



Evolutionary Cross-Modal Attention Framework for Multimodal Human Emotion Recognition Using Hybrid Deep Learning

Dhruva M S^{1*}, Sunitha R²

¹ Department of Computer Science and Engineering, BNM Institute of Technology, Visvesvaraya Technological University, Bengaluru 560070, India

² Department of Artificial Intelligence and Machine Learning, BNM Institute of Technology, Visvesvaraya Technological University, Bangalore 560070, India

Corresponding Author Email: msdhruva@gmail.com

Copyright: ©2026 The authors. This article is published by IETA and is licensed under the CC BY 4.0 license (<http://creativecommons.org/licenses/by/4.0/>).

<https://doi.org/10.18280/jesa.590925>

ABSTRACT

Received: 28 June 2026

Revised: 17 September 2026

Accepted: 26 September 2026

Available online: 30 September 2026

Keywords:

multimodal emotion recognition, cross-modal attention, particle swarm optimization, vision transformer, BERT, multimodal fusion

Recognizing human emotions is an important aspect of emotional computing, enabling intelligent systems to observe human behavior across multiple communication channels. However, current multimodal approaches still rely on static fusion strategies that provide little or no consideration for dynamic interdependencies among heterogeneous modalities. This paper presents an evolutionary cross-modal attention framework (ECMAF) to recognize human emotion from multimodal data comprising visual, audio, and textual modalities. The framework uses modality-specific deep encoders to extract discriminative features across the three modalities and construct a multimodal emotion vector. The cross-modal attention mechanism is applied to extract inter-modal relationships, and particle swarm optimization (PSO) is used to adaptively optimize the weights assigned to the modalities and their hyperparameters before emotion recognition. The experiment conducted on the CMU-MOSEI and IEMOCAP datasets gave 89% and 85% accuracy, respectively, with Macro F1-scores of 88% and 84%, respectively. With respect to the experimental settings mentioned above, ECMAF proved to be better than the baseline algorithms, with ablation studies suggesting that cross-modal attention and PSO-adaptive fusion made significant contributions. Additional comparative studies and ablation analyses confirm that ECMAF consistently outperforms current state-of-the-art multimodal emotion recognition (MER) systems with respect to accuracy and adaptability in performing fusion operations.

1. INTRODUCTION

Human emotion recognition (HER) has become a key research area within affective computing, focusing on enabling intelligent systems to identify and understand a user's emotional state to support natural human-computer interaction. The accurate identification of emotion is critical in real-world applications, including healthcare monitoring, intelligent tutoring systems, virtual assistants, autonomous vehicles, social robotics, and personalized recommendation systems. Earlier emotion recognition systems used only one type of signal as input to detect emotions, such as facial expressions, speech, or text. Even though unimodal emotion recognition achieved positive results in laboratory-controlled settings, these systems are typically not very robust because emotions are complex and often influenced by environmental conditions, occlusion, speaker variability, and ambiguous contexts [1]. The development of deep learning has been instrumental in improving emotion recognition using only one modality by automatically extracting features from diverse data. Examples include convolutional neural networks (CNNs) that have been developed to identify features of faces [2], long short-term memory (LSTM) and bidirectional LSTM

(BiLSTM) that have been developed to detect features from voice over time; and transformer-based language models (e.g., BERT) that have been developed to teach how to get semantic meaning from types of natural language (text) [3]. Currently, relying on a single modality is insufficient; emotional data is often shared across multiple channels.

Multimodal emotion recognition (MER) is an emerging area of study that addresses the limitations of unimodal systems by integrating complementary data from visual, acoustic, and textual sources [4]. By using multiple data sources simultaneously to derive emotion data, multimodal systems tend to achieve higher recognition accuracy and be more robust to challenging conditions than unimodal systems. Few examples of multimodal fusion approaches have been designed by various researchers. These explicitly model interactions among unimodal, bimodal, and trimodal feature sets, demonstrating the efficacy of comprehensive feature fusion [5]. Furthermore, few mechanisms have addressed the need to improve computational efficiency while preserving cross-modal interactions through low-rank decompositions [6]. The emergence of transformer models has ushered in a new era of applications in MER. Few researchers have introduced a class of pairwise, directionally cross-modal

attention mechanisms that can model long-range dependencies across disparate modalities without temporal alignment [7]. Recently, hierarchical attention and self-distillation within transformer-based architectures have been shown to yield superior contextual representational learning for emotion recognition from conversational environments [8]. Cross-modal transformer architectures have been successfully developed and combined with pre-trained encoders. They learn the semantic relationships between modalities while aligning features across modalities to produce integrated multimodal representations [9]. Studies have demonstrated that using adaptive cross-modal attention improves multimodal representation learning and emotion classification in complex conversational scenarios [10]. However, most existing attention-based fusion methods assume fixed modality contributions across all training conditions, implicitly assuming that all modalities are equally reliable in producing emotional responses across varying degrees of emotional context. Despite impressive achievements by transformer-style multimodal frameworks, several important research challenges remain unaddressed. The novelty of evolutionary cross-modal attention framework (ECMAF) lies in its cross-modal attention-based interaction learning, together with particle swarm optimization (PSO)-based fusion optimization, rather than in cross-modal attention alone. Specifically, after obtaining modality-specific features and mapping them to a common latent space where cross-modal attention enhances their quality, PSO searches for the optimal parameter vector, which includes the weights for visual, audio, and textual fusion, as well as the learning rate and dropout rate. It is important to note that the modality weights are restricted to be positive and to sum to one. The PSO criterion includes both the validation accuracy and the Macro F1-score, thus providing the ability to select the optimal fusion based on overall correctness and balanced performance among classes. The critical point is that the PSO process is considered an outer optimization procedure and is not performed separately for each mini-batch. ECMAF not only serves as an evaluation benchmark but can also be perceived as a multimodal perception front-end for automatic intelligent systems. In the context of a human-machine interaction (HMI) system, the synchronized inputs from the visual, audio, and textual modalities can be fed to their corresponding modality-based encoders, and further, cross-modal attention mechanisms and adaptive fusion are performed to obtain an emotion representation or emotion class prediction. These outputs can then be used for the subsequent decision-making or interaction process. The same architecture may be useful for intelligent monitoring systems, where the input related to emotions can be considered one of many perceptual inputs.

The remainder of this paper is organized as follows: Section 2 reviews the literature on MER systems and evolutionary algorithms. The proposed methodology and mathematical formulation, along with descriptions of the experimental setup and the measures being evaluated, are presented in Section 3. Section 4 provides an overview of the experimental setup and evaluation measures used to build the results. Finally, Section 5 discusses conclusions and future directions of research based on the results of this research.

2. RELATED WORK

Emotion recognition is very important in the current

stressful, busy, modern lifestyle. There exists literature that has worked on stress, depression detection, mental health analysis, etc. Here, we discuss a few works categorized as recognition of emotion through single feature types, strategies for fusing multiple modalities, transformer-based MER, cross-modal attention mechanisms, and evolutionary optimization for emotion recognition.

2.1 Recognition of emotion through single feature types

Most emotion recognition studies in the past have used only one modality to recognize emotion, such as facial expressions, speech, or writing. Although Deep Neural Networks have helped improve feature recognition across modalities within a single modality (unimodality), it has been shown that using multiple modalities to capture emotional signals across multiple channels fails in many cases because modalities can provide complementary emotional signals that are not captured by a single modality alone.

The development of an attention fusion architecture for emotion recognition using physiological signals, proposed by Gong et al. [11], demonstrated that attention mechanisms can be employed to learn relationships between modality-specific features and to create a modality-independent model. Wang et al. [12] demonstrated an improvement in the ability of the two modalities (i.e., EEG and Acoustic) to recognize emotion by using both as heterogeneous sources of emotional information in both models. Nonetheless, unimodal and bimodal systems, which lack redundancy, can still be affected by noise, occlusion, or the context in which they were created.

2.2 Strategies for fusing multiple modalities

The most recent MER systems integrate visual, audio, and written modalities into a single system to provide complementary emotional information across modalities. Existing fusion methods can be grouped into three general categories: early, late, and hybrid. Zadeh et al. [13] proposed a tensor fusion framework to combine unimodal, bi-modal, and tri-modal data using the Cartesian product operation. Liu et al. [14] proposed low-rank multimodal fusion methods to reduce the number of parameters required for multimodal fusion without compromising cross-modality comparisons, thereby reducing the computational intensity of tensor fusion.

Tang et al. [15] proposed a hierarchical multimodal fusion framework based on contrastive alignment techniques to enhance physiological emotion recognition. Their results showed that adaptive fusion techniques outperform static feature-aggregation methods, yielding more consistent features across modalities. Traditional fusion techniques assume that all modalities should be treated equally, regardless of their quality, and therefore are inflexible when working with disparate data quality.

2.3 Transformer-based multimodal emotion recognition

Transformer architectures have emerged as a highly popular choice for MER, thanks to their ability to model long-term dependencies and relationships among contextual signals. Tsai et al. [16] proposed the "Multimodal Transformer" (MulT) using directional pairwise cross-modal attention to model the relationship between the different modalities. Ma et al. [17] developed a transformer-based emotion recognition framework for conversation-based self-distillation, utilizing

hierarchical gated fusion to dynamically weight contributions from each modality.

Khan et al. [18] created MemoCMT, a cross-modal transformer architecture that fuses HuBERT and BERT representations using cross-attention mechanisms. A multimodal GRU with pairwise, directed cross-modal attention is developed by Qin et al. [19] to improve sentiment analysis performance by explicitly modeling modality dependencies. Tuerhong et al. [20] proposed an adaptive multimodal transformer with modality-exchange capabilities, enabling improved contextual feature learning. Lu and Feng [21] proposed an adaptive high-order transformer network that captures higher-order modality relationships to improve emotion recognition. Despite the impressive performance of transformer-based techniques, most approaches to date use static modality weighting and fixed fusion schemes.

2.4 Cross-modal attention mechanisms

Over the past several years, cross-modal attention has emerged as an effective means of modeling dependencies among multiple heterogeneous modalities. Zhao et al. [22] proposed a cross-modal gated feature enhancement framework to emphasize useful emotion-related features (cues) while downweighting redundant features. Dai et al. [23] utilized contrastive learning to improve cross-modal representation learning by fusing semantic features across modalities. While

all of these methods demonstrate the need for adaptive cross-modal interactions, many fail to account for the dynamic trustworthiness of each modality across different levels of interference, such as noise or missing data.

2.5 Evolutionary optimization for emotion recognition

Optimization techniques have been increasingly utilized to enhance the multimodal framework using feature selection, hyperparameter tuning, and adjustment of fusion weights. Traditional optimization techniques, such as grid search and Bayesian optimization, often incur high computational costs and may converge only to suboptimal solutions. Evolutionary optimization techniques such as GAs, PSOs, GWOs, and NSGA-II are known for their strong global searching ability in more complex optimization problems.

According to recent studies, evolutionary optimization has been shown to improve deep learning models using dynamically selected features and tuning hyperparameters [24, 25]. However, there is little research on integrating evolutionary optimization with cross-modal attention architectures for adaptive modality weighting. As a result, there remains a significant gap in research on MER frameworks that leverage both cross-modal attention and evolutionary optimization to flexibly adjust the contributions of each modality based on their contextual relevance and the quality of the associated data.

Table 1. Progress in multimodal emotion recognition (MER) models

Method	Input Modalities	Fusion Strategy	Attention Mechanism	Optimization/Training
Tzirakis et al. [2]	Audio + visual	End-to-end multimodal fusion	Not explicitly used	Gradient-based deep learning
TFN [3]	Text + visual + audio	Tensor fusion	No explicit cross-modal attention	End-to-end gradient-based learning
LMF [14]	Text + visual + audio	Low-rank tensor fusion	No explicit cross-modal attention	End-to-end gradient-based learning
MuT [16]	Text + visual + audio	Pairwise cross-modal transformer fusion	Directional cross-modal attention	Transformer-based gradient optimization
Cross-Modal Transformer [18]	Text + visual + audio	Cross-modal transformer fusion	Multi-head/cross-modal attention	Gradient-based optimization
ECMAF (proposed)	Visual + audio + text	Weighted multimodal fusion	Cross-modal attention	PSO-based adaptive fusion and hyperparameter optimization

Note: Tensor Fusion Network (TFN); Low-rank Multimodal Fusion (LMF); evolutionary cross-modal attention framework (ECMAF).

Table 1 shows the evolution of MER approaches towards multimodal fusion methods, indicating that the process of multimodal fusion has become structured. While Tensor Fusion Network (TFN) utilizes a tensor fusion method to model inter- and intra-modality interactions, Low-rank Multimodal Fusion (LMF) aims to reduce the computational complexity of tensor fusion. MuT, however, applies directional cross-modal transformers to learn interactions of different modality sequences. This set of approaches shows that interaction modeling between modalities is crucial, yet all of them apply a gradient-based approach to training the multimodal fusion architecture.

2.6 Gaps in previous research

There are several gaps in previous research that this study may address. First, the majority of MER systems currently use fixed (static) fusion strategies and do not provide an adaptive way to combine modalities based on their relative reliability.

Second, existing transformer-based systems do not adequately account for the continually changing importance of the various modalities during the fusion phase. Third, there has been very little research examining how to combine evolutionary optimization algorithms with cross-modal attention mechanisms to achieve adaptive multimodal fusion. Fourth, the majority of previous studies on MER have focused on improving prediction performance at the expense of developing robust models that can work effectively even when one or more modalities are damaged or unavailable.

This study proposes an ECMAF for multimodal HER using hybrid deep learning. This framework will employ a combination of modality-specific deep feature extraction, cross-modal attention learning, and evolutionary optimization to dynamically learn optimal fusion weights for each modality and recognition performance across all potential emotional states. Lastly, the generalizability of the developed framework across different datasets and its computational efficiency remain unexplored in much of the current literature.

3. PROPOSED MODEL

The ECMAF network proposed for multimodal HER incorporates the three input modalities (video streams, speech signals, and text transcripts) mentioned above to improve emotion classification. Heterogeneous input, containing both visual and auditory signals, is obtained from three modalities: visual input provides visual signals associated with facial expressions and other visual cues, auditory input provides

acoustic signals such as pitch, tone, and speech intensity, and the textual input provides semantic and contextual information based on the spoken or written information contained within. The proposed architecture for an ECMAF for multimodal HER using hybrid deep learning comprises six distinct layers as shown in Figure 1: data acquisition and pre-processing, modality-specific feature extraction, temporal representation learning, cross-modal attention fusion, evolutionary optimization, and emotion classification and explainability.

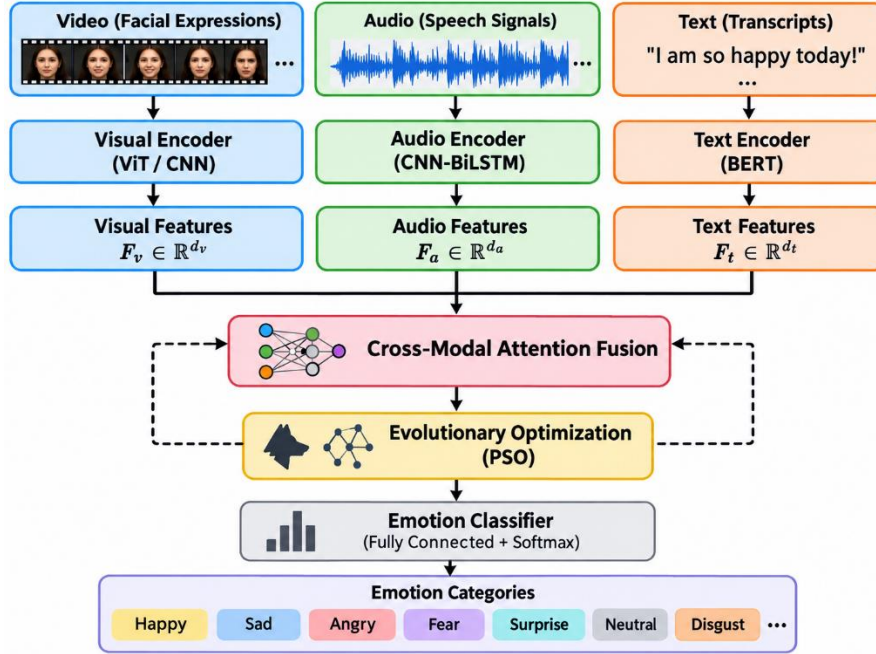


Figure 1. Evolutionary cross-modal attention framework (ECMAF)

Modalities are processed separately with an individual deep learning encoder for each modality. Visual data uses a ViT or CNN model to analyze video streams and extract discriminative spatial features, denoted as F_v . Audio data uses a CNN-BiLSTM approach to identify both temporal and spectral features, forming audio features F_a . Textual data uses the BERT model to produce contextual representations, denoted as F_t , which are then sent to the cross-modal attention fusion module, the primary component of the proposed architecture. The cross-modal attention fusion module learns the interdependent relationships between the three modalities of video, audio, and text feature extraction.

This process generates a comprehensive multimodal aggregation based on how the cross-modal relationships are modelled. To further increase adaptability, an evolutionary optimization module that employs PSO or GWO is also used. The iterative process of updating fusion weights and attention parameters using the optimizer improves emotional recognition accuracy with each iteration. The broken feedback paths indicate that updates will occur continuously as the fusion method is adapted during training. The newly created multimodal representations then flow into a fully connected layer, followed by a softmax layer to determine the final emotion class. The framework will provide multi-dimensional emotion recognition for happy, sad, angry, fear, neutral, and disgust states. By leveraging modality-specific deep feature extraction, cross-modal attention-aware fusion, and evolutionary optimization, a reliable and accurate emotional recognition framework is proposed. This will be especially

useful when working with noisy, partial, or multiple types of data sources.

This section presents the mathematical model for the proposed ECMAF. The ECMAF consists of modality-specific feature extraction, feature projection, cross-modal attention fusion, evolutionary optimization, emotion classification, and minimizing loss. The aim of the mathematical model is to effectively integrate heterogeneous emotional signals from visual, acoustic, and textual sources and to adaptively optimize the importance of each source modality.

Let a multimodal dataset be represented with regard to Eq. (1).

$$DD = \{(X_i^v, X_i^a, X_i^t, y_i)\}_{i=1}^N \quad (1)$$

Here, N indicates the total training samples, X_i^v refers to the visual inputs, X_i^a indicates the audio inputs, X_i^t denotes the textual inputs, $y_{i \in \{1,2,C\}}$ indicates the corresponding emotion label, C is the total emotion classes. The objective of the proposed model is to learn the mapping function as Eq. (2).

$$f: (X^v, X^a, X^t) \rightarrow y \quad (2)$$

This is to predict emotional categories by using all information, relating to multiple modalities so that information is correlated across modalities. This establishes the supervised learning problem of MER.

3.1 Modality-specific feature extraction

All modalities are processed independently using modality-specific deep encoders to learn the emotional features of each modality. For example, the visual encoder is described as Eq. (3).

$$F_v = E_v(X^v) \quad (3)$$

Here, $F_v \in R^{dv}$ and d_v denotes the dimensionality of visual features, E_v is defined as the vision transformer (ViT) encoder and F_v^{dv} will be the extracted feature vector for visual modality. The audio signal is first converted into Mel-spectrogram representations and processed using a CNN-BiLSTM encoder is given as Eq. (4).

$$F_a = E_a(X^a) \quad (4)$$

Here, $F_a \in R^{da}$ and d_a is the audio feature dimension, the CNN-BiLSTM encoder is represented as E_a and the acoustic characteristics of a sound are represented as F_a^{da} . A BERT-based encoder E_t is employed to obtain contextual textual embeddings as given in Eq. (5).

$$F_t = E_t(X^t) \quad (5)$$

Here, $F_t \in R^{dt}$ and d_t denotes the text feature dimension, the textual representation is obtained using the BERT model as E_t and F_t^{dt} represents a contextual embedding of a text. The encoders specific to each modality are used to capture complementary information from the face, speech, and language, which together create MER.

3.2 Feature projection

Due to the different dimensions of the features extracted, they need to be projected into a common latent feature space before they are fused between modalities. The projection of features is described as Eqs. (6) to (8).

$$H_v = W_{vF_v} + b_v \quad (6)$$

$$H_a = W_{aF_a} + b_a \quad (7)$$

$$H_t = W_{tF_t} + b_t \quad (8)$$

Here, $W_v \in R^{d \times dv}$, $W_a \in R^{d \times da}$, $W_t \in R^{d \times dt}$ are trainable projection matrices, W_v, W_a, W_t are the learnable projection matrices, b_v, b_a, b_t are the bias vectors, H_v, H_a, H_t^d are the features project to the shared latent space. The projected features are given in Eq. (9).

$$H_v, H_a, H_t \in R^d \quad (9)$$

By projecting the features from different sources into a common latent space, it is effectively used to interact across different types of modalities for the purposes of fusing them together.

3.3 Cross-modal attention fusion

The cross-modal attention mechanism is a method for capturing interdependencies video, text and audio. The

attention function is written as follows:

$$Attention(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (10)$$

Here, Q is the query matrix, K is the key matrix, V is the value matrix, d_k is the key dimension used for normalizing the outputs of the attention function. The attention representations between two modalities are computed using Eqs. (11) to (13).

$$A_{va} = Attention(H_v, H_a, H_a) \quad (11)$$

$$A_{vt} = Attention(H_v, H_t, H_t) \quad (12)$$

$$A_{at} = Attention(H_a, H_t, H_t) \quad (13)$$

After the attention, the modality representations are processed using Eqs. (14) to (16).

$$Z_v = H_v + A_{va} + A_{vt} \quad (14)$$

$$Z_a = H_a + A_{va} + A_{at} \quad (15)$$

$$Z_t = H_t + A_{vt} + A_{at} \quad (16)$$

Here, Z_v, Z_a, Z_t are the visual, audio, and textual representations that have been enhanced by their corresponding attentions. Thus, visual representation with an attention enhanced by its audio and textual representations.

3.5 Evolutionary fusion optimization

The optimization of the multimodal fusion method utilizes the PSO technique to evaluate the contribution of each mode. The fusion weight values are defined using the vector provided in Eq. (17).

$$w = [w_v, w_a, w_t]^T \quad (17)$$

Subject to, $w_v + w_a + w_t = 1$. For the fused representation, we have Eq. (18).

$$F_{fusion} = w_v Z_v + w_a Z_a + w_t Z_t \quad (18)$$

The requisite PSO particle representation is given in Eq. (19).

$$x_i = [w_v, w_a, w_t, \eta, \lambda]^T \quad (19)$$

Here, w_v, w_a, w_t are weights associated with each modality, η is the learning rate, and λ is the dropout rate. The particle's position and velocity are determined using the following Eq. (20).

$$v_i^{t+1} = \omega v_i^t + c_1 r_1 (p_i - x_i^t) + c_2 r_2 (g - x_i^t) \quad (20)$$

Here, ω is the inertia weight, c_1 and c_2 define the acceleration coefficients, r_1 and r_2 are random variables, p_i is the personal best position, and g is the global best position. The PSO uses this adaptive technique to create an optimal set of fusion weights that will maximize recognition accuracy.

The objective of optimization is to jointly maximize the classification accuracy and the Macro F1-score as given in Eq.

(21).

$$J = \alpha Accuracy + \beta MacroF1 \quad (21)$$

Subject to: $\alpha + \beta = 1$. Here, α and β denote weighting factors that control how much each metric's evaluation score contributes to the overall performance score. Thus, the optimal parameter vector is defined as Eq. (22).

$$x^* = \arg \max_x J(x) \quad (22)$$

Here, x^* defines the optimal parameter values for maximizing overall classification performance. However, x^* optimally determined modality weights must be distinguished from the attention coefficients that arise from the cross-modal attention mechanism. Attention learning is the process that controls the interactions between modalities at the feature level using the query-key-value function in Eqs. (10) to (13), while PSO is responsible for determining the contribution of Z_v , Z_a , and Z_t during their fusion process. As a result, the attention mechanism operates at a different level from PSO.

3.5 Emotion classification

Following the extraction of the optimal multimodal feature representation, these features undergo processing using two dense layers with ReLU activation and a softmax classifier. The formula for classification is presented in Eq. (23).

$$\hat{y} = Softmax(W_c F_{fusion} + b_c) \quad (23)$$

Here, the classification weight matrix is denoted by W_c ; the bias Vector is represented by b_c . Thus, the Class Probabilities are calculated using Eq. (24).

$$P(y = c | F_{fusion}) = \frac{\exp(z_c)}{\sum_{k=1}^C \exp(z_k)} \quad (24)$$

Here, z_c denotes the classifier logit associated with emotion class c . The class with the greatest class probability will be chosen as the final prediction.

3.6 Categorical cross entropy loss for multimodal emotion recognition

The parameters in the network have been optimised via a categorical cross-entropy loss function using Eq. (25).

$$L_{CE} = - \left(\frac{1}{N} \right) \sum_{i=1}^N \sum_{c=1}^C y_{ic} \log(\hat{y}_{ic}) \quad (25)$$

Here, L_{CE} is the categorical cross entropy loss function is identified by N being the number of samples in the training mini-batch and the number of emotion classes being C . The predicted probability and true probability values for each class are denoted by $\hat{y}_{ic}$ and y_{ic} , respectively.

This represents the final objective of the optimisation as shown in Eq. (26).

$$\theta^* = \arg \min_{\theta} L_{CE} \quad (26)$$

Here, $\theta = \{W_v, W_a, W_t, W_c, b_c, f\{w\}\}$, the minimization of the loss function allows the Network to learn multidimensional

features that are modality specific; enable adaptive cross-modality interaction; enable optimum fusion weights for accurate prediction of multimodal emotionality. The overall framework jointly learns modality-specific representations, cross-modal interactions, and optimal fusion weights to improve multimodal emotion recognition performance.

4. RESULTS AND DISCUSSIONS

The proposed framework is developed and implemented using Python 3.11 and PyTorch 2.3. Access to pretrained transformer models is provided by the Hugging Face Transformers library. The PSO algorithm is used via the PySwarms library. The experiments reported in this paper are conducted on a workstation equipped with an NVIDIA RTX 4090 GPU with 24 GB of memory, an Intel Core i9 processor, and 64 GB of RAM, running Ubuntu 22.04 LTS. The ViT-base model is used to process the visual data and had a patch size of (16×16) and an embedding dimensionality of 768, while a hybrid CNN-BiLSTM model composed of three convolutional layers followed by two BiLSTM layers (with 256 hidden units each) is utilized to process audio data in order to capture the spectral and temporal characteristics of audio. The textual data is encoded with BERT-Base Uncased, with a hidden dimension of 768. The distinct extracted modality-specific features are projected into a shared latent space to enable multimodal interactions, and then processed by a cross-modal attention module with 8 attention heads and a hidden dimension of 512. A dropout rate of 0.3 is used in order to help prevent overfitting. Emotion classification is performed using 2 fully connected neural network layers with ReLU activation functions, followed by a softmax output layer.

PSO is used as the evolutionary optimization method to adaptively optimize the modalities' fusion weights and the chosen hyperparameters. Each particle represented a potential solution for the fusion weights across the visual, audio, and textual modalities, along with the hyperparameters used to train the model. A Swarm with 30 particles and a maximum of 100 iterations is used; an inertia weight of 0.7, a cognitive acceleration coefficient of 1.5, and a social acceleration coefficient of 1.5 are used. The search space for the modality weights is constrained so that the modality weights could never be negative. The AdamW optimizer is used to train the model; learning rate is 10^{-4} , Weight Decay is 0, and Batch Size = 32. The model is trained for up to 100 epochs. To prevent overfitting, an early stopping criterion with a patience of 10 epochs is added. During training, the categorical cross-entropy loss is minimized to optimize the network parameters.

AdamW optimizer with a learning rate of 1×10^{-4} and no weight decay along with a batch size of 32 was used for training the suggested ECMAF model. The CNN part of CNN-BiLSTM audio encoder is made up of 128 filters and then followed by 2 layers of BiLSTM with 256 hidden nodes each. 0.3 was taken as the value for dropout rate to avoid overfitting. PSO algorithm was set for evolutionary fusion optimization such that there were 30 particles and 100 maximum iterations, inertia weight was 0.7, cognitive coefficient was 1.5 and social coefficient was 1.5.

The performance of the proposed ECMAF is evaluated using two standard benchmark datasets: CMU-MOSEI and IEMOCAP. These datasets are primarily used because they feature synchronized visual, acoustic, and textual modalities, enabling a thorough assessment of multimodal emotion

recognition systems. CMU-MOSEI [26] is a large-scale multimodal dataset collected from over 1,000 speakers, containing more than 23,500 annotated video segments. Each video segment is aligned in three modalities: video, audio, and text. The dataset contains a wide range of speaking styles, recording conditions, and emotional expressions.

IEMOCAP [27] comprises about 12 hours of audiovisual recordings of dyadic interactions from 10 actors across 5 sessions. The dataset contains multimodal recordings and hand-annotated emotion labels. Because of its conversational quality and realistic expressions of emotion, IEMOCAP has become a de facto benchmark for multimodal emotion recognition. In each dataset, the emotion labels are translated into six common categories: happy, sad, angry, neutral, fear, and disgust, to maintain similarity between the two datasets.

To ensure a consistent level of data quality & compatibility across data sources, preprocessing is performed for each modality. The Visual Modality (video data) had frames extracted at a speed of 25 fps. Face Detection is performed using a Multi-task Cascaded Convolutional Network and resized to 224×224 pixels. Once the face is extracted, the data is normalized to the range $[0, 1]$ for proper model training. The audio modality recorded speech at 16 kHz, and the audio is normalized to 0 dB using a spectral-subtractive noise-reduction process. The log-Mel Spectrogram is created at a window size of 25ms, a hop size of 10 ms, with 128 Mel band filters. The Textual Modality had transcripts lowercased, special characters removed, and any irrelevant symbols removed. Finally, BERT is used for tokenization with a max sequence length of 128 tokens.

To keep the development and testing phases independent, the datasets are split into training, validation and test subsets in the ratio of 70/10/20, respectively. The five-fold cross-validation method is only used for the 70% portion of the training dataset. Therefore, the training subset will be further split into five folds where in each round four of them are used for training while the other one is used for validation. The procedure is performed iteratively till all of the five folds are used for validation once. The independently created 10% validation subset and 20% test subset are excluded from the five-fold cross-validation process.

The validation data for both model development and PSO-based parameters optimization are independent from the test data used in the final evaluation. In fact, the test data are used neither for parameter optimization nor for model selection or PSO fitness evaluation, but only in the final evaluation step. This way of splitting the data guarantees that no information from the test data affects the optimization process. In order to provide a uniform classification environment for the two datasets of CMU-MOSEI and IEMOCAP, the original emotions annotations of both datasets were mapped into six common emotions categories: Happy, Sad, Angry, Neutral, Fear, and Disgust. This was done using semantic equivalence between original and common label categories. The original labels that did not belong to the six-label setting were discarded from the classification process. Table 2 contains the original dataset labels and their mapping into common emotions with the final number of samples for each category.

The proposed framework is evaluated on several widely used classification metrics: accuracy, precision, recall, macro F1-score, weighted F1-score, and area under the curve (AUC) using the receiver operating characteristic (ROC) curve. Accuracy is defined as the proportion of correctly classified samples; precision and recall are defined as the classifier's

ability to minimize false positives and false negatives, respectively; and the F1 metric represents the harmonic mean of precision and recall, which is particularly helpful for imbalanced data sets.

Table 2. Original dataset labels and their mapping into common emotions

Dataset	Original Label	Final Common Class
CMU-MOSEI	Happy	Happy
CMU-MOSEI	Sad	Sad
CMU-MOSEI	Angry	Angry
CMU-MOSEI	Neutral	Neutral
CMU-MOSEI	Fearful	Fear
CMU-MOSEI	Disgusted	Disgust
IEMOCAP	Happiness	Happy
IEMOCAP	Sadness	Sad
IEMOCAP	Anger	Angry
IEMOCAP	Neutral	Neutral
IEMOCAP	Fear	Fear
IEMOCAP	Disgust	Disgust

The evaluation metrics are calculated as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (27)$$

$$Precision = \frac{TP}{TP + FP} \quad (28)$$

$$Recall = \frac{TP}{TP + FN} \quad (29)$$

$$F1 - score = 2 \times \frac{Precision \times recall}{Precision + recall} \quad (30)$$

where, TP , TN , FP , and FN denote true positives, true negatives, false positives, and false negatives, respectively.

4.1 Performance analysis on CMU-MOSEI

The baseline models used here to show the comparison are implemented with same environment setup as proposed ECMAF model. The performance comparison on the CMU-MOSEI dataset shows that ECMAF consistently outperforms all Baseline Methods across all evaluation metrics. This performance gap is shown in the data presented in Table 3 and Figure 2. For example, CNN and CNN-BiLSTM both use only one modality for emotional expression and yielded accuracies of 72% and 74%, respectively, indicating the limitations of unimodal frameworks for emotion extraction. Transformers (e.g., BERT, MulT, and Cross-Modal Transformer) performed better than both CNN and CNN-BiLSTM due to their ability to learn both contextual dependencies and cross-modal dependencies. For instance, the Cross-Modal Transformer achieves much higher accuracy than either CNN or CNN-BiLSTM, with 83% and a Macro F1-score of 81%, respectively.

The ECMAF framework achieved the highest performance among the evaluated algorithms, with an accuracy of 89%, precision of 88%, recall of 87%, Macro F1-score of 88%, and AUC of 94%. Compared to the strongest baseline, ECMAF achieved a 6 percentage point improvement in accuracy and a 7 PP improvement in the Macro F1-score. The integration of modality-specific deep encoders, cross-modal attention

fusion, and PSO-based adaptive optimization improves accuracy and performance across all evaluated modalities. Using a cross-modal attention mechanism, ECMAF captures interdependence between heterogeneous modalities. The

evolutionary optimization module allows dynamic adjustment of each modality's contribution based on its relevance or reliability to the task at hand.

Table 3. Performance targets on CMU-MOSEI

Method	Accuracy (%)	Precision (%)	Recall (%)	Macro F1-Score (%)	AUC (%)
CNN	72	71	70	70	78
CNN-BiLSTM	74	73	72	72	80
BERT	76	75	74	74	82
TFN	79	78	77	77	84
LMF	80	79	78	78	85
MuT	82	81	80	80	87
Cross-Modal Transformer	83	82	81	81	88
Proposed ECMAF	89	88	87	88	94

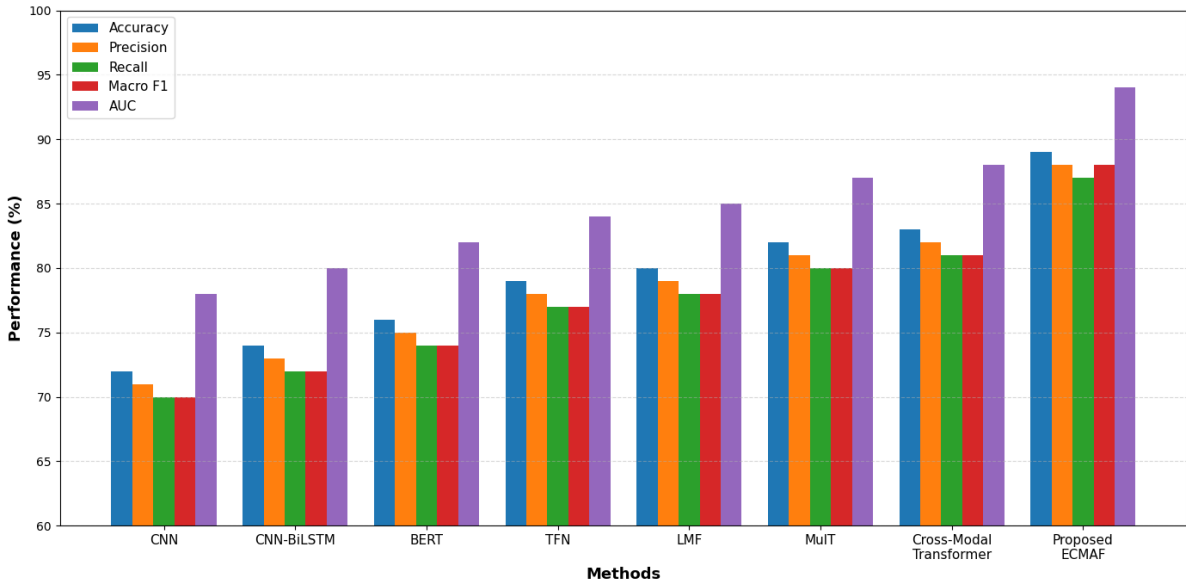


Figure 2. Performance comparison on CMU-MOSEI dataset

Table 4. Performance targets on IEMOCAP

Method	Accuracy (%)	Precision (%)	Recall (%)	Macro F1-Score (%)	Weighted F1 (%)
CNN	68	67	66	66	68
CNN-BiLSTM	70	69	68	68	70
BERT	72	71	70	70	72
TFN	74	73	72	72	74
LMF	75	74	73	73	75
MuT	77	76	75	75	77
Cross-Modal Transformer	78	77	76	76	78
Proposed ECMAF	85	84	83	84	86

4.2 Performance analysis on IEMOCAP

The comparative results from the IEMOCAP dataset are illustrated in Table 4 and Figure 3. Across all baseline models, multimodal methods outperform unimodal approaches. For example, traditional CNN methods achieved around 68% to 70% accuracy, while transformer-based multimodal models, such as MuT and Cross-Modal Transformer, yielded approximately 77% and 78%, respectively, on the same dataset. The best-performing method proposed in this research is ECMAF, achieving 85% accuracy, 84% precision, 83% recall, 84% Macro F1-score, and 86% weighted F1-score. Compared to the baseline Cross-Modal Transformer, ECMAF improved accuracy by 7 percentage points and Macro F1-score by 8 percentage points. These results suggest that adaptive

fusion techniques are effective in managing the complexities of emotional expression during conversation. The superior performance of ECMAF is attributed to its ability to leverage visual, acoustic, and textual modality information and to apply dynamic optimization via PSO.

Table 5 shows the ablation results of ECMAF for the IEMOCAP dataset. The entire ECMAF model is able to achieve an Accuracy of 85.0% and a Macro F1-score of 84.0%. Removing the PSO optimization will drop the accuracy and Macro F1-scores to 82.5% and 81.5%, respectively. The removal of the adaptive PSO fusion will drop the accuracy and Macro F1-scores to 82.0% and 81.0%, respectively. Further removal of the cross-modal attention mechanism drops the accuracy and Macro F1-scores to 81.0% and 79.5%, respectively. As for individual modality ablation

experiments, removal of the text modality drops the accuracy and Macro F1-scores to 78.0% and 77.0%, respectively. This is followed by removal of the visual modality (accuracy

80.0%, Macro F1-score 79.0%), and then the audio modality (accuracy 81.0%, Macro F1-score 80.0%).

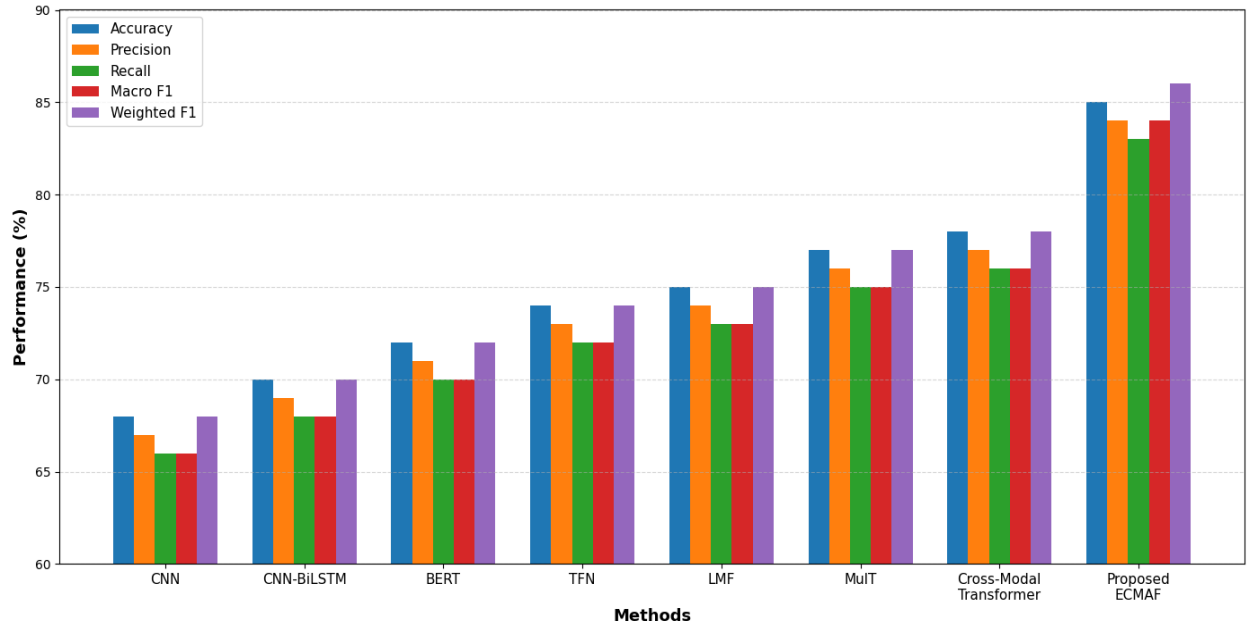


Figure 3. Performance comparison on the IEMOCAP dataset

Table 5. Ablation study of evolutionary cross-modal attention framework (ECMAF) on the IEMOCAP dataset

Configuration	Fusion Strategy	Accuracy (%)	Macro F1-Score (%)
Full ECMAF	PSO-based adaptive fusion	85.0	84.0
Without PSO	Fixed/adaptive fusion without PSO optimization	82.5	81.5
Without Cross-Modal Attention	PSO-based adaptive fusion	81.0	79.5
Fixed Fusion Weights	Fixed fusion	82.0	81.0
Without Visual Modality	PSO-based adaptive fusion	80.0	79.0
Without Audio Modality	PSO-based adaptive fusion	81.0	80.0
Without Text Modality	PSO-based adaptive fusion	78.0	77.0

5. CONCLUSION

This study proposes an ECMAF for recognizing human emotions across multiple modalities such as visual, audio, and textual. Current methods typically apply fixed strategies to combine multiple modalities. However, different modality types have varying importance across emotional contexts, and most existing approaches do not adapt to this variability. The proposed framework is assessed for effectiveness on CMU-MOSEI and IEMOCAP using metrics such as accuracy, precision, recall, F1-score, Macro-F1, and AUC.

The experiments conducted on the CMU-MOSEI and IEMOCAP databases yielded an accuracy of 89% and 85%, respectively, with Macro F1-scores of 88% and 84%. Using the IEMOCAP database, ECMAF performed at 85% accuracy against the baseline of Cross-Modal Transformer at 78%. The results from the ablation studies show that cross-modal attention and PSO fusion are critical in the framework. Future research can focus on lightweight models for real-time emotion recognition, self-supervised multimodal representation learning, and addressing missing modalities. Additionally, extending the current framework to multilingual or cross-cultural emotion datasets and using advanced multi-objective evolutionary optimization algorithms will improve the model's generalization and robustness for real-world use.

REFERENCES

- [1] Khanra, M., Roy, S.S. (2026). Data-centric review of multimodal emotion recognition: Datasets, feature extraction, data fusion and limitations. *Frontiers in Robotics and AI*, 13: 1810478. <https://doi.org/10.3389/frobt.2026.1810478>
- [2] Tzirakis, P., Trigeorgis, G., Nicolaou, M.A., Schuller, B., Zafeiriou, S. (2017). End-to-end multimodal emotion recognition using deep neural networks. *IEEE Journal of Selected Topics in Signal Processing*, 11(8): 1301-1309. <https://doi.org/10.1109/JSTSP.2017.2764438>
- [3] Devlin, J., Chang, M.W., Lee, K., Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, USA, pp. 4171-4186. <https://doi.org/10.18653/v1/N19-1423>
- [4] Younis, E.M.G., Mohsen, S., Houssein, E.H., et al. (2024). Machine learning for human emotion recognition: A comprehensive review. *Neural Computing and Applications*, 36: 8901-8947. <https://doi.org/10.1007/s00521-024-09426-2>
- [5] Du, Z., Wang, Y., Yang, X., Im, S.K., Wang, Z. (2026). MISA-GMC: An enhanced multimodal sentiment

- analysis framework with gated fusion and momentum contrastive modality relationship modeling. *Mathematics*, 14(1): 115. <https://doi.org/10.3390/math14010115>
- [6] Sun, L., Lian, Z., Liu, B., Tao, J. (2024). Efficient multimodal transformer with dual-level feature restoration for robust multimodal sentiment analysis. *IEEE Transactions on Affective Computing*, 15(1): 309-325. <https://doi.org/10.1109/TAFFC.2023.3274829>
- [7] Gan, C., Tang, Y., Fu, X., Zhu, Q., Jain, D.K., García, S. (2024). Video multimodal sentiment analysis using cross-modal feature translation and dynamical propagation. *Knowledge-Based Systems*, 299: 111982. <https://doi.org/10.1016/j.knosys.2024.111982>
- [8] Jin, H., Yang, T., Yan, L., Wang, C., Song, X. (2025). Multimodal emotion recognition in conversations using transformer and graph neural networks. *Applied Sciences*, 15(22): 11971. <https://doi.org/10.3390/app152211971>
- [9] Wu, Y., Mi, Q., Gao, T. (2025). A comprehensive review of multimodal emotion recognition: Techniques, challenges, and future directions. *Biomimetics*, 10(7): 418. <https://doi.org/10.3390/biomimetics10070418>
- [10] Dhruva, M.S., Sunitha, R., Rekha, K.S., et al. (2026). Optimized cross-modal data fusion framework for robust emotion recognition multimodal using hybrid deep learning techniques. *International Journal of Intelligent Engineering & Systems*, 19(2): 330. <https://doi.org/10.22266/ijies2026.0228.21>
- [11] Gong, L., Chen, W., Li, M., Zhang, T. (2024). Emotion recognition from multiple physiological signals using intra- and inter-modality attention fusion network. *Digital Signal Processing*, 145: 104278. <https://doi.org/10.1016/j.dsp.2023.104278>
- [12] Wang, Q., Wang, M., Yang, Y., Zhang, X. (2022). Multimodal emotion recognition using EEG and speech signals. *Computers in Biology and Medicine*, 149: 105907. <https://doi.org/10.1016/j.compbiomed.2022.105907>
- [13] Zadeh, A., Chen, M., Poria, S., Cambria, E., Morency, L.P. (2017). Tensor fusion network for multimodal sentiment analysis. In *Proceedings of Empirical Methods in Natural Language Processing*, Copenhagen, Denmark, pp. 1103-1114. <https://doi.org/10.18653/v1/D17-1115>
- [14] Liu, Z., Shen, Y., Lakshminarasimhan, V.B., Liang, P.P., Zadeh, A., Morency, L.P. (2018). Efficient low-rank multimodal fusion with modality-specific factors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Melbourne, Australia, pp. 2247-2256. <https://doi.org/10.18653/v1/P18-1209>
- [15] Tang, J., Ma, Z., Gan, K., Zhang, J., Yin, Z. (2023). Hierarchical multimodal fusion of physiological signals for emotion recognition with scenario adaptation and contrastive alignment. *Information Fusion*, 95: 215-228. <https://doi.org/10.1016/j.inffus.2023.102129>
- [16] Tsai, Y.H.H., Bai, S., Liang, P.P., Kolter, J.Z., Morency, L.P., Salakhutdinov, R. (2019). Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, pp. 6558-6569. <https://doi.org/10.18653/v1/P19-1656>
- [17] Ma, H., Wang, J., Lin, H., Zhang, B., Zhang, Y., Xu, B. (2024). A transformer-based model with self-distillation for multimodal emotion recognition in conversations. *IEEE Transactions on Multimedia*, 26: 776-788. <https://doi.org/10.1109/TMM.2023.3271019>
- [18] Khan, M., Tran, P.N., Pham, N.T., El Saddik, A., Othmani, A. (2025). MemoCMT: Multimodal emotion recognition using cross-modal transformer-based feature fusion. *Scientific Reports*, 15: 5473. <https://doi.org/10.1038/s41598-025-89202-x>
- [19] Qin, Z., Luo, Q., Zang, Z., Fu, H. (2025). Multimodal GRU with directed pairwise cross-modal attention for sentiment analysis. *Scientific Reports*, 15: 10112. <https://doi.org/10.1038/s41598-025-93023-3>
- [20] Tuerhong, G., Fu, F., Wushouer, M. (2025). Adaptive multimodal transformer based on exchanging for multimodal sentiment analysis. *Scientific Reports*, 15: 27265. <https://doi.org/10.1038/s41598-025-11848-4>
- [21] Lu, Y., Feng, H. (2025). Multimodal emotion recognition via adaptive high-order transformer network. *PLoS ONE*, 20(10): e0333674. <https://doi.org/10.1371/journal.pone.0333674>
- [22] Zhao, S., Ren, J., Zhou, X. (2025). Cross-modal gated feature enhancement for multimodal emotion recognition in conversations. *Scientific Reports*, 15: 30004. <https://doi.org/10.1038/s41598-025-11989-6>
- [23] Dai, W., Zheng, D., Yu, F., Zhang, Y., Hou, Y. (2025). A novel approach to for multimodal emotion recognition: Multimodal semantic information fusion. *arXiv preprint arXiv:2502.08573*. <https://doi.org/10.48550/arXiv.2502.08573>
- [24] Mirjalili, S., Mirjalili, S.M., Lewis, A. (2014). Grey wolf optimizer. *Advances in Engineering Software*, 69: 46-61. <https://doi.org/10.1016/j.advengsoft.2013.12.007>
- [25] Kennedy, J., Eberhart, R. (1995). Particle swarm optimization. In *Proceedings of ICNN'95 - International Conference on Neural Networks*, Perth, WA, Australia, pp. 1942-1948. <https://doi.org/10.1109/ICNN.1995.488968>
- [26] Zadeh, A., Liang, P.P., Poria, S., Cambria, E., Morency, L.P. (2018). Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, Melbourne, Australia, pp. 2236-2246. <https://doi.org/10.18653/v1/P18-1208>
- [27] Busso, C., Bulut, M., Lee, C.C., et al. (2008). IEMOCAP: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42(4): 335-359. <https://doi.org/10.1007/s10579-008-9076-6>