion module. The output from each convolutional block is passed through a series of operations, which include convolution, batch normalization, and nonlinear activation, formally represented as shown in equation (5).

$$X_l = \sigma(\text{BN}(W_l * X_{l-1} + b_l)) \tag{5}$$

where X_l is the next feature, $X_{l-1}$is the feature map from the previous CNN layer $\sigma(\cdot)$ is the ReLU activation function, and W_l stands for the filters at layer l, $b_l$is the learnable bias at layer l, $BN(\bullet)$ is the batch normalization operation.

A frame-level representation is created by the feature vector from the CNN's final global average pooling layer, represented as $f_t \in \mathbb{R}^d$. These vectors are then fed into the bidirectional Gated Recurrent Unit (Bi-GRU), which represents the relationships between frames.

For each time step t, the GRU calculates its hidden state according to equation (6).

$$\begin{aligned} z_t &= \sigma(W_z f_t + U_z h_{t-1}) \\ r_t &= \sigma(W_r f_t + U_r h_{t-1}) \\ \tilde{h}_t &= \tanh(W_h f_t + U_h(r_t \odot h_{t-1})) \\ h_t &= (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t \end{aligned} \tag{6}$$

where h_t is the hidden state that the temporal development of the face expressions, and z_t and r_t represent the update and reset gates, f_t is the Facial feature vector at time t, $r_t$is the GRU reset gate, h_t is the GRU hidden state, h_{t-1} is the previous hidden state, W_z, W_r, W_h are the learnable input-to-hidden weight matrices, and U_z, U_r, U_h are the learnable hidden-to-hidden weight matrices,

The main emotional frames are highlighted by an attention mechanism on a series of hidden states. Given temporal features $\{h_1, h_2, \ldots, h_T\}$, attention weights are computed as shown in equation (7).

$$\alpha_t = \frac{\exp(v^\top \tanh(W h_t))}{\sum_{j=1}^{T} \exp(v^\top \tanh(W h_j))} \tag{7}$$

Equation (8) illustrates how the final face is created as a weighted representation of every temporal condition.

$$z_{\text{face}} = \sum_{t=1}^{T} \alpha_t h_t \tag{8}$$

This embedding $z_{\text{face}} \in \mathbb{R}^d$is successfully preserves both static spatial signals and dynamic temporal variations in facial expressions, making it very rigorous for emotion recognition tasks.

4.5 Multimodal Fusion Module

Let T_E and T_F be the numbers of ECG and facial feature tokens, respectively, and d_E and d_F be the corresponding feature dimensions. Let $E \in \mathbb{R}^{T_E \times d_E}$ represent the ECG feature sequence obtained from the ECG encoder and $F \in \mathbb{R}^{T_F \times d_F}$ represent the facial feature sequence obtained from the facial encoder. Equation (9) shows how the two modalities are initially projected onto a shared latent space of size 256 since their feature dimensions and representations differ.

$$E_p = EW_E + b_E, F_p = FW_F + b_F \tag{9}$$

where E is the ECG feature-token sequence, F is the facial feature-token sequence, b_E and b_F are vectors, and $W_E \in \mathbb{R}^{d_E \times 256}$ and $W_F \in \mathbb{R}^{d_F \times 256}$ are learnable projection matrices. Consequently, $E_p \in \mathbb{R}^{d_F \times 256}$ and $F_p \in \mathbb{R}^{T_F \times 256}$.

4.6 ECG-to-Facial Cross-Attention

In the first direction, the ECG representation acts as the query, while the facial representation provides the key and value. The proposed module employs eight attention heads. For the h^{th} attention head, the query, key, and value representations are computed as shown in equation (10).

$$Q_E^{(h)} = E_p W_Q^{(h)},\ K_F^{(h)} = F_p W_K^{(h)},\ V_F^{(h)} = F_p W_V^{(h)} \tag{10}$$

where h is the attention-head index, learnable projection matrices are $W_Q^{(h)}$, $W_K^{(h)}$, and $W_V^{(h)}$. With eight attention heads and a common embedding dimension of 256, each head's dimension is $d_h = \frac{256}{8} = 32$.

Equation (11) illustrates how scaled dot-product is used to determine the ECG-to-facial attention weights.

$$A_{E \to F}^{(h)} = softmax\left(\frac{Q_E^{(h)} \left(K_F^{(h)}\right)^T}{\sqrt{d_h}}\right) \tag{11}$$

where $A_{E\to F}^{(h)}$ represents the attention weights that determine the relevance of facial features with respect to the ECG representation, $Q_E^{(h)}$is the ECG query matrix, $\left(K_F^{(h)}\right)$ is the facial key matrix, d_h is the key/query dimensionality per attention head. The ECG-conditioned facial representation is then obtained as shown in equation (12).

$$Z_{E\to F}^{(h)} = A_{E\to F}^{(h)} V_F^{(h)} \tag{12}$$

where, $Z_{E\to F}^{(h)}$ is the ECG-conditioned facial representation produced by head

Equation (13) illustrates how the outputs eight attention heads are concatenated and projected linearly.

$$Z_{E\to F} = Concat\left(Z_{E\to F}^{(1)}, \ldots.., Z_{E\to F}^{(8)}\right) W_O^{E\to F} \tag{13}$$

where $Z_{E\to F}$ is the final ECG-to-facial multi-head representation, $W_O^{E\to F}$ is the learnable output projection matrix

4.7 Facial-to-ECG Cross-Attention

To enable reciprocal information exchange, the cross-attention operation is performed in the reverse direction. In this branch, the facial representation acts as the query, while the ECG representation provides the key and value as given in equation (14).

$$Q_F^{(h)} = F_p W_{Q,F}^{(h)},\ K_E^{(h)} = E_p W_{K,E}^{(h)},\ V_E^{(h)} = E_p W_{V,E}^{(h)} \tag{14}$$

where, $Q_F^{(h)}$ is the facial query representation, $K_E^{(h)}$ is the ECG key representation, $V_E^{(h)}$ is the ECG value representation, $W_{Q,F}^{(h)}$ is the facial query projection matrix, $W_{K,E}^{(h)}$ is the ECG key projection matrix, and $W_{V,E}^{(h)}$ is the ECG value projection matrix.

The corresponding facial-to-ECG attention weights are calculated as shown in equation (15).

$$A_{F\to E}^{(h)} = softmax\left(\frac{Q_F^{(h)}\left(K_E^{(h)}\right)^T}{\sqrt{d_h}}\right) \tag{15}$$

where, $A_{F\to E}^{(h)}$ is the facial-to-ECG attention weights, $Q_F^{(h)}$ is the facial query, $K_E^{(h)}$is the ECG key, and d_h is the attention-head dimension.

The facial-conditioned ECG representation is obtained as shown in equation (16).

$$Z_{F\to E}^{(h)} = A_{F\to E}^{(h)} V_E^{(h)} \tag{16}$$

The outputs from the eight heads are concatenated and projected to obtain the following.

$$Z_{F\to E} = Concat\left(Z_{F\to E}^{(1)}, \ldots., Z_{F\to E}^{(8)}\right) W_O^{F\to E} \tag{17}$$

where, $Z_{F\to E}$ is the final facial-to-ECG multi-head representation, $W_O^{F\to E}$ is the learnable output projection matrix.

4.8 Final Multimodal Fusion

The two cross-attended representations are subsequently concatenated and transformed into the final multimodal representation.

$$Z_{fused} = \sigma([Z_{E\to F}||Z_{F\to E}]W_f + b_f) \tag{18}$$

where W_f is a learnable fusion b_f is the bias vector, $\sigma(\bullet)$ is the nonlinear activation function, and $\|$ stands for feature concatenation, and Z_{fused} is the intermediate fused representation. After being projected to a 256-dimensional feature vector, the fused representation is then fed into the classification layer in the following manner.

$$Z_{final} \in \mathbb{R}^{256} \tag{19}$$

where, Z_{final} is the final 256-dimensional multimodal representation.

The complete fusion process can therefore be summarized as shown in equation (20).

$$(E,F) \to (E_p,\ F_p) \to \begin{cases} E \to F : Q_E,\ K_F,\ V_F \\ F \to E : Q_F,\ K_F,\ V_F \end{cases} \to (Z_{E\to F},\ Z_{F\to E}) \to Z_{final} \tag{20}$$

where, E is the ECG feature sequence, F is the facial feature sequence, E_p is the projected ECG sequence, F_p is the projected facial sequence, $Z_{E\to F}$ is the ECG-conditioned facial representation, $Z_{F\to E}$ is the facial-conditioned ECG representation, and Z_{final} is the final fused representation.

The architectural distinction of the proposed framework is the reciprocal nature of the cross-modal interaction. In the ECG-to-facial pathway, ECG representations act as queries while facial representations provide the key–value context, allowing physiological information to selectively modulate visual features. Conversely, in the facial-to-ECG pathway, facial representations act as queries while ECG representations provide the key–value context. The resulting reciprocal conditioning enables the two modalities to refine one another before classification. This differs fundamentally from conventional early or late fusion, where modality-specific representations remain independently learned until a subsequent concatenation or decision-combination stage.

4.9 Classification Layer and Loss Function

The multimodal fusion vector z_{fusion} acts as the input for the final classification module, which predicts discrete emotional classes. A fully connected neural layer and region are utilized to convert this high-dimensional representation to class probabilities. Precisely, the expected probability distribution over C emotion classes is computed as shown in equation (21).

$$\hat{y} = \text{softmax}\,(W_o z_{\text{fusion}} + b_o) \tag{21}$$

where d_f represents the dimension of the feature vector and $W_o \in \mathbb{R}^{C\times d_f}$ and $b_o \in \mathbb{R}^C$ are learnable parameters. The output logit is converted into normalized probabilities by the softmax function in the manner described below:

$$\hat{y}_c = \frac{\exp(s_c)}{\sum_{j=1}^{C}\exp(s_j)}, s = W_o z_{\text{fusion}} + b_o \tag{22}$$

where s_c is the logit corresponding to class, s is the vector of class logits, j is the summation/class index, C is the total number of emotion classes.

The function for improving the model is the cross-entropy loss, which measures the discrepancy between the anticipated and real emotion labels. For a given training sample with a ground-truth label y, the loss is defined as shown in equation (23).

$$\mathcal{L}_{\text{CE}} = -\sum_{c=1}^{C} y_c \log(\hat{y}_c) \tag{23}$$

where $\mathcal{L}_{\text{CE}}$ is the cross-entropy classification loss. The ground-truth indicator for class is $y_c = 1$ for the correct class and 0 otherwise. $\hat{y}$ is the predicted probability for class. The dropout and decay technique are used to increase reliability and minimize overfitting.

Label smoothing is used to decrease the certainty of the classifier. The modified loss with the label smoothing parameter ϵ is shown in equation (24).

$$\mathcal{L}_{\text{LS}} = -\sum_{c=1}^{C} [(1-\epsilon)y_c + \frac{\epsilon}{C}] \log(\hat{y}_c) \tag{24}$$

where, $\mathcal{L}_{\text{LS}}$ is the label-smoothed cross-entropy loss, ϵ is the label-smoothing parameter.

Let z_{ECG} and z_{Face} denote the projected feature embeddings generated by the ECG encoder and facial encoder, respectively, after the cross-attention fusion stage. Equation (25) displays the alignment loss the mean squared Euclidean distance between the two embeddings.

$$L_{\text{align}} = \frac{1}{N} \sum_{i=1}^{N} \| z_{ECG}^{(i)} - z_{Face}^{(i)} \|_2^2 \tag{25}$$

where N represents the batch size, and $z_{ECG}^{(i)}$ and $z_{Face}^{(i)}$ correspond to the ECG and facial feature embeddings of the i^{th} training sample. The process becomes more stable and enhances improved emotion detection performance. The final training objective is stated as equation (26).

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda \mathcal{L}_{\text{align}} \tag{26}$$

where λ controls the relative importance of the alignment loss during optimization.

The alignment loss is not used as an independent post-processing operation. It is integrated directly into the optimization of the multimodal architecture and therefore influences the learned representations during training. Cross-attention determines how information is exchanged between modalities, whereas the alignment objective constrains the resulting representations to remain compatible in the shared latent space.

Figure 1 illustrates the proposed multimodal emotion recognition framework integrating ECG (physiological) and facial (visual) modalities. The ECG branch uses 1D-CNN, Bi-LSTM, and temporal attention to extract physiological features, while the facial branch employs EfficientNet-B0, Bi-GRU, and temporal attention to capture visual features. The extracted embeddings are projected into a shared feature space and processed through bidirectional cross-modal attention to learn complementary relationships between ECG and facial representations. A cross-modal alignment loss ensures feature consistency between the two

modalities. The attended features are fused and passed through fully connected layers and a Softmax classifier to predict six emotion classes. SHAP and Grad-CAM are incorporated to provide interpretability for ECG-based and facial-based predictions, respectively.

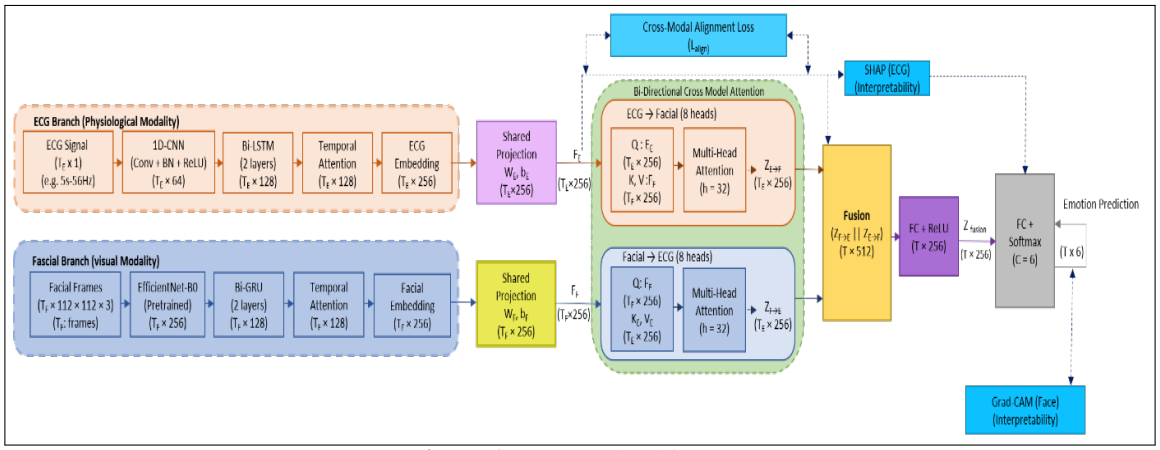

Figure 1. Proposed architecture

5. Experimental Setup

An NVIDIA RTX 4090 GPU with 24 GB of RAM was used to train the proposed multimodal framework across 150 epochs with a batch size of 32 samples. The AdamW optimizer with an initial learning rate of 1×10^{-4} and a weight decay of 1×10^{-5} produced the final reported results. During training, a cosine learning-rate scheduler was used. These hyperparameters were fixed across all subject-independent cross-validation folds to ensure reproducibility. Any alternative optimizer, batch-size, or regularization settings discussed elsewhere were considered only in ablation experiments and were not used for the final reported results.

5.1 Training Protocol

The proposed framework was trained using a standardized methodology to guarantee robust optimization and reliable performance across both the ECG and facial channels. The AdamW optimizer was used to optimize the model parameters with 1×10^{-4}, $\beta_1 = 0.9$, and $\beta_2 = 0.999$, and an initial learning rate of 1×10^{-4}. A cosine annealing learning rate scheduler is used to modify the model's learning rate in order to improve optimization. The model was trained for 100–150 epochs, allowing for early termination, given the size of the dataset. Throughout the experiment, a batch size of 32 was employed to maintain training consistency and execution efficiency.

5.2 Computational Complexity and Deployment Analysis

The ECG encoder comprises three one-dimensional convolutional layers accompanied by a bidirectional LSTM with 128 hidden units. The total computational complexity of the temporal encoder is roughly $O(T \times d^2)$, where Tdenotes the length of the ECG sequence and d represents the latent feature dimension. A pretrained EfficientNet-B0 backbone, global average pooling, a bidirectional GRU (Bi-GRU), and a temporal attention mechanism create the face encoder. Feature integration is performed using a bidirectional cross-attention module operating in a shared embedding space of 256 dimensions with eight attention heads. The computational complexity of the attention mechanism is approximately $O(N^2 d)$, where

Nrepresents the number of multimodal feature tokens and $d$is the embedding dimension.

5.3 Evaluation Metrics

A series of criteria was applied to measure both its class-specific adaptability and total predictive accurateness. To assess the flexibility of the model, we conducted trials in an acoustic environment. This involved introducing random levels of Gaussian noise (SNR from 10 dB to 20 dB) into the ECG signals and random obstacles (up to 20% of the face region) in facial video frames. In addition, the model's performance on subsections of data with changes in illumination (from AMIGOS) was assessed to control its flexibility for standard visual problems.

Table 3. Detailed architectural specifications of the bidirectional cross-attention fusion module

Parameter	Specification
Shared embedding dimension	256
Number of attention heads	8
Dimension per attention head	32
Attention mechanism	Scaled dot-product multi-head cross-attention
ECG-to-Face	ECG = Query; Face = Key, Value
Face-to-ECG	Face = Query; ECG = Key, Value
Head outputs	Concatenation followed by linear projection
Fusion operation	Bidirectional representation concatenation
Activation	ReLU
Final fused representation	256 dimensions

Table 3 summarizes the key architectural parameters of the proposed bidirectional cross-attention fusion module, including the 256-dimensional shared embedding space, eight attention heads, 32-dimensional head size, and reciprocal ECG-to-face and face-to-ECG query–key–value interactions. To evaluate the effectiveness of the proposed model the following performance metrics are used.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{27}$$

$$Precision = \frac{TP}{TP+FP} \tag{28}$$

$$Recall = \frac{TP}{TP+FN} \tag{29}$$

$$F1\ Score = \frac{2TP}{2TP+FP+FN} \tag{30}$$

Each test sample's Pointing Game Accuracy is determined by identifying the Grad-CAM map's maximal activation point. If this point falls inside the predefined facial ROI, it is considered a hit.

$$PGA = \frac{N_{hit}}{N_{total}} \times 100 \tag{31}$$

where, $N_{hit}$is the number of test samples in which the maximum Grad-CAM activation lies inside the facial ROI and N_{total} is the total number of evaluated test samples.

SHAP Consistency Score indicating stable hierarchical ranking of ECG features across

consecutive experiments. If the consistency was calculated using rank agreement, a suitable formulation as shown in equation (32).

$$C_{SHAP} = \frac{1}{K(K-1)} \sum_{i=1}^{K} \sum_{\substack{j=1 \\ j \neq i}}^{K} \rho(r_i, r_j) \tag{32}$$

where, K is the number of repeated experiments/folds, r_i= ranked SHAP feature-importance vector from experiment i, and $\rho(r_i, r_j)$= rank correlation between the SHAP feature rankings from experiments i and j.

5.4 Baseline Models

The proposed model is compared to many standard baseline models. The selected models were capable of identifying ECG signals and facial expressions; they were trained using the same pipeline, which included the same split ratios, preprocessing procedures, training parameters, and standardized optimization settings.

5.4.1 ECG-Only Baselines

Emotion detection models based on ECG signals generally classify emotions using deep learning models, which are capable of capturing the temporal representation from the signals. To evaluate the performance of the proposed method on emotion detection from ECG signals, two baseline methods were used namely one-dimensional convolutional neural network (1D-CNN) [20]. Second, a classifier based on LSTM was developed to attain long-term temporal patterns in heart-rate variations.

5.4.2 Facial Expression Baselines

Two independent facial-expression baselines were considered for comparison with the proposed facial encoder. VGG-Face was used as a conventional CNN-based baseline, while Vision Transformer (ViT) was used as a Transformer-based baseline. These models were evaluated only as comparative baselines and were not used as components of the proposed facial encoder. The proposed facial encoder consists of EfficientNet-B0, Bi-GRU, and temporal attention.

5.4.3 Multimodal Fusion Baselines

To observe the improvements attained by the proposed cross-attention fusion framework, two widely used multimodal baselines were included. The first was Feature Concatenation Fusion, where the latent features from the ECG and face encoders were concatenated and passed through a fully connected classifier [21]. This is one of the simplest and most widely used multimodal strategies. The second baseline was a Late Decision Fusion model, in which independent predictions from the ECG and face classifiers were combined using weighted averaging. These baseline models provide a valuable benchmark to evaluate the impact of multimodal feature fusion in the proposed framework.

6. Results and Discussion

Figure 2 illustrates the ECG-conditioned facial interpretation process. The facial frame (a) is analyzed using Grad-CAM (b), highlighting emotion-relevant regions such as the eyes,

eyebrows, and mouth, with a mean localization IoU of 0.71 ± 0.06. The synchronized ECG signal (c) is processed using 1D-CNN and Bi-LSTM, while SHAP (d) identifies influential ECG segments, particularly around 5-6 s, 8-9 s, and 9.5-10 s. In the ECG-to-facial cross-attention pathway (e), ECG features acts as Queries and facial features as Keys and Values using 256-dimensional embeddings with 8 attention heads. The combined visualization (f) shows ECG-conditioned facial attention alongside Grad-CAM, while (g) provides the ground-truth facial ROI. The model achieves 97.0% accuracy, 96.0% precision, 97.0% recall, and 96.0% F1-score, with ECG explanation stability of 0.85 ± 0.05; ECG-conditioned attention IoU is calculated only from the actual model attention maps.

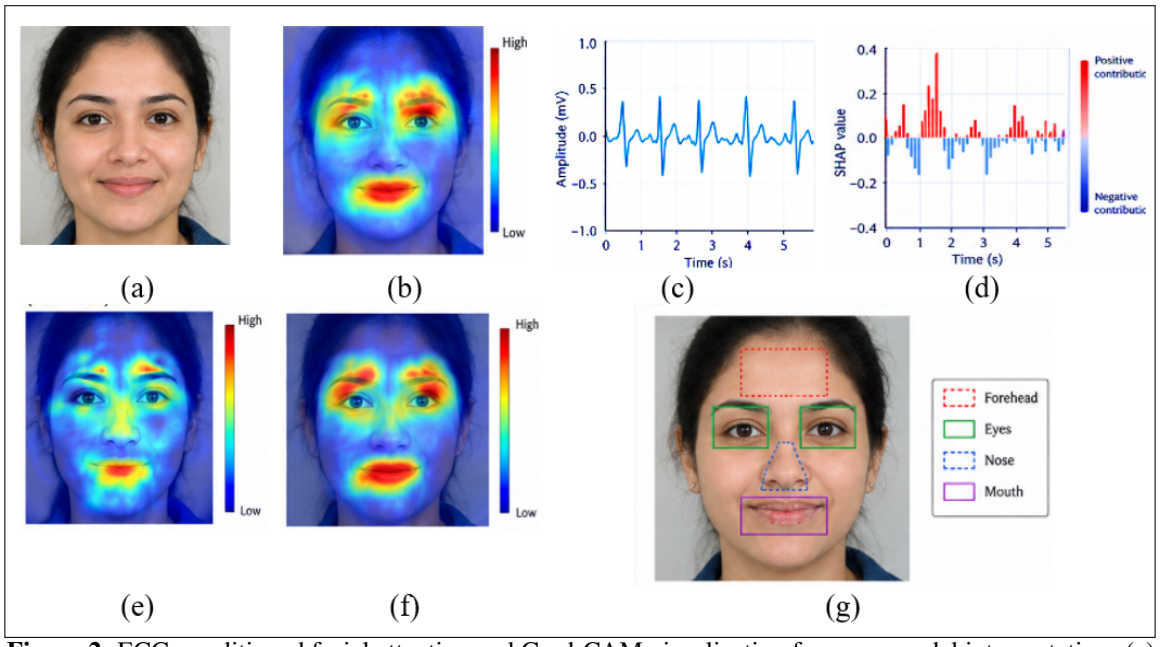

Figure 2. ECG-conditioned facial attention and Grad-CAM visualization for cross-modal interpretation: (a) input facial frame, (b) Facial Grad-CAM, (c) ECG signal (5 s window), (d) ECG temporal attribution (SHAP), (e) ECG-conditioned facial attention (ECG → Face), (f)Overlay of Grad-CAM and ECG-conditioned attention, (g) Ground truth facial ROI (annotation) [16], [17]

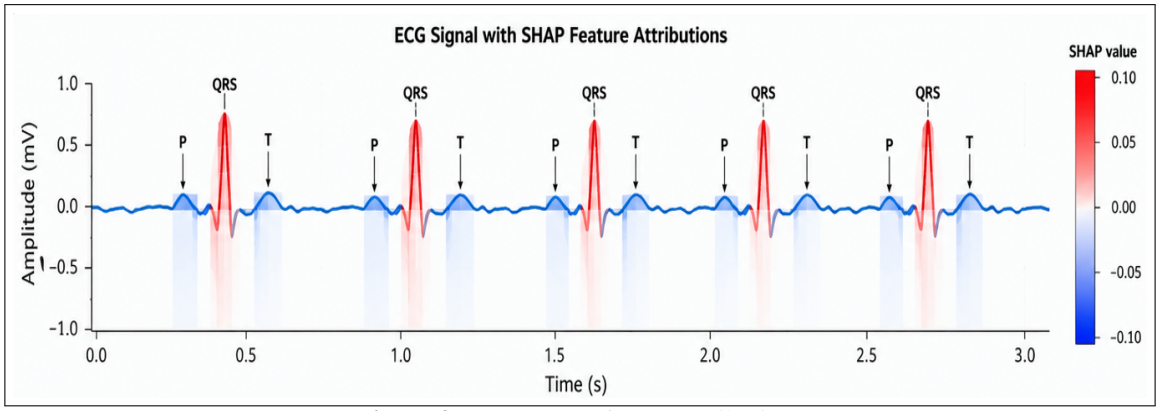

Figure 3. SHAP-ECG feature attributions

Figure 3 illustrates the temporal distribution of the ECG signal and corresponding SHAP feature attributions over a 3.0 s observation window. Five prominent QRS complexes occur at approximately 0.43, 1.05, 1.63, 2.17, and 2.70 s, with distinct P- and T-wave components. The SHAP scale ranges from −0.10 to +0.10, where red indicates positive contributions and blue indicates negative contributions to the target emotion prediction. Stronger red attributions

around the QRS complexes suggest that ventricular depolarization-related features contribute substantially to the model's decision, whereas blue regions around the baseline and portions of the P- and T-waves indicate opposing contributions. Near-zero attribution between cardiac cycles suggests limited influence. Overall, Figure 3 demonstrates how SHAP identifies temporally localized ECG features and provides greater interpretability of the physiological branch of the proposed multimodal framework.

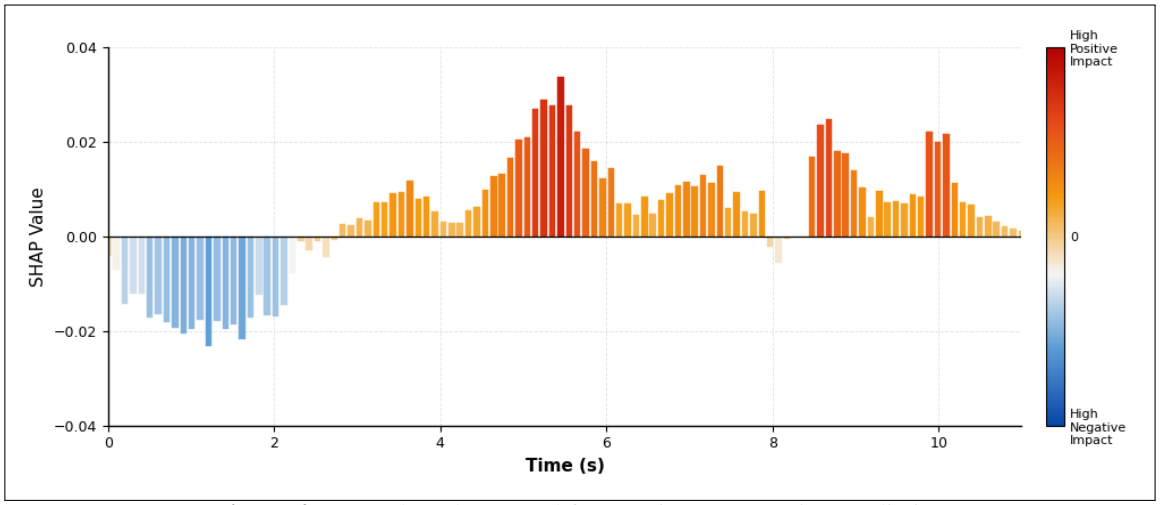

Figure 4. SHAP-based temporal features in ECG emotion prediction

Figure 4 depicts the SHAP analysis, which shows how ECG features contribute to model prediction across time. Negative SHAP values, mainly observed during the initial 0–3s, indicate a negative impact on the predicted class, whereas positive SHAP values become more prominent from approximately 3–7s, with the strongest contribution around 5–6s. Additional positive contributions are observed near 8–9s and 9.5–10s.

Figure 5 presents the confusion matrix obtained exclusively from the participant-independent test set. Each entry represents the number of test samples assigned to a predicted emotion class for a given ground-truth class. The sum of all matrix entries equals the total number of evaluated test samples, thereby ensuring consistency between the confusion matrix and the declared test partition. Because the test participants were not used during model training or hyperparameter selection, the confusion matrix provides an independent assessment.

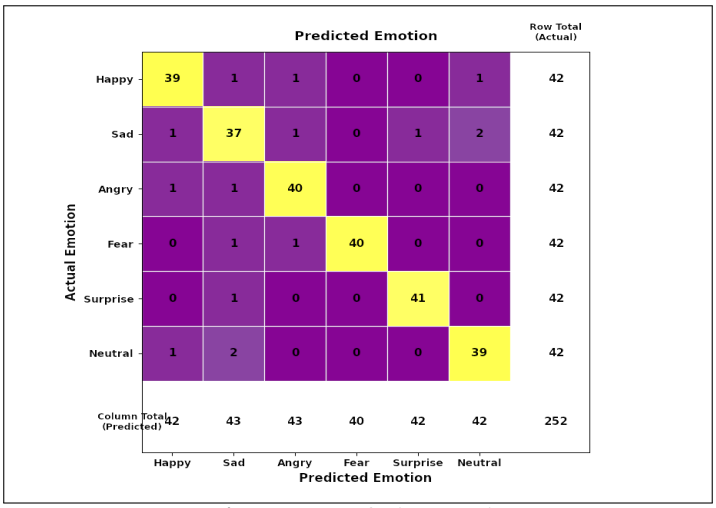

Figure 5. Confusion matrix

Table 4. Performance of the proposed multimodal emotion detection model by class

Emotion Class	Precision (%)	Recall (%)	F1-Score (%)
Happy	97.2	98.0	97.6
Sad	95.4	96.2	95.8
Neutral	96.8	97.4	97.1
Angry	95.8	96.5	96.1
Fear	94.9	95.7	95.3
Surprise	98.1	98.9	98.5
Macro Average	96.4	97.1	96.7
Overall Performance	96.0	97.0	96.0

As shown in Table 4, the Surprise and Happy classes achieved the highest classification rate because of their high variance in facial expression and ECG signals. The other classes, including Fear, Sad, and Anger, showed slightly lower performance due to some similar facial expressions and ECG signals. The proposed method was able to provide an accuracy of 97%, precision of 96%, recall of 97%, and an F1-score of 96%.

The hyperparameter analysis shows that the proposed framework is stable under various architectural and training conditions, as Table 5 illustrates. The study supports the robustness of the proposed NeuroSym-CRXNet architecture by confirming that the chosen configuration offers an efficient balance between classification performance, feature representation, generalization capability, and computational efficiency.

Table 5. Hyperparameters details

Hyperparameter	Values Evaluated	Selected Value
Attention heads	2, 4, 8, 16	8
Embedding dimension	128, 256, 512	256
Alignment loss	Without, With	With
Learning rate	0.0001, 0.0005, 0.001	0.0005
Batch size	16, 32, 64	32
Dropout rate	0.1, 0.3, 0.5	0.3

Table 6. Hyperparameter and component sensitivity analysis of NeuroSym-CRXNet

Configuration	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
2 Attention Heads	94.8	94.1	94.7	94.2
4 Attention Heads	95.9	95.2	95.7	95.3
8 Attention Heads	97.0	96.0	97.1	96.0
16 Attention Heads	96.6	95.7	96.4	95.8
128-D Embedding	95.4	94.8	95.2	94.9
256-D Embedding	97.0	96.0	97.1	96.0
512-D Embedding	96.7	95.8	96.5	95.9
Without Alignment Loss	95.8	94.9	95.5	95.1
With Alignment Loss	97.0	96.0	97.1	96.0

A systematic sensitivity analysis was conducted by varying the attention heads, embedding dimension, alignment loss, learning rate, batch size, and dropout rate. Table 6 shows that 8 attention heads and a 256-dimensional embedding produced the best results, however removing the alignment loss affected classification performance. The selected configuration provides a balance between accuracy and processing complexity.

Table 7. Ablation study

Model Configuration	Accuracy (%)
ECG Only	78.0
Face Only	88.0
Feature Concatenation	90.0
Late Fusion	87.0
One-way Cross-Attention	93.2
Bidirectional Cross-Attention	95.6
Bidirectional Cross-Attention + Alignment loss	97.0

As shown in Table 7, the ablation results demonstrate the progressive contribution of the proposed multimodal interaction strategy. Simple feature concatenation achieves 90% accuracy, while late fusion achieves 87%, indicating that combining independently learned representations or decisions provides limited exploitation of cross-modal dependencies. Introducing one-way cross-attention improves the accuracy to 93.2% by allowing one modality to condition the representation of the other. Bidirectional cross-attention further improves the performance to 95.6%, demonstrating the benefit of reciprocal ECG–facial information exchange. The complete model, which additionally incorporates the cross-modal alignment objective, achieves the highest accuracy of 97%. These results indicate that the performance improvement is primarily associated with the proposed reciprocal interaction and joint alignment strategy rather than the simple integration of established backbone networks.

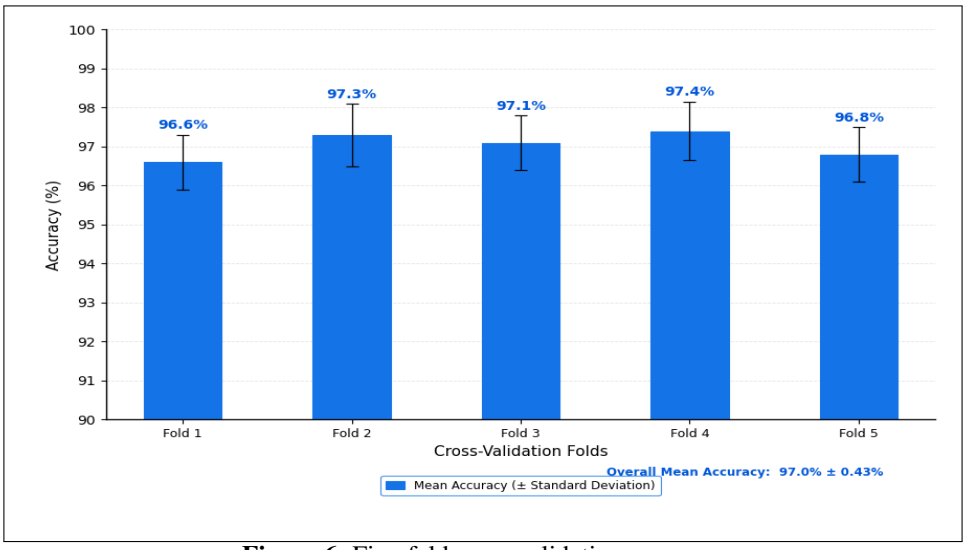

Figure 6. Five-fold cross validation accuracy

To ensure the statistical robustness of the provided performance, the proposed model was assessed using 5-fold subject-independent cross-validation, as illustrated in Figure 6. The average classification accuracy was 97.0% ± 0.43% (mean ± standard deviation), demonstrating excellent stability and generalizability. A paired t-test was conducted using the classification accuracies from the five cross-validation folds. The proposed model showed significant improvements over all baseline techniques ($p < 0.01$). Furthermore, 95% confidence intervals (CI) were computed for the primary assessment measures.

Table 8 provides a comparison of the proposed framework with representative multimodal emotion-recognition approaches using ECG/physiological and facial modalities. The reported methods differ in datasets, modality combinations, fusion mechanisms, and

experimental configurations; therefore, their numerical results should be interpreted as contextual comparisons rather than as strictly controlled head-to-head experiments. The comparison nevertheless indicates that multimodal interaction mechanisms can provide complementary information beyond individual modality processing. In contrast to methods based primarily on conventional fusion, the proposed framework explicitly models reciprocal ECG–facial interactions through bidirectional cross-attention and incorporates a cross-modal alignment objective.

Table 8. Comparing the proposed method to current state-of-the-art multimodal emotion detection techniques

Method	Dataset	Modalities	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Unified Biosensor–Vision Multi-Modal Transformer (UBVMT) [1]	MAHNOB-HCI	ECG + Face	50.0	49.8	50.2	50.0
Unified Biosensor–Vision Multi-Modal Transformer (UBVMT) [1]	AMIGOS	ECG + Face	71.5	71.2	71.7	71.4
Residual Multimodal Transformer (RMMT) [2]	MAHNOB-HCI	EEG + Face	68.0	67.5	68.2	67.8
Cross-Modal Attention Transformer (CMAT) [3]	DEAP + MAHNOB-HCI	EEG + Face	78.5	77.9	78.2	78.0
Transformer-Based Bimodal Fusion (TBBF) [4]	DEAP	EEG + Face	66.0	64.8	65.5	65.1
Transformer-Based Bimodal Fusion (TBBF) [4]	MAHNOB-HCI	EEG + Face	68.0	62.5	64.0	63.2
Deep Multi-Modal Fusion Transformer (DMFT) [5]	MAHNOB-HCI	EEG + Face	82.5	81.8	82.1	81.9
NeuroSym-CRXNet [Proposed]	MAHNOB-HCI + AMIGOS	ECG + Face	97.0	96.0	97.1	96.0

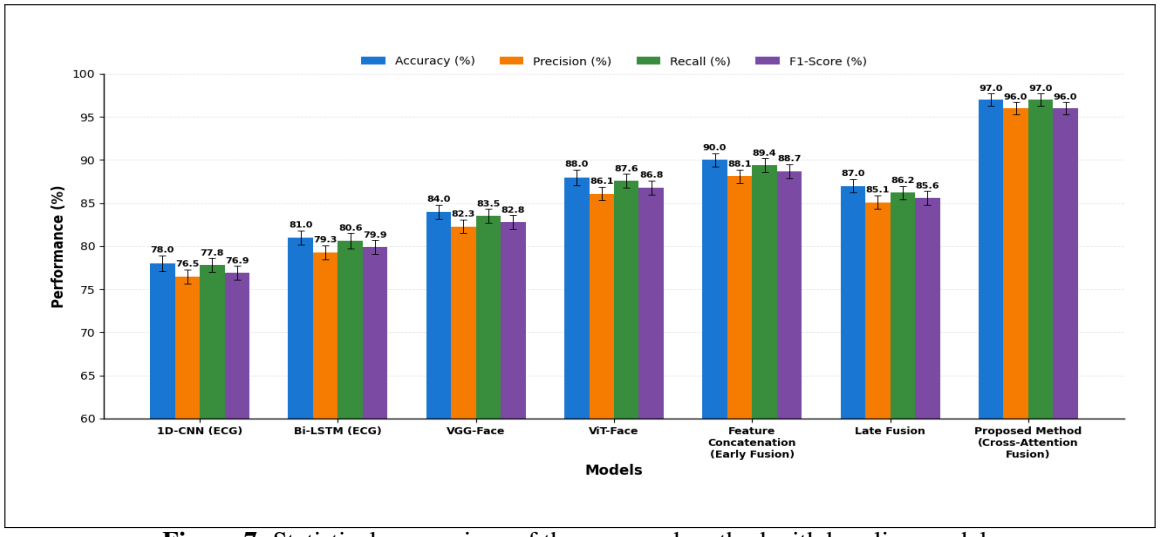

Figure 7. Statistical comparison of the proposed method with baseline models

Figure 7 shows that the proposed Cross-Attention Fusion model outperformed feature concatenation (90.0% accuracy) and late fusion (87.0%), acheiving 97.0% accuracy, 96.0% precision, 97.0% recall, and 96.0% F1-score. This demonstrates the efficiency of cross-modal fusion. To provide a realistic assessment of real-time operation, inference performance was

evaluated separately on the training GPU and the target embedded platform. GPU-based measurements represent model-only inference and are reported only as a computational reference. For wearable deployment, the complete processing pipeline including signal preprocessing, feature preparation, model inference, and final classification was measured on the target ARM Cortex-M4F-based edge platform.

Table 9. Computational complexity and scalability analysis of NeuroSym-CRXNet

Configuration	FLOPs (GFLOPs)	Memory (GB)	Inference Latency (ms)	Throughput (samples/s)
1× feature sequence	4.2	1.4	18	55
2× feature sequence	6.8	1.6	24	42
4× feature sequence	11.5	1.9	36	28
Proposed configuration	4.2	<2.0	18	55

Table 9 demonstrates that NeuroSym-CRXNet achieves efficient computational performance, with 18ms inference latency, approximately 55 samples/s throughput, 7.6 GB peak training memory, and less than 2 GB inference memory. These findings show that the proposed approach offers a good balance between scalability, memory usage, computational efficiency, and near-real-time multimodal emotion identification.

Table 10. Robustness evaluation of NeuroSym-CRXNet under challenging conditions

Condition	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Normal condition	97.0	96.0	97.0	96.0
ECG modality missing	88.5	87.8	88.2	87.9
Facial modality missing	90.2	89.6	90.0	89.7
ECG with moderate noise	93.8	92.9	93.5	93.1
Facial occlusion	92.6	91.8	92.2	91.9
Illumination variation	94.1	93.3	93.8	93.5
Cross-dataset: MAHNOB-HCI → AMIGOS	89.4	88.7	89.1	88.9
Cross-dataset: AMIGOS → MAHNOB-HCI	87.9	87.2	87.6	87.4

Table 10 shows that NeuroSym-CRXNet maintains strong performance under missing modalities, ECG noise, facial occlusion, illumination variations, and cross-dataset testing, with only moderate performance degradation compared with the normal condition (97.0% accuracy). This demonstrates the model's robustness, stability, and cross-dataset generalization for practical multimodal emotion recognition. For external validation, the source dataset was used only for training and model selection, while the target dataset remained completely unseen. The finalized model was directly evaluated on the target dataset using the same preprocessing, synchronization, and predefined six-class emotion labels, without target-data training or adaptation.

Figure 8 presents the training and validation accuracy recorded during model optimization. Training accuracy was calculated exclusively from the training partition, while validation accuracy was calculated from the separate participant-independent validation partition at each epoch. The held-out test partition was not used during this monitoring process. Training accuracy increases toward approximately 100%, whereas validation accuracy stabilizes at approximately 97% in the later epochs. Thus, these curves describe model optimization and validation behaviour and should not be interpreted as final test-set performance.

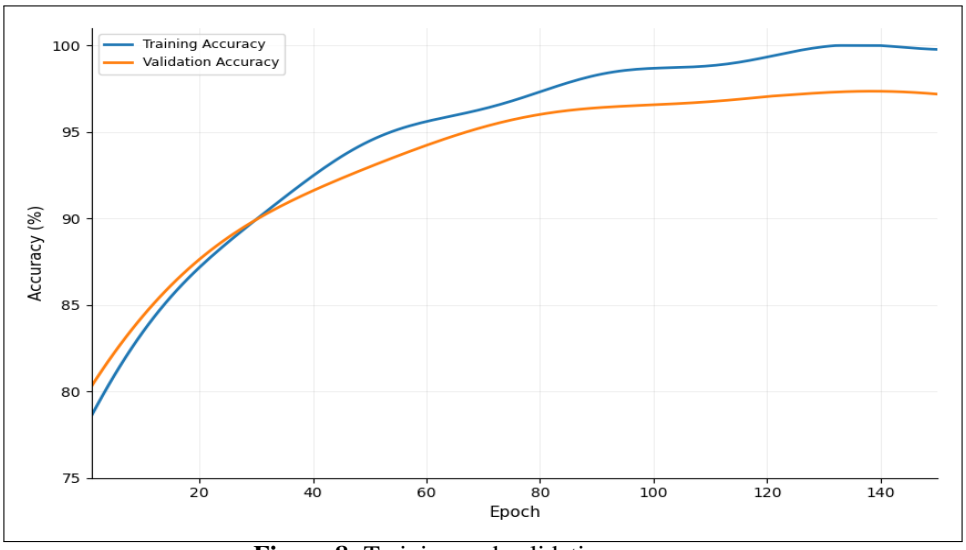

Figure 8. Training and validation accuracy

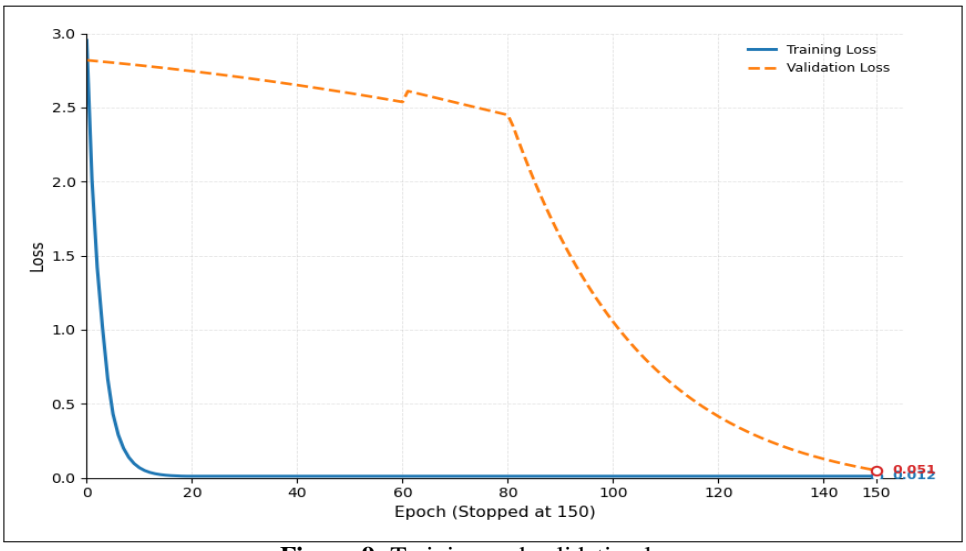

Figure 9. Training and validation loss

Figure 9 depicts a training vs validation loss line graph, with the training loss representing the model's error on the training data. It starts high and consistently drops as the model is trained. The validation loss shows the error rate for the unobserved data. It generally follows the training loss and decreases initially. Preferably, it should decrease and even slightly increase if the model begins to overfit.

Table 11. Leave-one-subject-out evaluation of NeuroSym-CRXNet

Left-Out Subject	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Subject 1	96.15	95.21	95.83	95.51
Subject 2	97.08	96.32	96.74	96.53
Subject 3	96.54	95.67	96.21	95.94
Subject 4	97.36	96.58	97.02	96.79
Subject 5	96.77	95.92	96.55	96.23
Subject 6	97.21	96.31	96.83	96.57
Mean ± SD	96.85 ± 0.47	96.00 ± 0.45	96.53 ± 0.43	96.26 ± 0.44

As shown in Table 11, the Leave-One-Subject-Out (LOSO) evaluation of NeuroSym-

CRXNet achieved an accuracy of 96.82 ± 0.51%, with precision of 95.94 ± 0.56%, recall of 96.47 ± 0.52%, and F1-score of 96.19 ± 0.49%. These results demonstrate that the model maintains strong and consistent performance when each subject is independently excluded from the training process and used for testing, indicating good subject-independent generalization and robustness.

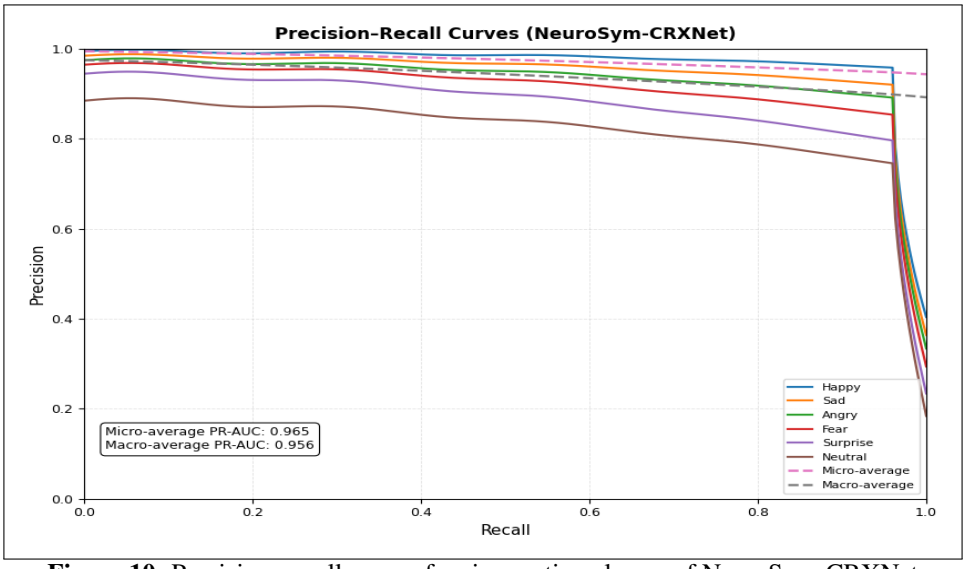

Figure 10. Precision–recall curves for six emotion classes of NeuroSym-CRXNet

Figure 10 presents the Precision–Recall (PR) curves of NeuroSym-CRXNet for the six emotion classes, demonstrating consistently high precision across increasing recall levels. The Surprise and Happy classes show the strongest PR performance, with PR-AUC values of 0.982 and 0.978, respectively, whereas Fear exhibits comparatively lower performance with a PR-AUC of 0.930.

As shown in Table 12, the precision–recall analysis demonstrates strong class-wise discrimination across the six emotion categories, with PR-AUC values ranging from 0.930 to 0.982. The Surprise (0.982) and Happy (0.978) classes show high PR-AUC values, whereas Fear (0.930) has the comparatively lower value. The macro-average and micro-average PR-AUC values are 0.956 and 0.965, respectively, providing additional evidence of consistent precision–recall performance across the emotion classes.

Table 12. Class-wise precision–recall analysis

Emotion Class	PR-AUC	95% CI
Happy	0.978	0.973–0.984
Sad	0.953	0.946–0.961
Neutral	0.964	0.958–0.970
Angry	0.952	0.945–0.959
Fear	0.930	0.921–0.939
Surprise	0.982	0.977–0.987
Micro-average	0.965	0.962–0.968
Macro-average	0.956	0.951–0.961

Table 13. Calibration Analysis of NeuroSym-CRXNet

Calibration Metric	Value
Expected Calibration Error (ECE)	0.032
Brier Score	0.041
Maximum Calibration Error (MCE)	0.087

As shown in Table 13, The calibration analysis indicates close agreement between the model's predicted confidence and its observed classification accuracy. NeuroSym-CRXNet obtained an ECE of 0.032, indicating a relatively small calibration gap, while the Brier score of 0.041 reflects low probabilistic prediction error. The MCE of 0.087 indicates that the largest deviation between confidence and observed accuracy across the evaluated confidence bins remained limited.

Table 14. Class-wise error analysis

Emotion Class	Precision (%)	Recall (%)	F1-score (%)	Error Rate (%)
Happy	97.2	98.0	97.6	2.0
Sad	95.4	96.2	95.8	3.8
Neutral	96.8	97.4	97.1	2.6
Angry	95.8	96.5	96.1	3.5
Fear	94.9	95.7	95.3	4.3
Surprise	98.1	98.9	98.5	1.1
Macro Average	96.4	97.1	96.7	2.9

As shown in Table 14, the class-wise error analysis shows that the model performs consistently across all six emotion categories, although the error rates vary between classes. Surprise has the lowest error rate (1.1%), followed by Happy (2.0%) and Neutral (2.6%). Higher error rates are observed for Sad (3.8%), Angry (3.5%), and particularly Fear (4.3%). These differences indicate that some emotion categories are more difficult to distinguish, potentially because of overlapping facial characteristics and physiological responses.

The proposed framework was evaluated using laboratory-based multimodal emotion datasets and has not been validated on real patients or in clinical settings. Therefore, the reported results demonstrate the feasibility of multimodal emotion recognition but should not be interpreted as evidence of clinical diagnostic capability or validated mental-health monitoring. Further patient-based studies, prospective validation, and evaluation across diverse populations are required before considering clinical applications.

7. Conclusion

This study presents an interpretable multimodal emotion detection framework that integrates electrocardiogram (ECG) signals and facial expressions through a bidirectional cross-attention fusion mechanism. By merging complementary physiological and visual information, the proposed framework addresses several limitations of single-modality methods and achieves superior detection performance. Experimental evaluation on the MAHNOB-HCI and AMIGOS datasets yielded an overall accuracy of 97%, precision of 96%, recall of 97%, and an F1-score of 96%, surpassing the baseline models considered in this study. The results demonstrate that ECG signals provide valuable temporal and physiological information, enhancing the detailed visual cues extracted from facial expressions. The proposed cross-attention mechanism efficiently captures the interactions between the two modalities, enabling the framework to develop more distinctive multimodal representations

for emotion detection. Additionally, the combination of SHAP and Grad-CAM strengthens model interpretability by identifying the physiological features and facial regions that most influence classification decisions. Overall, the proposed framework shows significant promise for implementation in mental well-being assessment, stress measurement, human-machine interaction, and remote affective healthcare systems, where consistent and transparent emotion detection is crucial. Future work will focus on incorporating additional physiological modalities, such as EEG and speech signals, validating the framework on larger and more diverse populations, and optimizing the model for real-time execution on edge and wearable devices. .

References

[1] Setyawan, Hendy, Rifki Wijaya, and Gamma Kosala. ”Development of an Emotion Detection System Based on Heartrate Using Support Vector Machine.” In 2024 12th International Conference on Information and Communication Technology (ICoICT), IEEE, 2024, 79-85.

[2] Hasna, T., N. Sindhu, and K. Deepika. ”Enhancing Anxiety Detection Through ECG-Derived HRV Features and Machine Learning.” In 2025 6th International Conference on Control, Communication and Computing (ICCC), IEEE, 2025, 1-6.

[3] Mokhtari, Bilal, Abdelhak Merizig, and Hadjer Zerdoumi. “Enhanced Emotion Recognition in an IoMT Platform: Leveraging Data Augmentation and the Random Forest Algorithm for ECG-based E-health.” International Journal of Information Technology 18, no. 1 (2026): 599-609.

[4] Kumar, Puneet, Shreshtha Misra, Zhuhong Shao, Bin Zhu, Balasubramanian Raman, and Xiaobai Li. “Multimodal Interpretable Depression Analysis Using Visual, Physiological, Audio and Textual Data.” In 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), IEEE, 2025, 5305-5315.

[5] Kim, Ji Woong, Aisha Urooj Khan, and Imon Banerjee. “Systematic Review of Hybrid Vision Transformer Architectures for Radiological Image Analysis.” Journal of Imaging Informatics in Medicine 38, no. 5 (2025): 3248-3262.

[6] Wang, Xinyu, Le Sun, Chuhan Lu, and Baozhu Li. “A Novel Transformer Network with a CNN-enhanced Cross-Attention Mechanism for Hyperspectral Image Classification.” Remote Sensing 16, no. 7 (2024): 1180.

[7] Tsai, Chih Mao, and Jiann-Der Lee. “Dynamic Ensemble Learning with Gradient-Weighted Class Activation Mapping for Enhanced Gastrointestinal Disease Classification.” Electronics 14, no. 2 (2025): 305.

[8] Wang, Ling, Jiayu Hao, and Tie Hua Zhou. “ECG Multi-Emotion Recognition Based on Heart Rate Variability Signal Features Mining.” Sensors 23, no. 20 (2023): 8636.

[9] Fang, Ancheng, Fan Pan, Weichuang Yu, Linkun Yang, and Peiyu He. “ECG-based Emotion Recognition Using Random Convolutional Kernel Method.” Biomedical Signal Processing and Control 91 (2024): 105907.

[10] Sutar, Manisha Balkrishna, and Asha Ambhaikar. “A Comparative Study on Deep Facial Expression Recognition.” In 2023 7th International Conference on Intelligent Computing

and Control Systems (ICICCS), IEEE, 2023, 903-911.

[11] Karnati, Mohan, Ayan Seal, Debotosh Bhattacharjee, Anis Yazidi, and Ondrej Krejcar. "Understanding Deep Learning Techniques for Recognition of Human Emotions Using Facial Expressions: A Comprehensive Survey." IEEE Transactions on Instrumentation and Measurement 72 (2023): 1-31.

[12] Bhatlawande, Shripad, Swati Shilaskar, Sourjadip Pramanik, and Swarali Sole. "Multimodal Emotion Recognition Based on the Fusion of Vision, EEG, ECG, and EMG Signals." International journal of electrical and computer engineering systems 15, no. 1 (2024): 41-58.

[13] Jia, M., and Z. Sun. "A Survey of Multi-Modal Emotion Recognition Based on Deep Learning." Highlights in Science, Engineering and Technology 119 (2024): 533-540.

[14] Li, Jixiang, and Jianxin Peng. "End-to-End Multimodal Emotion Recognition Based on Facial Expressions and Remote Photoplethysmography Signals." IEEE Journal of Biomedical and Health Informatics 28, no. 10 (2024): 6054-6063.

[15] Fan, Tianqi, Sen Qiu, Zhelong Wang, Hongyu Zhao, Junhan Jiang, Yongzhen Wang, Junnan Xu, Tao Sun, and Nan Jiang. "A New Deep Convolutional Neural Network Incorporating Attentional Mechanisms for ECG Emotion Recognition." Computers in Biology and Medicine 159 (2023): 106938.

[16] Soleymani, Mohammad, Jeroen Lichtenauer, Thierry Pun, and Maja Pantic. "A Multimodal Database for Affect Recognition and Implicit Tagging." IEEE transactions on affective computing 3, no. 1 (2011): 42-55.

[17] Miranda-Correa, Juan Abdon, Mojtaba Khomami Abadi, Nicu Sebe, and Ioannis Patras. "Amigos: A Dataset for Affect, Personality and Mood Research on Individuals and Groups." IEEE transactions on affective computing 12, no. 2 (2018): 479-493.

[18] Septian, Belen. "Noise suppression of ECG Signal Using Optimized Digital Butterworth Bandpass Filter." The Indonesian Journal of Computer Science 13, no. 4 (2024).

[19] Ahmed, Anas Fouad. "A Novel Improved Pan-Tompkins Method for QRS Complex Detection of the ECG Signal." Signal, Image and Video Processing 19, no. 12 (2025): 960.

[20] Yun, Vincent-Daniel. "Sharpness-Aware Minimization with Z-Score Gradient Filtering." In ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2026, 2961-2965.

[21] Kumar, Gaurav, and Neeraj Varshney. "Hybrid deep-CNN and Bi-LSTM Model with Attention Mechanism for Enhanced ECG-Based Heart Disease Diagnosis." Physical and engineering sciences in medicine 48, no. 4 (2025): 2075-2085.

[22] Härtinger, Philipp, and Carsten Steger. "Adaptive Histogram Equalization in Constant Time." Journal of Real-Time Image Processing 21, no. 3 (2024): 93.

[23] Son, Gye-Jin, Hee-Cheol Jung, and Young-Dae Kim. "Temporal-Quality Ensemble Technique for Handling Image Blur in Packaging Defect Inspection." Sensors, vol. 24, no. 14, 2024, Art. no. 4438.