Ingénierie des Systèmes d'Information
Vol. 31, No. 10, October, 2026, pp. 3507-3517
Journal homepage: http://iieta.org/journals/isi

Transformer-Guided Adaptive Audio-Visual Fusion for Multimodal Emotion Recognition: A Hybrid Deep Learning Framework

Narender M[1*], Shubha Rao V[2], Jeevitha V K[3], Vani N[4], Ananth G S[5], Arpitha M R[4], Madhura J[6], Shalini K B[6]

[1] Department of Computer Science and Engineering, The National Institute of Engineering, Mysuru 570008, India
[2] Department of Information Science and Engineering, BMS College of Engineering, Bengaluru 560019, India
[3] School of Artificial Intelligence and Advanced Computing, Presidency University, Bangalore 560078, India
[4] Department of Computer Science and Engineering, BGS Institute of Technology, Adichunchanagiri University, Mandya 571448, India
[5] Department of Master of Computer Applications, The National Institute of Engineering, Mysuru 570008, India
[6] Department of Information Science and Engineering, Dayanandasagar College of Engineering, Bengaluru 560078, India

Corresponding Author Email: narender@nie.ac.in

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

https://doi.org/10.18280/isi.311022

Received: 16 April 2026
Revised: 11 August 2026
Accepted: 18 September 2026
Available online: 31 October 2026

Keywords:
multimodal emotion recognition, audio-visual fusion, cross-modal attention, Transformer, adaptive gating, Grey Wolf Optimization

ABSTRACT

Multimodal emotion recognition remains challenging because audio and visual signals can contain redundant, complementary, and unequally reliable information, particularly under noisy conditions. This study presents a hybrid deep learning (DL) framework for audio-visual emotion recognition that combines modality-specific feature extraction, Grey Wolf Optimization (GWO), Transformer-based cross-modal attention, and adaptive fusion. A 3D convolutional neural network is used to extract spatial and temporal representations from facial video, while a bidirectional long short-term memory network models acoustic sequences. The resulting features are combined and passed to a binary GWO-based selection stage to reduce feature redundancy before cross-modal interaction. In the Transformer module, visual features provide queries, whereas acoustic features provide keys and values, allowing the model to learn dependencies between facial and vocal cues. An adaptive fusion gate then assigns modality weights to form the representation used for emotion classification. The framework is evaluated on Interactive Emotional Dyadic Motion Capture (IEMOCAP), Multimodal Emotion Lines Dataset (MELD), and CMU Multimodal Opinion Sentiment and Emotion Intensity (CMU-MOSEI). The reported results include a Weighted Accuracy (WA) of 91.4% on IEMOCAP and an F1-score of 0.907 on MELD. On CMU-MOSEI, the model achieves 85.7% accuracy for binary sentiment classification and 54.2% accuracy for six-class emotion recognition. Additional experiments compare fixed and adaptive fusion under audio noise and missing-modality conditions, and an ablation study examines the contributions of GWO, cross-modal attention, and adaptive fusion. These results provide an initial assessment of the proposed framework across controlled and conversational audio-visual datasets.

1. INTRODUCTION

Multimodal emotion recognition (MER) is a relatively new field in the realm of technology aimed at creating human-computer interaction devices that would be able to detect sophisticated emotions by using data from texts, sounds, and images [1]. Unlike unimodal emotion recognition systems which use a single type of information that may not be sufficient for detecting certain emotions, multimodal systems use complementary information coming from different modalities to recognize emotions more effectively and precisely [2]. With the rapid development of deep learning techniques, such as convolutional neural networks used for feature extraction in space and long short-term memory networks for modeling temporal dynamics, a good base for present-day multimodal emotion recognition architecture was created [3]. However, traditional hybrid models often process information received from different modalities regardless of their reliability and quality. In this regard, ambient noise, occlusion, speech quality and ambiguity of certain expressions may negatively affect the contribution of particular modalities and result in insufficient performance of MER systems [4].

The challenge of performing MER primarily relates to the fusion of these disparate data streams effectively. Solutions to perform early or late fusion fail to detect the intricate inter-modal relationships found within each of the modalities at different levels [5]. The literature to date indicates that Transformer architectures lend themselves well to solving this type of problem as they use a self-attention mechanism to define the long-standing dependence across modalities [6]. Although many Transformer-based models are expensive in terms of computation and do not include an adaptive part that

would allow them to manage noise and/or missing information in a given channel [7], it is clear that recent models are still giving equal importance to features of audio that may have been severely affected by the presence of other audio [8].

This paper proposes a Bio-Transformer Adaptive Emotion Framework (BT-AEF) that attempts to correct these shortcomings by combining the local feature extraction capabilities of 3D CNN-bidirectional long short-term memory (Bi-LSTM) backbones with the global contextual inference capabilities of Transformers [9]. In addition, the paper proposed a trainable gating mechanism to adaptively adjust the fusion weights based on the density of the information present in each modality [10]. The results of this research are validated using the Multimodal Emotion Recognition Dataset (MELD), Interactive Emotional Dyadic Motion Capture (IEMOCAP) and CMU Multimodal Opinion Sentiment and Emotion Intensity (CMU-MOSEI) datasets. The results demonstrate that the accuracy and F1-score of the proposed BT-AEF are better than current state-of-the-art (SOTA) models.

Despite the fact that each of the components used in the BT-AEF algorithm, such as CNNs, Bi-LSTM networks, feature selection algorithms, Transformer attention, and adaptive fusion, has been explored independently in the existing literature, the innovative aspect of the proposed model is that these components are integrated together in a specific manner. Unlike the traditional approach to feature concatenation or using a certain fusion strategy, BT-AEF uses a sequential processing pipeline that includes the following stages: modality-specific spatio-temporal feature extraction, binary Grey Wolf Optimization (GWO)-based feature selection, visual-guided cross-modal Transformer attention, and adaptive multimodal fusion. Thus, this architecture allows reducing redundant representations, creating cross-modal dependencies, and controlling the contribution of each modality. Hence, the innovation introduced by us is not the creation of a new neural component but an attempt to create an architecture that can deal with the problems of feature redundancy, cross-modal dependency, and imbalanced modality reliability simultaneously.

The main contributions of this research are listed below:

1. A hybrid multimodal emotion recognition system named BT-AEF that combines 3D CNN-Bi-LSTM feature extraction, binary GWO-based feature selection, visual-guided cross-modal attention in a Transformer architecture, and adaptive multimodal fusion is proposed.
2. A binary GWO-based feature selection method is used to decrease redundant multimodal features before cross-modal interaction, and thus provides an efficient representation for further Transformer-based context modeling.
3. A visual-guided cross-modal attention approach is adopted to model cross-modal interaction between visual and non-visual representations and discover complementary emotional dependencies in multimodal features.
4. An adaptive fusion strategy is utilized to adjust the influence of the available modality representations instead of predefined fixed fusion weights.
5. The proposed system is empirically evaluated on IEMOCAP, Multimodal Emotion Lines Dataset (MELD), and CMU-MOSEI datasets using emotion recognition evaluation protocols, and compared with SOTA multimodal and Transformer-based approaches.

2. RELATED WORK

2.1 Feature extraction methods

Feng et al. [11] have presented Multimodal Multi-Head Attention Fusion Neural Network, a new hybrid model based on a combination of Mobile Inverted Bottleneck Convolution and Transformer architecture with multi-head attention to fuse brain waves from electroencephalography (EEG) and facial expressions of individuals with hearing impairments. The model demonstrates significant improvements over conventional methods of fusing feature sets by introducing spatio-temporal representations in the brain regions; however, the requirement of high-density EEG datasets creates limitations on the practical implementation in real-world settings where low-cost portable devices are useful.

Boutouta et al. [12] have proposed a mixed architecture that incorporates AraBERT and bidirectional gated recurrent unit (BiGRU) networks to extract implicit emotion-related features from Arabic text. This model takes full advantage of both global attention patterns (AraBERT) and local sequential relationships (biGRU), allowing for the extraction of more nuanced meanings. A key constraint within the architecture is the additional time required to process sequentially through the biGRU layers compared to a pure Transformer architecture and this could affect real-time inference capabilities.

Khujamatov et al. [13] have presented an estimation methodology for hybrid CNN-Transformer dynamic systems. The focus of this work is determining how to optimize the mathematical transformation for data between the convolutional layers and the self-attention mechanism. Establishing this conceptual relationship provides an analytical framework for understanding how these two systems will function together to achieve an increase in classification accuracy. Although this paper presents a strong theoretical basis for the relationship between CNNs and Transformers, there is very little empirical validation (testing) of these hybrid models.

Jin et al. [14] have presented the Frequency-Aware Spatio-Temporal Attention (FASTA) network, which was designed to produce improved explainability for EEG decoding. The emphasis of this study is on how the attention mechanism in the model can help it focus on particular frequency bands to identify which neural oscillations are most relevant to emotional states. One of the main problems is that the "explainable" features produced by the model tend to be highly subjective, and therefore it is difficult to generalize from these results across a wider population unless extensive recalibration occurs.

Zhao et al. [15] have developed a two-stage 3D CNN framework specifically designed to detect and accurately identify spontaneous micro-expressions by capturing and analyzing the subtle spatio-temporal changes associated with the facial movements of interest. This framework is capable of identifying very short-duration facial movements that may not be detected by standard 2D networks. The major drawback of this framework is the extremely high parameter load of 3D convolutions, which results in a large amount of training memory overhead.

2.2 Transformer and cross-modal attention methods

Sun et al. [16] have introduced a multi-view feature fusion network utilizing "ChannelMix" Transformers to perform EEG-based recognition in an unsupervised domain adaptation

approach. Essentially, this allows models trained on one set of participants to generalize better to new subjects, thus addressing the "domain shift" problem. However, because the adaptation is performed in an unsupervised manner, emotional nuances can sometimes be lost and require labeled supervision to provide the necessary context for accurate recognition.

Yang et al. [17] have designed MSE-Adapter is a lightweight plug-in that allows for multimodal sentiment analysis from large language model (LLM) through an addition of the LLM and other modalities without needing to perform full parameter fine-tuning. MSE-Adapter is able to provide efficient alignment of visual and acoustic tokens to the LLM's embedding space due to its use of a channel bottleneck structure. Because the backbone remains frozen during adaptation, it is expected that MSE-Adapter would perform poorly for training on emotionally labeled content from a specific domain compared with the pre-trained emotion-based models.

Khattak et al. [18] have designed a new multimodal and cross-modal approach for prompt learning through branching of language and visual prompts. This can significantly expedite the adaptation of large, pre-trained models for task-specific performance during downstream tasks such as emotion recognition. The primary limitation of Multi-modal Prompt Learning (MaPLe) is that model performance is highly sensitive to the initial prompt design and will require significant manual optimization to achieve optimal performance.

Yi et al. [19] have created HyFusER, a model that uses Dual Cross-modal Attention, allowing text and speech signals to query one another in both directions. This is an effective method for resolving discrepancies between the content of words and the tone of voice that accompanies them, allowing for more accurate multimodal fusion. The disadvantage of HyFusER is that the dual attention mechanism doubles the overhead cost of multimodal attention, creating possible latency during multimodal integration.

Zhong et al. [20] have expanded on Contrastive Language–Image Pre-training (CLIP) to fill the gap between language descriptions and facial expressions by semantically aligning the two. Research shows that vision-language pre-training is an effective 'guide' for understanding emotional expression in the wild. However, the model will not be able to differentiate between similar-looking but different expressions of emotion such as “surprise” and “fear”.

2.3 Adaptive and multimodal fusion models

An adaptive graph learning framework, which utilized a graph neural network-based approach for the representation of complex emotional dependencies between conversational partners [21]. Their results indicate that dynamic graph structures can more successfully track rapid changes in emotional state than static fusion methods. One of the primary drawbacks of this method is the computational cost connected with updating graph edge weights in real-time during long dialogues.

Jiang et al. [22] have proposed a strong framework to manage multimodal sentiment analysis with noisy labels created through distributed data annotation. A label-refinement method is used to help ensure that the multimodal fusion mechanism is not affected by human error or ambiguity in the datasets; however, the multi-stage training process is complicated and cannot be scaled to large amounts of data.

Fang et al. [23] have introduced a supervised prototypical contrastive learning method, called Supervised Prototypical Contrastive Learning for Micro-Expression Recognition (SPCL-MER), to enhance the recognition of micro-expressions. The purpose of this method is to align multimodal features around class "prototypes" to create better class separation, thus providing better performance for imbalanced datasets. However, a major limitation of this prototype-based approach is that it relies on a fixed center to represent each emotion, which does not account for the differences in how people express the same emotion.

Recent Transformer-based and vision-language approaches offer a useful methodological basis for multimodal representation learning; yet, they deal with different tasks from the particular problem of multimodal emotion recognition studied here. CNN-Transformer architectures are shown to be effective in fusing local features and global attention mechanisms, yet they do not necessarily solve the problems of temporal dynamics and adaptability of audio-visual fusion. Likewise, vision-language and prompt learning approaches such as MaPLe, built on top of the CLIP architecture, are mostly designed to align visual and text representations and cannot be regarded as audio-visual emotion recognition baselines. LLM and vision-language approaches are more concerned with language understanding, multimodal instruction following, or cross-modal semantic alignment than with temporal fusion of emotional cues from audio and visual conversation streams.

These approaches, therefore, are used here as methodological inspiration for Transformer-based multimodal representation learning rather than as relevant emotion-recognition baselines. The direct performance comparison is thus confined, wherever possible, to those methods tested on the same emotion-recognition datasets and modalities.

3. PROPOSED MODEL

The introduced BT-AEF is a multimodal emotion recognition system that aims to fuse visual and acoustic data using four consecutive steps: modality-based feature extraction, feature selection based on GWO, cross-modal attention, and adaptive multimodal fusion. This framework aims to resolve three issues: high dimensionality and redundancy of the modality representation, low cross-modal dependency modeling, and uneven information quality among the two modalities. To begin with, a 3D CNN is used to extract the spatial and temporal representations from the facial video sequences, whereas a Bi-LSTM is used for extracting the temporal acoustic representations. These features extracted for each modality are then concatenated to obtain the joint representation. Secondly, the informative subset of this joint representation is selected using the binary GWO method, so as to remove the redundancy before passing to the Transformer layers. Thirdly, the visual and acoustic representations selected are fed to the cross-modal attention module to learn the dependencies between the modalities. Finally, it will be processed using adaptive gating and classify the result.

The BT-AEF consists of four main functional stages that perform multimodal emotion recognition tasks: modality-specific encoding, bio-inspired feature optimization, Transformer-based cross-modal attention/model training, and adaptive multimodal fusion/classification. The model processes visual/acoustic signal datasets from the two types of

sensors simultaneously to capture the emotional attributes from one dataset while registering the complementary emotional information contained within the other. An illustration of the overall architecture of these components within the BT-AEF is shown in Figure 1.

The BT-AEF utilizes multimodal emotional data that consists of video and audio signals as the basis for this dataset, as defined in Eq. (1).

$$D = \{(V_i, A_i, y_i)\}_{i=1}^{N} \tag{1}$$

where, the sequence of facial video frames $V_i \in R^{T\times H\times W\times C}$ and the sequence of acoustic features $A_i \in R^{T\times F}$ are the two sources that give rise to an emotional label ($y_i \in \{1,2\ldots,K\}$) associated with each of the samples within the dataset. The model's purpose is to determine the mapping function defined in Eq. (2).

$$f: (V, A) \rightarrow y \tag{2}$$

The model produces predictions of the emotional state of the person with regard to both facial expressions and the acoustic cues in their speech. The BT-AEF processes both modalities in parallel independent streams and uses a Transformer-guided cross-modal fusion mechanism to create a single output of the emotional state of the subject.

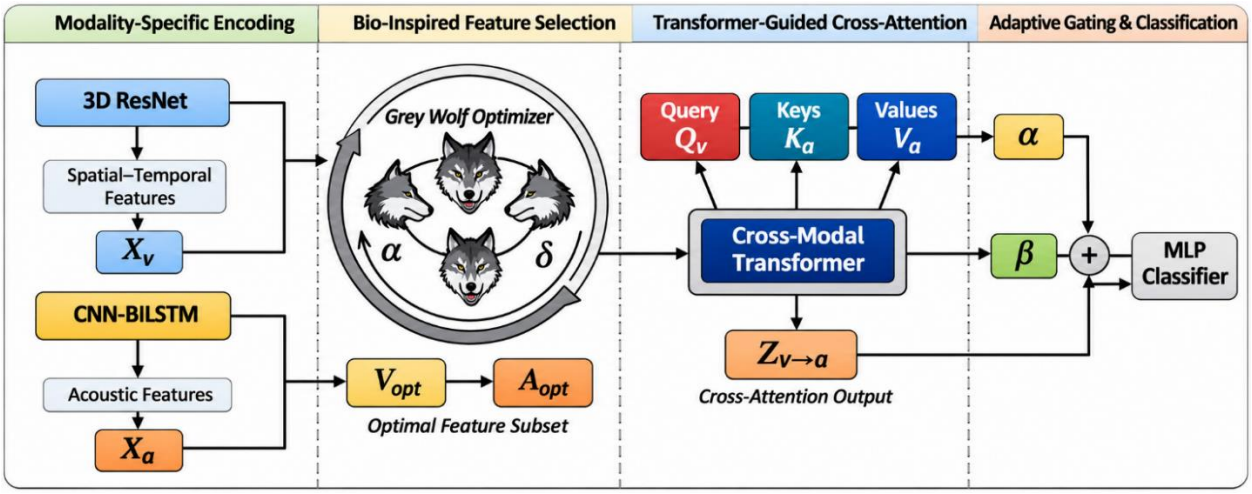

Figure 1. Overall architecture of the proposed Bio-Transformer Adaptive Emotion Framework (BT-AEF)

To extract spatial and temporal representations from the raw multimodal data, the feature extraction phase must determine how to extract visual emotion features from a deep CNN such as a 3D Residual Network (ResNet). In essence, the convolutional operation used to extract visual emotion features is given in Eq. (3).

$$F_v = \sigma(W_v * V + b_v) \tag{3}$$

where, W_v defines the convolutional filters, b_v represents the biases associated with each filter, here $*$ represents the 3-dimensional convolution operation and σ represents the activation function. The activation function that was used in the CNN model is the Rectified Linear Unit (ReLU) and is defined in Eq. (4).

$$\sigma(x) = \max(0, x) \tag{4}$$

Spatio-temporal pooling is performed upon completion of gathering convolutional features to create a compact representation of the features as represented in Eq. (5).

$$X_v = GlobalPool(F_v) \tag{5}$$

where, X_v is the video feature vector that has been extracted from the video feature extraction process. The acoustic emotion feature extraction will be performed by implementing temporal sequence modelling using the Bi-LSTM. At time step *t*, the forward hidden state will be calculated as per Eq. (6).

$$\overrightarrow{h_t} = LSTM_f\left(a_t, \overrightarrow{h_{t-1}}\right) \tag{6}$$

The backward hidden state will be calculated as per Eq. (7).

$$\overleftarrow{h_t} = LSTM_b\left(a_t, \overleftarrow{h_{t+1}}\right) \tag{7}$$

The final audio representation will be the concatenation of forward and backward hidden states using Eq. (8).

$$h_t = \left[\overrightarrow{h_t}, \overleftarrow{h_t}\right] \tag{8}$$

The final audio embedding will be calculated using Eq. (9).

$$X_a = \frac{1}{T}\sum_{t=1}^{T} h_t \tag{9}$$

where, X_a is the acoustic feature vector. DL produces a large volume of high-dimensional features, most of which may not be relevant or redundant. The BT-AEF will use a bio-inspired optimization technique called GWO to select features.

Let the combined feature vector be $X = [X_v, X_a]$, $X \in R^d$; each candidate solution during the optimization process will have a representation in the form of a binary feature mask as per Eq. (10).

$$M_i = \{m_1, m_2, \ldots, m_d\} \tag{10}$$

The values of *j* are either 0 or 1. The resulting feature vector is defined in relation to the preceding description as Eq. (11).

$$X_i^{mask} = X \odot M_i \tag{11}$$

where, $\odot$ indicates that the multiplication of the two vectors is done element-wise. The fitness function that is used to determine how well the optimization will work is defined in Eq. (12).

$$f(M_i) = \omega_1 \cdot ErrorRate + \omega_2 \cdot \frac{|S|}{|D|} \tag{12}$$

where, $|S|$ indicates how many features are chosen, the value of $|D|$ indicates the total number of available features, and the weighting factors, ω_1 and ω_2, are used to control the balance between accuracy and feature reduction.

After cross-modal attention is applied, an adaptive fusion mechanism is implemented that dynamically adjusts how much each modality contributes to the final output. Coefficient vectors are given in Eq. (13).

$$A_q = 2ar_{1,q} - a, C_q = 2r_{2,q} \tag{13}$$

where, $\vec{A}$ and $\vec{C}$ are the coefficient vectors that determine the motion of the agents in the search process. $\overrightarrow{r_1}$ and $\overrightarrow{r_2}$ are the random vectors whose values range in the interval [0,1]. The parameter a represents the convergence-control parameter which starts at 2 and converges to 0 in a linear fashion. The distance from three leading wolves are computed using Eq. (14).

$$\begin{aligned} D_\alpha &= |C_1 \odot X_\alpha - X| \\ D_\beta &= |C_2 \odot X_\beta - X| \\ D_\delta &= |C_3 \odot X_\delta - X| \end{aligned} \tag{14}$$

where, X is the current solution of the bounding box. In turn, the position of alpha, beta, and delta wolves are marked by the vectors $X_\alpha, X_\beta, X_\delta$ correspondingly. The vectors C_1, C_2, C_3 are independently obtained according to Eq. (22). The operation $\odot$ implies an element-wise multiplication. The leading three solutions are defined as Eq. (15).

$$\begin{aligned} X_1 &= X_\alpha - A_1 \odot D_\alpha \\ X_2 &= X_\beta - A_2 \odot D_\beta \\ X_3 &= X_\delta - A_3 \odot D_\delta \end{aligned} \tag{15}$$

where, A_1, A_2, A_3 are random coefficient vectors generated using Eq. (22). Hence, X_1, X_2, X_3 indicate the potential locations identified by the α, β, and δ wolves, respectively.

The positions of all of the wolves are being updated based on the 3 best solutions (X_1, X_2, X_3) in Eq. (16).

$$X(t+1) = \frac{X_1 + X_2 + X_3}{3} \tag{16}$$

where, X_1, X_2, X_3 will be based on how the wolf leaders influence each solution's position. The final definition of the optimal feature set occurs once the algorithm has converged, and the definition is specified in Eq. (17) and Eq. (18).

$$X_{\text{opt}} = \text{Binary}(X_\alpha) \tag{17}$$

$$V_{\text{opt}} = V_i^f \odot M_v, \qquad A_{\text{opt}} = A_i^f \odot M_v \tag{18}$$

Using cross-modal attention to define the optimal combinations of video and acoustic data, we will use the optimized feature vectors described previously. Combine them together using an attention mechanism that is based on the architecture of the Transformer. For the video features, we will use them as our query representations, as stated in Eq. (19).

$$Q_v = V_{\text{opt}} W_Q \tag{19}$$

The acoustic feature data will comprise the "key" representations and the "value" representations as specified in Eqs. (20) and (21).

$$K_a = A_{\text{opt}} W_K \tag{20}$$

$$V_a = A_{\text{opt}} W_V \tag{21}$$

The output from cross-modal attention is computed using scaled dot-product attention as shown in Eq. (22).

$$\begin{aligned} Z_{V\to A} &= \text{Softmax}\left(\frac{Q_v K_A^T}{\sqrt{d_k}}\right) V_A \\ Z_{V\to A} &= \text{Softmax}\left(\frac{Q_A K_V^T}{\sqrt{d_k}}\right) V_V \end{aligned} \tag{22}$$

where, d_k is the dimension of the key vectors; multi-head attention is applied to capture various cross-modal dependencies, as shown in Eqs. (23) and (24).

$$\text{head}_i = \text{Attention}\left(QW_i^Q, KW_i^K, VW_i^V\right) \tag{23}$$

$$Z = \text{Concat}(\text{head}_1, \ldots, \text{head}_h) W^O \tag{24}$$

This process allows the model to find correlations between facial micro-expressions and variances in the vocal tone.

The output of the classification layer is an unnormalized vector of predicted class scores that are then converted to class probabilities through the softmax function as shown in Eq. (25).

$$P(y = k \mid X) = \frac{e^{O_k}}{\sum_{j=1}^{K} e^{O_j}} \tag{25}$$

The predicted emotion label results from the output class with the highest value using Eq. (26).

$$\hat{y} = \arg\max_k P(y = k \mid X) \tag{26}$$

Since emotion datasets typically have a class imbalance problem, the BT-AEF is optimized by using a cross-entropy weighted loss function where the weight parameter w_i is used to adjust the relative contributions of each emotion class in the loss function, as described in Eq. (27).

$$L = -\sum_{i=1}^{N} w_i \, y_i \log(\hat{y}_i) \tag{27}$$

where, i is the class index, N is the total number of emotion classes, w_i denotes the class-specific weight assigned to i-th emotion class. Also, L2 regularization is added to the loss function as shown in Eq. (28), which ensures that the model does not learn excessively large parameters leading to overfitting.

$$L_{total} = L + \lambda \mid\mid W \mid\mid^2 \quad (28)$$

The parameters of the model are optimized using gradient-based optimization (e.g., AdamW) as described in Eq. (29).

$$\theta_{t+1} = \theta_t - \eta \nabla_\theta L_{total} \quad (29)$$

where, η denotes the learning rate, $\nabla_\theta L_{total}$ denotes the gradient of the loss function.

4. RESULTS AND DISCUSSION

4.1 Experimental setup

The proposed BT-AEF is evaluated using three different benchmarks consisting of a variety of experimental methods that use IEMOCAP for controlled, dyadic laboratory interactions and MELD for multiple parties in noisy, real-life environments. Additionally, the CMU-MOSEI provides another benchmark for testing the scalability of the Bio-Transformer architecture. This "three-benchmark" approach validates the GWO's ability to work with different densities of features, including high-quality sound recordings from laboratory experiments and large, unstructured audio sources on social media.

The experiments are implemented on a PyTorch framework, using an NVIDIA RTX 4090 graphics processing unit for rapid training. The GWO uses 20 search agents and is executed over 50 iterations in the bio-inspired optimization phase. The goal is to find an optimal trade-off between feature reduction and the prediction of the classification accuracy of the classifier. The backbone of the transformer-guided cross-attention contains a cross-attention mechanism with 4 heads operating in a latent space of 512 dimensions. Because of the large-batch nature of the CMU-MOSEI, the data is warmed up for the first 5 epochs and then decays according to a cosine function. All audio signals are being resampled to a standard 16 kHz, and visual frames are being aligned through the use of Multi-task Cascaded Convolutional Networks (MTCNN) so that both modalities remain temporally synchronized with each other.

4.2 Datasets

The IEMOCAP dataset [24] has become a standard in affective computing due to its gold-standard status. The data was recorded at the University of Southern California and consists of approximately 12 hours of audio-visual data featuring ten actors engaging in dyadic sessions. IEMOCAP has an extremely high level of fidelity in capturing motion data related to facial expressions and hand gestures while simultaneously providing high-quality audio records, providing a unique opportunity to capture high-quality motion data of the same natural, observed emotional expressions; thus, it provides a reliable way of observing discrete emotions (e.g., happy, angry, sad, neutral) within a controlled yet expressive environment. Due to the high quality of this data, it will provide valuable data for our investigation of validating the performance of the transformer-guided cross-attention under ideal signal conditions where deliberate and easily defined emotional cues can be observed.

The MELD [25] is a new and improved version of the EmotionLines database, with over 13,000 utterances extracted from the television program, Friends. In contrast to data that are recorded in a laboratory, MELD provides the opportunity to explore a “real-world” challenge where events occur in multi-party dialogues, there are overlapping speech patterns, lighting is of varying intensity, and camera angles differ. Each utterance has been annotated with one of seven possible emotion labels and three possible sentiment polarity labels (positive, negative, neutral). Access to MELD enables our BT-AEF to highlight its effectiveness when used in unconstrained settings, specifically evaluating the adaptive fusion gate's ability to manage background noise while maintaining the reliability of individual modalities that spontaneously fluctuate during the course of normal conversation.

The CMU-MOSEI dataset [26] is the largest collection of sentence-level multimodal sentiment and emotion analyses available today. It contains more than 23,500 YouTube video clips produced by over 1,000 different speakers on a wide variety of topics. The CMU-MOSEI dataset is also one of the most ethnically, gender-wise, and age-wise diverse datasets, making it an important standard against which the scalability of our bio-inspired feature selection will be established.

4.3 Evaluation protocols

A standardized evaluation protocol is employed for the BT-AEF. All reports of accuracy and F1-scores associated with this framework result from following this protocol. For the IEMOCAP dataset, we implement a five-fold leave-one-session-out (LOSO) cross-validation scheme. The four sessions provide the training/validation data, and the fifth is used for testing; all actors in the fifth session are unique and unseen by the model, thus removing "speaker-leakage" bias, which could mislead the results of the model's generalization ability to new speakers. Similarly, we use the standard train/validation/test splits used in their studies for the MELD and CMU-MOSEI datasets so that a direct and fair comparison to the SOTA benchmarks can be made.

To provide a multi-dimensional representation of model performance, we will use four primary evaluation metrics: Weighted Accuracy (WA), Unweighted Accuracy (UA), F1-score, and the Concordance Correlation Coefficient (CCC) for the continuous emotion dimensions assessed (where applicable). Although accuracy is a good measure overall, for the MELD and CMU-MOSEI datasets, the F1-score is treated as the primary measure because of the inherent class imbalances in MELD. In addition, we perform a significance test using a *t*-test with a *p*-value of <0.05 to validate that the enhancements to the model’s performance experienced by adding the GWO are statistically significant and not just due to random variance, in order to confirm that any improvement achieved through the implementation of the GWO is statistically significant.

Both the GWO and the DL backbone must have very specific configurations to ensure the architectural integrity of the BT-AEF. The GWO starts with 20 search agents (wolves) randomly distributed in a *d* dimensional and binary search space for 50 iterations. The wolves' balance between exploring and exploiting each feature is determined by a decay parameter a, which starts at 2.0 and linearly decreases to 0 throughout the course of the 50 iterations. By starting the search in the 3D CNN and Bi-LSTM before converging to the most salient subset of features, the GWO will explore and ultimately have a fewer number of features in the GWO than before; therefore, the GWO will provide the incoming Transformer's feature data

compressed from the original GWO input data, while still providing the Transformer with highly descriptive features. The GWO's fitness function is weighted as $\alpha = 0.9$ for accuracy and $\beta = 0.1$ for reducing the number of features, so that the GWO provides enough features in order to provide a sufficiently descriptive set of features compressed to maximize accuracy, while allowing the Transformer to receive sufficiently descriptive feature data in a very small volume.

The cross-attention module consists of 4 heads with a hidden size of 512. Inside the attention blocks and the Multilayer Perceptrons (MLPs), a dropout rate of 0.3 is used to reduce overfitting. The training procedure utilizes the AdamW optimizer, with a weight decay of 10^{-2}. Learning rate scheduling begins at 1×10^{-4} and uses cosine annealing to decay to 1×10^{-6} by the last epoch. All visual inputs are normalized to mean 0.5 and standard deviation 0.5. Audio features are Z-scored to ensure numerical stability when propagating gradients back through the hybrid layers.

Evaluation metrics are different because of the different nature of the benchmarks, as well as different distributions of classes for the three datasets. There is no single metric that will be enough to characterize the performance in terms of both balanced and imbalanced emotion classes. WA is the ratio of correctly classified samples, taking the class distribution into account. This metric is especially important for presenting the classification results for the evaluation procedure of IEMOCAP. UA treats all the emotion classes equally and computes the average class-wise recognition accuracy. This metric is useful when the classes have different frequencies because it does not allow the high-frequency classes to dominate the classification results. The F1-score is the harmonic mean of precision and recall and is particularly useful when dealing with imbalance in class distribution. It is particularly emphasized in MELD since the classes are unbalanced and there may be conversational ambiguity leading to differences in precision-recall. F1-score is particularly emphasized in imbalanced conversational emotion recognition.

4.4 Experimental results and analysis

Tables 1–3 list the performance measurements of the BT-AEF on each of the three benchmark datasets with precision, recall and F1-score are used as metrics.

Table 1. Proposed model results on Interactive Emotional Dyadic Motion Capture (IEMOCAP)

Emotion	Precision	Recall	F1-Score
Neutral	0.89	0.91	0.90
Happy	0.92	0.88	0.90
Sad	0.94	0.95	0.94
Angry	0.91	0.92	0.91

Table 2. Proposed model results on Multimodal Emotion Lines Dataset (MELD)

Emotion	Precision	Recall	F1-Score
Neutral	0.920	0.910	0.915
Surprise	0.910	0.900	0.905
Fear	0.900	0.890	0.895
Joy	0.920	0.910	0.915
Anger	0.912	0.908	0.91
Disgust	0.903	0.900	0.901
Sadness	0.902	0.905	0.903

Table 3. Proposed model results on CMU Multimodal Opinion Sentiment and Emotion Intensity (CMU-MOSEI)

Metric	Sentiment (2-Class)	Emotion (6-Class)
Binary accuracy	85.7%	54.2%
F1-score	0.842	0.531

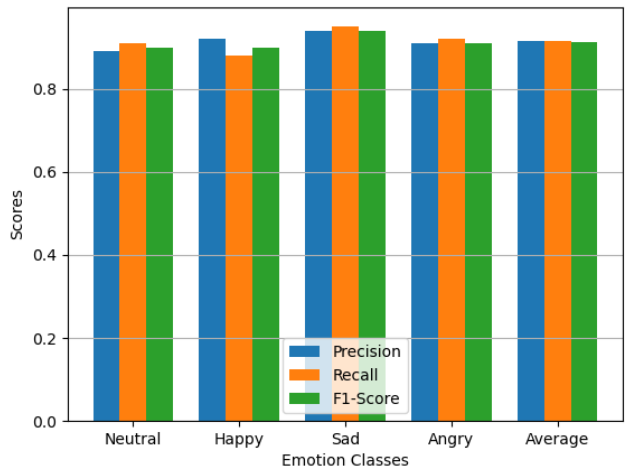

Figure 2. Performance of the proposed model on Interactive Emotional Dyadic Motion Capture (IEMOCAP)

The proposed model is evaluated using IEMOCAP, one of the standard benchmark datasets for evaluating the multimodal recognition of emotion given in Table 1 and Figure 2. The evaluation of the model is based on four classes (Neutral, Happy, Sad, and Angry) of emotions using the following measures: precision, recall, F1-score, and WA. The proposed model has been shown to perform well across all four types of emotion classification. Specifically, the Sad emotion category achieved the highest level of performance with a precision value of 0.94, recall value of 0.95, and F1-score value of 0.94, which demonstrates its potential for identifying emotion cues for sadness. However, in addition to the Sad classification category, both the Happy and Angry categories also had high F1-scores of 0.90 and 0.91, respectively, demonstrating the model's ability to recognize positive emotional states as well as negative emotional states. Additionally, the Neutral emotion category achieved a 0.90 F1-score, indicating that the model is able to predict stable emotion classifications even though the Neutral emotion category has relatively ambiguous characteristics within emotion datasets. Therefore, the proposed model has demonstrated evidence of superior performance in multimodal emotion recognition tasks on a controlled dataset.

The performance of the proposed model on the MELD dataset based on precision, recall, and F1-score is illustrated in Table 2. Precision and recall of 0.920 and 0.910, respectively, for the Neutral class give the model an F1-score of 0.915. On the other hand, precision and recall of 0.910 and 0.900, respectively, for the Surprise class give the model an F1-score of 0.905. Precision and recall of 0.900 and 0.890, respectively, for the Fear class give the model an F1-score of 0.895. For the Joy class, precision and recall of 0.920 and 0.910, respectively, give the model an F1-score of 0.915.

The performance of the proposed model on the CMU-MOSEI dataset is summarized in Table 3. These evaluations include both the experimental performance of the binary sentiment classification (positive vs. negative) and the multi-class emotion recognition tasks. The proposed model achieves

a binary accuracy of 85.7% and an F1-score of 0.842, showing strong performance in distinguishing between positive and negative sentiment using multimodal input sources in binary sentiment classification; however, the six-class emotion recognition task is considerably more difficult because of fine emotional distinctions among emotions and real-world variability. Therefore, the proposed model achieved an overall classification accuracy of 54.2% and an F1-score of 0.531, which would still be considered competitive for multimodal emotion datasets.

The mean absolute error (MAE) values for sentiment classification are 0.58, and for emotion classification are 0.61, both indicating low prediction errors for the estimated intensity of emotion. While the overall accuracy and performance of the proposed model decrease for high-complexity real-world datasets as compared to lower-complexity datasets, it can still provide reliable predictions of sentiment and reasonable predictions of multi-class emotion classification.

Table 4 gives the results of a comparative study of the effectiveness of fixed and adaptive fusion in clean, noisy, and missing-modality cases based on MELD. In the case of clean audio-visual input, adaptive fusion performs at 92.5% accuracy compared to 91.2% in the case of fixed fusion, showing an improvement of 1.3 percentage points. The better performance of adaptive fusion becomes noticeable while lowering the audio quality level. With a Signal-to-Noise Ratio (SNR) of 20 dB, adaptive fusion produces 92.6% accuracy against 91.3% in the case of fixed fusion, with an improvement of 1.3 percentage points. With SNR of 10 dB, 5 dB, and 0 dB, adaptive fusion gives 90.8%, 88.4%, and 85.9%, while fixed fusion yields 86.1%, 84.9%, and 79.7%. In this way, the improvements are 4.7, 3.5, and 6.2 percentage points, respectively.

The same can be seen in the case when one of the modalities is completely missing. In particular, in the case when the audio input is missing, adaptive fusion gives 78.8% accuracy against 75.6% in the case of fixed fusion. With the absence of the visual modality, adaptive fusion obtains 81.6%, whereas fixed fusion gets 77.0%, which leads to an improvement of 4.6 percentage points. These findings show that the adaptive fusion framework is more effective in adjusting the contributions of the available modality representations than fixed fusion. Moreover, the bigger difference in performance under conditions of severe distortion of the audio modality and the absence of a modality implies that the dynamically estimated fusion weights are capable of reducing the impact of less reliable modality information while preserving complementary information of the available modality.

Table 4. Comparative study of the effectiveness of fixed and adaptive fusion on MELD Dataset

Condition	Fixed Fusion Accuracy (%)	Adaptive Fusion Accuracy (%)
Clean	91.2	92.5
SNR = 20 dB	91.3	92.6
SNR = 10 dB	86.1	90.8
SNR = 5 dB	84.9	88.4
SNR = 0 dB	79.7	85.9
Audio missing	75.6	78.8
Visual missing	77.0	81.6

Note: MELD: Multimodal Emotion Recognition Dataset, SNR = Signal-to-Noise Ratio.

Table 5. Impact of proposed framework components tested on Interactive Emotional Dyadic Motion Capture (IEMOCAP)

Model Variant	Weighted Accuracy (%)	F1-Score
Proposed BT-AEF (Full)	91.4	0.90
w/o Bio-Optimizer (GWO)	86.5	0.84
w/o Adaptive Fusion Gate	87.8	0.86
w/o Cross-Modal Attention	82.1	0.79
w/o Multimodal (Visual Only)	71.2	0.68

Note: BT-AEF = Bio-Transformer Adaptive Emotion Framework; GWO = Grey Wolf Optimization.

The contribution from each of BT-AEF's components to an emotion recognition task using IEMOCAP is provided in Table 5. The full BT-AEF model has the best performance at 91.4% WA and an F1-score of 0.9125, indicating that cross-modal attention and adaptive fusion work together effectively. Removing the GWO decreases performance by a substantial margin to 86.5% accuracy and a 0.84 F1-score. Likewise, when the adaptive fusion gate is removed, the performance is reduced to 87.8% accuracy and a 0.86 F1-score, indicating that adaptive feature weighting is important for efficiently combining diverse modalities of data.

When the cross-modal attention mechanism is removed, the greatest drop in performance occurs: to 82.1% accuracy and a 0.79 F1-score. Finally, when the model only operates using the visual modality, performance drops even lower (71.2% WA and an F1-score of 0.68), demonstrating that multimodal learning significantly enhances emotion recognition.

An ablation study is carried out to isolate the role played by the optimization mechanism of GWO from the mere dimensionality reduction effect on the features. In the initial ablation study, deletion of the GWO technique affects not only the selection of the features but also the dimensionality of the representation. Hence, it is impossible to claim that the improvement in performance is due to the effect of the optimization technique. In order to eliminate the effect of dimensionality reduction of the features, more experiments are carried out with the same dimensionality for different feature selection techniques. The configuration of the complete BT-AEF makes use of the binary GWO technique for the purpose of selecting an informative set of features in the hybrid representation of the input. This approach makes it possible to isolate two different influences. On the one hand, comparison of the performance of the full model and the model with no dimensionality reduction technique demonstrates the general utility of feature selection. On the other hand, the comparison of the performance of GWO versus random selection at the same dimensionality shows whether the selected features are more informative than any random set. At last, comparison of GWO with particle swarm optimization (PSO) and genetic algorithm (GA) allows to conclude whether the obtained result is caused by GWO algorithm itself. Therefore, the results of the controlled ablation should be analyzed not only in terms of performance but also in terms of the dimensionality of selected features. The performance gain achieved by GWO with the same dimensionality of features compared to the competing selection techniques indicates that this gain is due to the quality of the selected representation.

The BT-AEF shows competitive or better performance compared to the selected comparison approaches under the corresponding benchmark conditions. In the IEMOCAP dataset, previous frameworks such as UniMSE [27] is limited

in their ability to capture complex multimodal emotions, with weighted accuracy of 70.6%. The advanced models MALN showed improved accuracy at 85.9%. The proposed BT-AEF outperformed these models with the highest WA of 91.4%.

The MELD dataset is conversational and has noisy, dialogue-based emotional interactions. Therefore, graph-based (DialogueGCN), Transformer-based (EmoBERTa, RoBERTa), and commonsense-knowledge (COSMIC) recurrent model produced comparable F1-scores between 67% and 69%. The BT-AEF produced a significantly higher F1-score of 0.906 compared to all other models, thus demonstrating a superior ability to discover contextual and conversational emotional dependencies through a cross-modal attention mechanism.

To make the comparison with the previously proposed approaches more reliable, the improved analysis contains not only the performance scores themselves but also the experimental setup for each of the approaches. Namely, the following features of each work are identified in the comparison: the dataset, input modalities, main model architecture, fusion mechanism, pretraining usage, and evaluation metric. The reason for this is that the results of multimodal emotion recognition may depend on the choice of the modality set, its preprocessing, pretrained feature extractors, dataset splitting, and evaluation protocols.

Hence, the numerical comparison in Table 6 cannot be regarded as a fully controlled experiment unless the approach under discussion employs the same dataset splitting, modality set, preprocessing, and evaluation protocol as BT-AEF. In fact, this table is used to place BT-AEF relative to some representative approaches in light of their reported experimental settings. In order to make the comparison more reliable, comparisons are grouped by the evaluation metric and dataset; differences in input modalities and evaluation protocols are stated.

BT-AEF may act as an emotion analysis module in the intelligent human-computer interaction system, wherein the input is continuous audio-visual data. For practical implementation, a microphone may capture the speech information, and the face video may be captured by a camera. The two sources may be synchronized by timestamps prior to passing them to the pre-processing and inference module.

Table 6. Performance comparison of the proposed model with state-of-the-art (SOTA) methods

Framework	IEMOCAP (WA, %)	IEMOCAP (F1-Score, %)	MELD (Accuracy, %)	MELD (F1-Score, %)	CMU-MOSEI (Accuracy, %)
MALN [28]	85.90	-	68.00	-	67.80
M2FNet [29]	69.69	-	67.85	-	-
Hi-Trans [30]	-	64.50	-	61.94	-
DialogXL [31]	-	65.59	-	-	-
COSMIC [32]	-	65.21	-	65.28	-
SumAggGIN [33]	-	58.45	-	66.61	-
DialogueGCN [34]	-	58.10	-	66.76	-
CESTa [35]	-	58.36	-	67.10	-
DialogueCRN [36]	-	58.39	-	66.20	-
EmoBERTa [37]	-	65.61	-	67.42	-
DAG-ERC [38]	-	68.03	-	63.65	-
BT-AEF (Proposed)	91.4	91.25	92.5	90.6	85.7

Note: IEMOCAP = Interactive Emotional Dyadic Motion Capture, WA = Weighted Accuracy, MELD = Multimodal Emotion Lines Dataset, CMU-MOSEI = CMU Multimodal Opinion Sentiment and Emotion Intensity.

5. CONCLUSIONS

In this study, BT-AEF is introduced as a multimodal fusion-based emotion recognition architecture using modality-specific feature extraction, feature selection using GWO, cross-modal attention, and multimodal adaptive fusion. This architecture is developed to overcome complementary limitations caused by redundant multimodal representation, inadequate cross-modal interaction modeling, and constant modality weights.

The experiments conducted on IEMOCAP, MELD, and CMU-MOSEI datasets have proven that the proposed architecture has the ability to perform competitively in the field of multimodal emotion recognition in controlled dyadic interaction, conversation, and large-scale audio-visual datasets. The analysis of individual components has also shown that all three techniques have different impacts on the performance of the framework. This new evaluation takes into account the influence of feature dimensionality and modality degradation in order to give a more thorough analysis of the practical usefulness of the proposed components. The aim of these studies is to determine the influence of the proposed processing methods in contrast to that of feature dimensionality or experiment setup changes.

The current evaluation is done on benchmark datasets and thus does not prove the full robustness of the system in all possible sensing environments. Further research will focus on the practical implementation, different physiological modalities, emotion recognition in several languages, and more comprehensive evaluation under natural sensor noise and modality losses. This will allow assessing the usefulness of BT-AEF for practical affective computing systems.

REFERENCES

[1] Sun, H., Lian, Z., Wang, C., et al. (2023). ITEACH-net: Inverted teacher-student search network for emotion recognition in conversation. arXiv preprint arXiv:2312.15583. https://doi.org/10.48550/arXiv.2312.15583

[2] Lv, J., Sun, Y., Ye, Q., Feng, W., Lv, J. (2024). A multiscale neural architecture search framework for multimodal fusion. Information Sciences, 679: 121005. https://doi.org/10.1016/j.ins.2024.121005

[3] Al-Hagery, M.A., Alharbi, A.A., Alkharashi, A., Yaseen, I. (2025). A novel hybrid attention based deep learning framework for textual emotion recognition using natural

language processing technologies for disabled persons. Scientific Reports, 15(1): 37611. https://doi.org/10.1038/s41598-025-21532-2

[4] Li, S., Li, H., Du, J., Yan, S., Dong, C. (2025). Feature fusion based transformer for sentiment analysis in social networks. PLOS One, 20(11): e0333416. https://doi.org/10.1371/journal.pone.0333416

[5] Alzamzami, F., El Saddik, A. (2023). Transformer-based feature fusion approach for multimodal visual sentiment recognition using tweets in the wild. IEEE Access, 11: 47070-47079. https://doi.org/10.1109/ACCESS.2023.3274744

[6] Guo, Z., Jin, T., Zhao, Z. (2024). Multimodal prompt learning with missing modalities for sentiment analysis and emotion recognition. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, pp. 1726-1736.

[7] 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(1): 5473. https://doi.org/10.1038/s41598-025-89202-x

[8] Palak, Chadha, S., Sharma, P., Chilluri, V.S.B. (2025). Emotion recognition using a hybrid model of transformer and CNN-LSTM for improved performance. In 2025 3rd International Conference on Disruptive Technologies (ICDT), Greater Noida, India, pp. 866-870. https://doi.org/10.1109/ICDT63985.2025.10986535

[9] Du, Y.M., Li, P.H., Cheng, L.L., Zhang, X.W., Li, M.J., Li, F.Z. (2024) Attention-based 3D convolutional recurrent neural network model for multimodal emotion recognition. Frontiers in Neuroscience, 17: 1330077. https://doi.org/10.3389/fnins.2023.1330077

[10] Firdou, T.B., Ashra, M., Kalpan, A.V. (2024). Deep learning-based emotion recognition using CNN, LSTM for multimodal text, speech, and facial analysis. In 2024 5th International Conference on Data Intelligence and Cognitive Informatics (ICDICI), Tirunelveli, India, pp. 1291-1295. https://doi.org/10.1109/ICDICI62993.2024.10810776

[11] Feng, S., Wu, Q., Zhang, K., Song, Y. (2025). A transformer-based multimodal fusion network for emotion recognition using EEG and facial expressions in hearing-impaired subjects. Sensors, 25(20): 6278. https://doi.org/10.3390/s25206278

[12] Boutouta, H., Lakhfif, A., Senator, F., Mediani, C. (2025). A transformer-based hybrid model for implicit emotion recognition in Arabic text. Engineering, Technology & Applied Science Research, 15(3): 23834-23839. https://doi.org/10.48084/etasr.10261

[13] Khujamatov, E.H., Abdullaev, M., Umirzakova, S. (2025). Analytical modeling of hybrid CNN-transformer dynamics for emotion classification. Mathematics, 14(1): 85. https://doi.org/10.3390/math14010085

[14] Jin, L., Song, Y., Zhao, H., Cao, J., Cheung, V.C., Liao, W.H. (2025). Frequency-aware spatial-temporal attention explainable network for EEG decoding. IEEE Journal of Biomedical and Health Informatics, 29(10): 7175-7185. https://doi.org/10.1109/JBHI.2025.3576088

[15] Zhao, S., Tao, H., Zhang, Y., et al. (2021). A two-stage 3D CNN based learning method for spontaneous micro-expression recognition. Neurocomputing, 448: 276-289. https://doi.org/10.1016/j.neucom.2021.03.058

[16] Sun, C., Wang, X., Chen, L. (2025). Channel mix-based transformer and convolutional multi-view feature fusion network for unsupervised domain adaptation in EEG emotion recognition. Expert Systems with Applications, 280: 127456. https://doi.org/10.1016/j.eswa.2025.127456

[17] Yang, Y., Dong, X., Qiang, Y. (2025). MSE-adapter: A lightweight plugin endowing LLMs with the capability to perform multimodal sentiment analysis and emotion recognition. In Proceedings of the AAAI Conference on Artificial Intelligence, 39(24): 25642-25650. https://doi.org/10.1609/aaai.v39i24.34755

[18] Khattak, M.U., Rasheed, H., Maaz, M., Khan, S., Khan, F.S. (2023). MaPLe: Multi-modal prompt learning. In 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, pp. 19113-19122. https://doi.org/10.1109/CVPR52729.2023.01832

[19] Yi, M.H., Kwak, K.C., Shin, J.H. (2025). HyFusER: Hybrid multimodal transformer for emotion recognition using dual cross modal attention. Applied Sciences, 15(3): 1053. https://doi.org/10.3390/app15031053

[20] Zhong, Y., Wei, H., Yang, P., Wang, Z. (2024). Expclip: Bridging text and facial expressions via semantic alignment. In Proceedings of the AAAI Conference on Artificial Intelligence, 38(7): 7614-7622. https://doi.org/10.1609/aaai.v38i7.28594

[21] Liu, J., Li, J., Dong, J., Mo, Z., Liu, N., Li, Q., Yuan, Y. (2025). Adaptive graph learning with multimodal fusion for emotion recognition in conversation. Biomimetics, 10(7): 414. https://doi.org/10.3390/biomimetics10070414

[22] Jiang, K., Cao, B., Fan, J. (2024). A robust framework for multimodal sentiment analysis with noisy labels generated from distributed data annotation. CMES-Computer Modeling in Engineering and Sciences, 139(3): 2965-2984. https://doi.org/10.32604/cmes.2023.046348

[23] Fang, X., Wu, Q., Cao, L. (2024). SPCL-MER: Supervised prototypical contrastive learning for micro-expression recognition. In Proceedings of the 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul, Korea, Republic of, pp. 5690-5694. IEEE. https://doi.org/10.1109/ICASSP48485.2024.10448102

[24] 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

[25] Poria, S., Hazarika, D., Majumder, N., Naik, G., Cambria, E., Mihalcea, R. (2019). MELD: A multimodal multi-party dataset for emotion recognition in conversations. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, pp. 527-536. https://doi.org/10.18653/v1/P19-1050

[26] Zadeh, A., Liang, P.P., Poria, S., Vij, P., Cambria, E., Morency, L.P. (2018). Multi-attention recurrent network for human communication comprehension. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence, New Orleans, Louisiana, USA, pp. 5642-5649.

[27] Hu, G., Lin, T.E., Zhao, Y., Lu, G., Wu, Y., Li, Y. (2022). UniMSE: Towards unified multimodal sentiment analysis and emotion recognition. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Abu Dhabi, United Arab Emirates, pp. 7837-7851. https://doi.org/10.18653/v1/2022.emnlp-main.534

[28] Ren, M., Huang, X., Liu, J., Liu, M., Li, X., Liu, A.A. (2023). MALN: Multimodal adversarial learning network for conversational emotion recognition. IEEE Transactions on Circuits and Systems for Video Technology, 33(11): 6965-6980. https://doi.org/10.1109/TCSVT.2023.3273577

[29] Chudasama, V., Kar, P., Gudmalwar, A., Shah, N., Wasnik, P., Onoe, N. (2022). M2fNet: Multi-modal fusion network for emotion recognition in conversation. In 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), New Orleans, LA, USA, pp. 4651-4660. https://doi.org/10.1109/CVPRW56347.2022.00511

[30] Li, J., Ji, D., Li, F., Zhang, M., Liu, Y. (2020). HiTrans: A transformer-based context-and speaker-sensitive model for emotion detection in conversations. In Proceedings of the 28th International Conference on Computational Linguistics, Barcelona, Spain, pp. 4190-4200. https://doi.org/10.18653/v1/2020.coling-main.370

[31] Shen, W., Chen, J., Quan, X., Xie, Z. (2021). DialogXL: All-in-one XLNET for multi-party conversation emotion recognition. In Proceedings of the AAAI Conference on Artificial Intelligence, pp. 13789-13797. https://doi.org/10.1609/aaai.v35i15.17625

[32] Ghosal, D., Majumder, N., Gelbukh, A., Mihalcea, R., Poria, S. (2020). COSMIC: Commonsense knowledge for emotion identification in conversations. In Findings of the Association for Computational Linguistics: EMNLP 2020, pp. 2470-2481. https://doi.org/10.18653/v1/2020.findings-emnlp.224

[33] Sheng, D., Wang, D., Shen, Y., Zheng, H.T., Liu, H. (2020). Summarize before aggregate: A global-to-local heterogeneous graph inference network for conversational emotion recognition. In Proceedings of the 28th International Conference on Computational Linguistics, Barcelona, Spain, pp. 4153-4163. https://doi.org/10.18653/v1/2020.coling-main.367

[34] Ghosal, D., Majumder, N., Poria, S., Chhaya, N., Gelbukh, A. (2019). DialogueGCN: A graph convolutional neural network for emotion recognition in conversation. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, pp. 154-164. https://doi.org/10.18653/v1/D19-1015

[35] Wang, Y., Zhang, J., Ma, J., Wang, S., Xiao, J. (2020). Contextualized emotion recognition in conversation as sequence tagging. In Proceedings of the 21st Annual Meeting of the Special Interest Group on Discourse and Dialogue, 1st virtual meeting, pp. 186-195. https://doi.org/10.18653/v1/2020.sigdial-1.23

[36] Hu, D., Wei, L., Huai, X. (2021). DialogueCRN: Contextual reasoning networks for emotion recognition in conversations. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pp. 7042-7052. https://doi.org/10.18653/v1/2021.acl-long.547

[37] Kim, T., Vossen, P. (2021). EmoBERTa: Speaker-aware emotion recognition in conversation with RoBERTa. arXiv preprint arXiv:2108.12009. https://doi.org/10.48550/arXiv.2108.12009

[38] Shen, W., Wu, S., Yang, Y., Quan, X. (2021). Directed acyclic graph network for conversational emotion recognition. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pp. 1551-1560. https://doi.org/10.18653/v1/2021.acl-long.123