IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Journal Européen des Systèmes Automatisés
Vol. 59, No. 1, January, 2026, pp. 275-284
Journal homepage: http://iieta.org/journals/jesa

Robust Emotion Recognition Multimodal Using an Optimized Cross-Modal Data Fusion Framework

Monika Sharma D[1*], Sonika Sharma D[2], Mohanish B M[3], Shashank Dhananjaya[4], Reshma J[5], Dhruva M S[6], Chaithra C P[6]

[1] Department of Electronics and Communication Engineering, B.M.S. College of Engineering, Bengaluru 560019, India
[2] Department of Computer Science and Engineering, B.M.S. College of Engineering, Bengaluru 560019, India
[3] Department of Artificial Intelligence and Machine Learning, BNM Institute of Technology, Bangalore 560070, India
[4] Department of Information Science and Engineering, The National Institute of Engineering, Mysuru 560008, India
[5] Department of Information Science and Engineering, Dayanandasagar College of Engineering, Bangalore 560078, India
[6] Department of Computer Science and Engineering, BGS Institute of Technology, Adichunchanagiri University, BG Nagara 571448, India

Corresponding Author Email: druvams5@gmail.com

https://doi.org/10.18280/jesa.590124

Received: 26 October 2025
Revised: 23 December 2025
Accepted: 30 December 2025
Available online: 31 January 2026

Keywords:
attention mechanism, cross-modal fusion, Convolutional Neural Networks, deep learning, Multimodal Emotion Recognition, Transformer Networks

ABSTRACT

The challenges associated with finding emotions in multimodal environments include heterogeneous data representations, cross-modal redundancy, and dynamic temporal dependencies. To address these challenges, we propose an optimized fusion-based cross-modal architecture for multimodal emotion recognition (OF-CMAER-CAT), which incorporates Convolutional Neural Networks (CNN), attention mechanism, and Transformer Networks is proposed in this paper. The proposed framework employs CNN-based modality-specific encoders to extract spatially discriminative features and uses an attention-based cross-modal fusion module to adaptively highlight significant inter-modal correlations while filtering out irrelevant and redundant information. The approach also utilizes Transformer Networks to appropriately model long-range temporal and contextual dependencies present in multimodal emotional signals. Additionally, the entire architecture is jointly optimized utilizing a single learning objective to produce robust and scalable emotion recognition systems. A variety of performance evaluations were completed utilizing several standard Multimodal Emotion Recognition (MER) datasets, and the proposed OF-CMAER-CAT solution exhibited a 94.1% average accuracy, a 93.6% F1 score, and a 94.8% precision when compared to other state-of-the-art fusion-based systems. The results validate that the OF-CMAER-CAT architecture is suitable for real world applications related to human-computer interaction (HCI) and affective computing.

1. INTRODUCTION

Emotion recognition is becoming one of the most important research areas within the fields of affective computing and HCI; this can be seen in the wide array of applications available for intelligent assistants, mental health evaluations, social robotics, and multimedia comprehension. Multimodal emotion recognition (MER) uses a variety of heterogeneous sources (speech, facial expressions, textual clues, etc.) to obtain more accurate indicators of emotional status than unimodal approaches. However, modeling the complex interactions between different modalities, temporal dependencies, and noise inherent to each modality remains one of the significant challenges faced by researchers in this area.

While many recent papers on multimodal learning have concentrated on the alignment of sequences and learning representations of them. Deep Learning models for emotion recognition using multimodalities [1] were discussed in few literatures. Few researchers demonstrated that paradigm of using transformers captures relationships across modalities in both temporally and semantically relevant contexts. From this initial research, several recent reviews have outlined how dominant the use of transformer-based techniques has become within MER, but also highlighted the remaining technical challenges faced by researchers regarding the optimization of fusion strategies and handling computational complexity [2]. In addition to transformer-based techniques, contrastive learning techniques are shown to help improve cross-modal semantics, though these techniques often require large amounts of data and careful selection of sampling strategies [3].

MER has been studied using fusion-centric models. Fusion-based methods, such as the graph convolutional networks architecture [4], leverage graph structures to capture intermodal relationships while also introducing an extra layer of complexity due to their reliance on multiple models. Low-rank fusion techniques such as LMF [5] and MM-LMF [6] reduce the number of parameters and computational load

needed for modeling multimodal interactions. However, they still have limitations when trying to accurately model the complex nature of emotions as a result of their reduced representational capabilities.

Dynamic fusion networks [7-8] allow for the adaptive weighting of modalities based on temporal cues; however, they require the specification of fusion heuristics that limit their ability to generalize to other situations. Newer modality-specific representation learning and hierarchical message-passing have also provided an alternative means of retaining emotional semantics better than previously discussed methods. While self-supervised multi-task learning frameworks, such as the study [9], provide enhanced robustness for features, they also add complexity to the training process. On the other hand, hierarchical multimodal message-passing models utilize increased intermodal communications but lack the scalability necessary for practical applications. Hierarchical multimodal message-passing models [10] allow for greater cross-modal communication, but they sacrifice the potential for scalability. Overall, the benefits of adaptive fusion and the need to address the ongoing challenge of developing a single, unified model capable of jointly optimising spatial feature extraction, cross-modal attention and temporal modelling merits continued research in this area.

Overall, these methods demonstrate the need for adaptive fusion and highlight the continued limitations of current technology regarding the joint optimization of spatial feature extraction, cross-modal attention, and temporal modeling.

In response to these issues and challenges, we propose OF-CMAER-CAT based on our findings: the Optimized Fusion–Based Cross-Modal Architecture for MER. OF-CMAER-CAT employs CNNs as spatial encoders, an attention mechanism for cross-modal fusion, and Transformative networks to model both temporal and contextual information. The primary difference between this work and previous works lies in the joint optimization of both the individual representational characteristics for each of the modalities and also the interactions that occur between them via an adaptive-fusion scheme in a manner that allows it to work very well for emotion identification in very diverse conditions as well as in very noisy and heterogeneous settings. OF-CMAER-CAT has been evaluated through extensive experimentation and performance evaluation, and the results have shown that OF-CMAER-CAT outperforms existing fusion-based and transformer-based MER models.

The main contributions of this work are summarized as follows:

- Designing an OF-CMAER-CAT, an optimized fusion–based cross-modal architecture that seamlessly integrates CNN-based spatial feature extraction, attention-driven cross-modal interaction modeling, and Transformer-based temporal context learning for robust MER.
- An adaptive cross-modal attention mechanism is introduced to dynamically weight inter-modal dependencies, effectively suppressing modality-specific noise and redundancy while enhancing emotionally salient features across modalities.
- Unlike existing fusion-centric or transformer-only approaches, the proposed model jointly optimizes modality-specific encoders, fusion layers, and temporal modeling components within a unified end-to-end training paradigm.
- Extensive experiments conducted on benchmark MER datasets demonstrate that OF-CMAER-CAT consistently outperforms state-of-the-art methods in terms of accuracy, precision, recall, and F1-score.

The rest of the paper is structured as follows. Section II provides the existing literature. Section III describes the presented OF-CMAER framework. Section IV reports the experimental details, datasets, preprocessing, and evaluation metrics. Lastly, Section VII summarizes the paper and implications for future research directions.

2. LITERATURE SURVEY

The development of MER is comprised of a multitude of interdependent research streams. It includes architectural design, cross-modal transformer (Attention-based) model, contrastive models for cross-modal alignment, conversational & graph-aware models, robustness, missing-modality handling, and lightweight techniques. Based on the survey results we summarized key papers to highlight the strengths and weaknesses of each research area in order to justify our proposed optimized cross-modal fusion framework.

Recently, there have been advancements made in MER methods towards implementing transformer-based methods, such as adaptive attention mechanisms, to model the complexity of the relationship between individuals and their emotions across modalities. The use of transformer-centric architectures has illustrated their capacity to capture expansive temporal and contextual relationships across heterogeneous data sets (modalities) over long distances. For example, in the study [11], MemoCMT is a cross-modal transformer that implements memory-guided attention to merge multiple modalities' features, resulting in increased robustness in recognizing emotional states from individuals. Similarly, Liu et al. [12] developed a model called TACFN, which used Transformer Blocks to adjust cross-modal fusion weights adaptively over time for increased generalizability, although this increases computational demands on the model.

As evidenced in many studies focused on joint multi-modal transformers for real-world emotion recognition, the use of Joint Multi-modal transformers has increased robustness to noise and misalignment of modalities using Evaluative In the wild datasets used by Waligora et al. [13]. In addition, transformer-based approaches to fusing Physiological Signals have also been done, where Feng et al. [14] fused EEG or Brain signals and Face Expressions to recognize emotions from subjects with hearing loss. Also, Ma et al. [15] fused micro expressions with EEG for detecting emotions that were not consciously stated. While these methodologies have proven to be effective, many of their approaches have been limited by the modality-specific nature of the methodologies and the inability to create Scalable Fusion Strategies.

Apart from the transformer-related literature, several variants of disentangled-based and attention-based techniques for fusing different types of data have been examined. Mahaseni and Khan [16] built on the concept of multimodal variational autoencoders that are both private keep modality-specific representations, allow sharing of representations and integrated them with LSTMs. It improved interpretability of results but at a higher cost of training complexity. Recursive attention strategies have also been studied using recursive joint cross-modal attention to continuously update multimodal representations, but they come at the cost of longer inference

time [17].

Wang et al. [18] was created MilMER, a machine learning method for Resolving Temporal Ambiguities through Multiple Instance Learning, so that researchers do not need to give researchers full instance-level annotations on multimodal sequences. Researchers have also adopted graph-based learning techniques and contrastive learning methods to create a body of work around emotion recognition from large datasets. Li et al. [19] have proposed framework combines joint modality fusion and graph-based contrastive learning in order to create discriminative representations across different modalities. However, the process of building the underlying graph for use with Joyful comes with additional costs associated with graph building (overhead). Prompt and adaptive learning methods have also been the focus of some researchers, such as Wafa et al. [20]. These authors demonstrated that combining prompt engineering with deep adaptive methods enhances the ability to scale up for the recognition of emotions from big data datasets.

Real-time and conversational emotional recognition has been explored using gated and model level fusion. A real-time audio-visual-text fusion framework has been developed for environmental emotional recognition by Gupta et al. [21]. On the other hand, a Cross-Modal Gated Feature Enhancer method for emotional recognition in conversation has been proposed by Zhao et al. [22]. Wang et al. [23] have examined linkages between transformer augmentation and modelling for speech emotional recognition, which is enabling improved temporal modelling. The earlier Multimodal attention-based frameworks [24, 25] as well as other domain-specific applications, including driver emotional recognition [26] and audio-video synergy models [27] provides a variety of fusion approaches. Exploratory studies continue to verify the advantages of multimodal learning across speech, text and facial modalities [28].

While these approaches demonstrate how transformers, attention mechanisms and adaptive fusion methods have been able to effectively achieve multimodality in emotional recognition, the majority of current methods are focused on a narrow range of modal combinations, primarily involve computationally expensive fusion processes and do not provide for unified optimization of the fusion process across multiple dimensions like spatial, cross-modal, and temporal. As such, there is an opportunity to develop a new method that would provide for a unified scalable architecture (OF-CMAER-CAT) that would integrate CNN-based feature extraction, attention-based optimized fusion, and transformer-based temporal modelling.

3. PROPOSED MODEL

OF-CMAER-CAT is an optimized fusion-based cross-modal architecture to develop robust and accurate MER using CNN, attention and transformers. The model will be designed to ensure effective fusion of heterogeneous modalities, which include visual facial cues, speech signals, and physiological/textual information, while minimising the effects of modal imbalances, temporal misalignments, noise in the data and cross-modal semantic inconsistencies. In contrast to standard early or late fusion methodologies, OF-CMAER-CAT provides an architecture that combines CNN feature extraction, cross-modal attention alignment, and transformer contextual modeling to provide for adaptive learning of inter-modal dependencies and dynamic optimization of the input contribution of each modal for fusion.

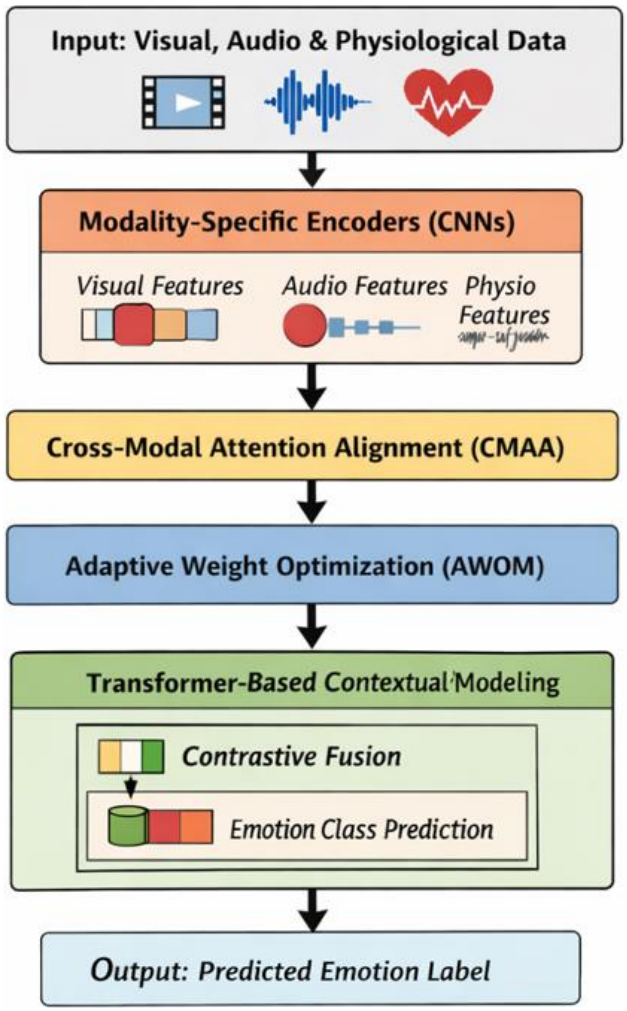

Figure 1. Design diagram of proposed OF-CMAER-CAT framework

The overall design shown in Figure 1 is comprised of a few key elements as discussed below. The proposed Optimized Fusion-Cross-Modal Architecture for Emotion Recognition using CNN, Attention, and Transformer Networks (OF-CMAER-CAT) employs a fully integrated process for learning noise-resilient emotional representations from heterogeneous multimodal data streams. First, independent multimodal raw inputs of visual frames, audio signals, and physiological signals are processed separately through CNN-based modality-specific encoders, which extract both high-level temporal and semantic representations as feature embeddings. The encoders reduce dimensionality and noise while preserving modality properties. After getting the extracted embedding sequences, the Cross-Modal Attention Alignment (CMAA) module aligns the heterogeneous modalities in a common semantic space with scaled dot-product attention, which allows attention given to each modality by the remaining modalities based on emotion relevance.

After alignment, the Adaptive Weight Optimization Module (AWOM) utilizes a set of statistical feature descriptors to produce an estimate of the reliability of each modality. It normalizes the estimated reliability score using softmax normalization to learn the normalized weights for the modalities. This process indicates how much informative modalities contribute to the fused representation compared to

the contribution of unreliable or noisy modalities. In addition to this inter-modality consistency enhancement, a contrastive learning objective is utilized that pulls semantically similar modality embeddings together in the latent space while pushing semantically dissimilar modality embeddings apart.

Next, the adaptively weighted and contrastively aligned features are input into a Transformer-based contextual modeling block that combines multi-head self-attention to capture long-range temporal dependencies as well as cross-modal contextual interactions. In doing so, this helps to improve emotion discrimination by modeling small temporal transitions between different emotions. Finally, after the fused representation has been processed through a feed-forward classification head, emotion categories will be predicted based on the output of a softmax layer. To keep the entire network stable and generalizable across multiple datasets, the joint training process uses a multi-task loss function that contains: classification loss, contrastive loss, reconstruction loss, entropy regularization, and weight decay.

To extract meaningful features from different sources of information, each of these sources of input is processed separately by a dedicated Convolutional Neural Network encoder. CNN has been shown to effectively capture not only local spatial characteristics but also temporal variations and modality-specific emotional indicators. A compact latent embedding representation, while at the same time preserving the important emotional attributes contained in the emotion dataset, is achieved through the use of an Encoder that takes raw input from the various types of data collected from a sensor. Redundancy in the data and the amount of processing that goes into fusing all the data from each sensor will be minimised through use of the Global Pooling technique to create two-dimensional (2D) representations of the sensor's data, providing retrospective views of both the time course of the data and its spatial distribution. These 2D representations will serve as an input to the next components of the algorithm for fusing modalities together, as defined by Eq. (1).

$$\mathrm{F}_i \;=\; E_i(M_i;\theta_i), \mathrm{F}_i \in R^{T_i \times d} \tag{1}$$

Here, M_i is the raw input for modality i, E_i is the encoder for modality i with a set of parameters θ_i, T_i is the number of time points for modality i, and d is the number of dimensions in the latent (hidden) feature space. The resulting encoded feature for each modality is represented by the matrix F_i in $R^{T_i X\, d}$, where the encoding of F_i is typically a large number of-dimensional time series. A global average embedding of F_i into a single vector $\bar{\mathrm{f}}_i$ is computed using Eq. (2), which simplifies the calculations performed during the process of fusing each modality, and provides a mechanism to establish optimal alignment between the global vectors obtained from each modality.

$$\bar{\mathrm{f}}_i \;=\; \frac{1}{T_i}\sum_{t=1}^{T_i} \mathrm{F}_i[t, R^{T_i}] \tag{2}$$

Here, temporal feature vector represented by $Fi[t;]$ at time t, number of time steps, T_i, and pooled embedding of modality i or $Fi[R(Ti);]$ are used to compute the query, key and value matrices for the cross-modality attention by utilizing linear projections through the attention mechanism. Through this process, the model can find relational contextual connections between modalities and gather supportive features from other modalities using Eq. (3).

$$Q_i \;=\; \mathrm{F}_i W_Q, K_i \;=\; \mathrm{F}_i W_K, V_i \;=\; \mathrm{F}_i W_V \tag{3}$$

Here, F_i is a matrix containing the encoded features and ($\mathrm{W_Q}$, $\mathrm{W_K}$, $\mathrm{W_V}$) is an R^{dXdK} type learnable projection at attention dimension dk. The matrices $\mathrm{Q_i}$, $\mathrm{K_i}$, and $\mathrm{V_i}$ are of the same dimension as $\mathrm{W_Q}$, $\mathrm{W_K}$, and $\mathrm{W_V}$, respectively, or R^{dXdK}. Scaled Dot-Product Attention computes a score of attention between modality i and modality j to allow modality i to focus its attention only on pertinent features for modality j. Hence, the interaction of modalities and their connections to the emotions will be enhanced through Eq. (4).

$$\mathrm{Attn}_{i\leftarrow j} \;=\; \mathrm{softmax}\left(\frac{Q_i K_j^{\top}}{\sqrt{d_k}}\right) V_j \tag{4}$$

Here, Query, Key, Value (Q_i, K_j, V_j), attention dimension $\mathrm{d_k}$, and $Softmax(.)$ are used to normalise similarity scores (cues) in attention-weighted representations $\big(Attn, (i,j)\big)$.

The attention is directed toward capturing complementary aspects of emotions between audio stimulus (e.g. talking) and visual stimulus (e.g. gaze) while filtering out irrelevant and contrary emotional stimulus (e.g. depressed, happy, angry) that occur within one or both attributes. By employing the CMAA mechanism, the model enhances the degree of accuracy and reliability with which it can produce representations of emotions, thus increasing the level of cross-modal cohesion and stability. The attentional outputs of each modality partner are aggregated into a single, aligned feature vector $\tilde{\boldsymbol{F}}_i$ during the Cross-Modal Aggregation process. The result is a complete integration of all three modalities' complementary information as represented in Eq. (5).

$$\tilde{\boldsymbol{F}}_i \;=\; \sum_{j\neq i} \alpha_{ij}\, Attn_{i\leftarrow j} \tag{5}$$

Here, $\boldsymbol{Attni}\leftarrow j$ is the attention output from modality j to modality i, αij is the normalized weight for modality j with respect to modality i, and $\tilde{\boldsymbol{F}}_i$ is an R^($T_i \times d$) aligned feature. The Attention Coefficients calculation provides weights for all partner modalities in order to assign importance to the most similar modality to the target modality as shown in Eq. (6). This is the basis for adaptive fusion in cross-modal interactions.

$$\alpha_{\mathrm{ij}} \;=\; \frac{\exp\left(\phi\left(\bar{\mathrm{f}}_{\mathrm{i}}, \bar{\mathrm{f}}_{\mathrm{j}}\right)\right)}{\sum_{\mathrm{k}\neq\mathrm{i}} \exp\left(\phi(\bar{\mathrm{f}}_{\mathrm{i}}, \bar{\mathrm{f}}_{\mathrm{k}})\right)}, \phi(\mathrm{u},\mathrm{v}) = \mathrm{u}^{\top} W_{\phi} \mathrm{v} \tag{6}$$

Here, $\bar{\mathrm{f}}_{\mathrm{i}}$ *and* $\bar{\mathrm{f}}_{\mathrm{j}}$ are global embeddings; W_ϕ is a learnable weight matrix; $\phi(u,v)$ is the compatibility function returning a scalar similarity score; and α_{ij} is normalized across all partner modalities. The Gated Residual Fusion method allows for cross-modal features $\tilde{\boldsymbol{F}}_i$ to be combined with the original modality $\boldsymbol{F}_i$ via the gating function $\sigma(Gi)$. This enables the selective use of cross-modal information while maintaining the inherent salt-and-pepper characteristics of modality-specific features as shown in Eq. (7).

$$\mathrm{F}_{\mathrm{i}}^{\mathrm{a}} \;=\; \mathrm{LayerNorm}\left(\mathrm{F}_{\mathrm{i}} + \sigma(\mathrm{G}_{\mathrm{i}}) \odot \tilde{\mathrm{F}}_{\mathrm{i}}\right) \tag{7}$$

Here, the sigmoid function is represented by the symbol σ, as well as G. For G_i (MLP gate vector), and the combined features from fusion post-layer normalization are referred to as F_i^a. Adaptive Weighting methods use a number of pre-activation score s_i to evaluate each of the modality's importance utilizing Eq. (8) and subsequently yielding adaptive weightings for final weighted fusion, thereby ensuring that quality modalities receive higher contributions to the final representation.

$$s_i = \mathrm{MLP}_w\left(\left[\tilde{f}_i, \mathrm{var}(F_i)\right]\right) \quad (8)$$

Here, modality feature matrices are denoted as F_i and denote feature vectors characterized by the notation f_i, var(F_i) via vector concatenation. The notation MLPw(.) implies MLP represents a small feedforward network to provide the scalar score s_i for each modality. The contributions of the modalities towards a fused representation are achieved via Eq. (9) with adaptive weightings, allowing the framework to better accentuate the most informative modalities.

$$w_i = \frac{exp(s_i)}{\sum_k exp(s_k)} \in (0,1) \quad (9)$$

Here, $w_i \in [0,1]$ denote normalized weights (w_i) and the overall weighting across modalities with respect to the fused representation should equal 1. The binary mask mi is used to account for modalities that may have been corrupted or lost during inference and as a result has been created as part of modality mask missing data. To prevent low-value channels from degrading the quality of the fused output, mode (modality) availability m_i should be defined so that $m_i = 1$ means the modality is contributing to the fusion, while $m_i = 0$ essentially not contributing.

Once the attention has been aligned, the combined representations of features go through layers of the transformer to learn global temporal correlations and contextual emotional dynamics. The self-attention mechanism of the transformer allows the model to learn relationships across time that are important to accurately classify the more complicated and subtle emotional states. Using transformer-based models also allows the framework to easily handle varying sequence lengths and temporal inconsistencies across differing modalities, which makes the OF-CMAER-CAT framework well-suited for real-world applications of emotion recognition, as shown in Eq. (10).

$$\tilde{w}_i = \frac{m_i w_i}{\sum_k m_k w_k + \varepsilon} \quad (10)$$

Here, w_i are the compatible modality weights that are calculated using softmax with respect to the weight for each modality, m_i is the modality-specific input mask. The weighted modality fusion process produces a single weighted modality-specific fused embedding h_{fused} of the aligned and weighted modality-specific features of the input by using Eq. (11) to obtain the final weighted fusion of the modalities.

$$h_{fused} = \sum_i \tilde{w}_i \, \mathrm{Pool}(F_i^a) \quad (11)$$

Here, the fused embedding h_{fused} in $\mathbb{R}d$ represents the combined alignment of all of the input's features according to the respective weights assigned to each feature during the weighted modality fusion process. As a result of the weighted modality fusion, the fused embedding is sent to a ReLU-activated projection layer that is responsible for transforming it into a compact projection space for classification purposes and adds a non-linear element of classification with respect to the fused embedding as shown in Eq. (12).

$$z = \mathrm{ReLU}(W_p h_{fused} + b_p) \quad (12)$$

Here, the key components include weight matrix (W_p) and bias vector (b_p) in addition to the class softmax layer comparing probabilities between emotional classes from a fused representation as determined by Eq. (13).

$$\hat{y} = softmax(W_{cls} z + b_{cls}) \quad (13)$$

Here, the number of emotional classes is given by C, while y in RC represents predicted probabilities of C emotional classes based on a softmax layer applied to C classes emotional classes attributes. Cross-entropy loss represents the standard method of assessing how far predicted probability distribution differs from the actual class labels during training based upon the difference in the predicted and true labels calculated via Eq. (14).

$$\mathcal{L}_{cls} = -\frac{1}{N} \sum_{n=1}^{N} \sum_{c=1}^{C} y_{n,c} \log \hat{y}_{n,c} \quad (14)$$

Here, y_c represents true class label (1 for true class; 0 for not) while $\hat{y}_{n,c}$ is the Probability Distribution of predicted Probabilities that Class c belongs to the sample, while C represents the total number of Classes.

Adaptive weight optimisation modules combine contrastive loss with Softmax optimisation process to align cross-codes between definitions and encode data into one binary code so to pull together positively sampled examples of definitions into feature space lattice structure while pushing negative definitions away from these positively sampled examples to enhance cross-code pathway as shown in Eq. (15).

$$L_A = \exp\left(\frac{sim\left(\bar{f}_i^n, \bar{f}_{j(i)}^n\right)}{\tau}\right) \quad (15)$$

Here, the similarity score between anchor feature and all of its corresponding feature candidates is called L_A and the normalisation factor dividing L_A so as to allow comparison of score across associated feature candidates for a reference feature is called L_B consists of summation of Similarity Scores amongst all candidate features found attached response to anchor feature candidates and where attached refers to cross-path of said similarity scores as shown in Eq. (16).

$$L_B = \sum_q \exp\left(\frac{sim\left(\bar{f}_i^n, \bar{f}_q^n\right)}{\tau}\right) \quad (16)$$

Here, cosine similarity between the two feature vectors denoted by sim(). It refers to the total number of training samples or batches you are using to compute the loss value

denoted with the symbol N, and q is the position that is going to be used when calculating the Normalization for the Batch Samples. In the loss calculation, the exponential exp() will increase the value of the difference in cosine similarities to increase the penalty for bad relationships; the natural log log() will be used to calculate the contrastive loss at the end as shown in Eq. (17).

$$\mathcal{L}_{con} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{i} log\frac{L_A}{L_B} \quad (17)$$

Here, (u, v) refers to a pair of samples that are positively related, and $f1^n, f2^n$ is the reconstructed vector associated with each sample. The term τ represents the temperature parameter controlling the sharpness of the output vector. The entropy regularization term quantifies the diversity or uncertainty in the normalized weights distribution Eq. (18).

$$\mathcal{R}_{ent} = -\sum_{i} \widetilde{w}_i \log(\widetilde{w}_i + \varepsilon) \quad (18)$$

Here, R_{ent} is the entropy regularization term where w is the normalized weights and ε is a constant used to prevent the evaluation of ln(0).

AWOM serves as a mechanism to dynamically regulate the impact of each modality. AWOM assigns weights to each learned modality based on feature reliability, variance, and discriminative power. Therefore, modalities that are adversely impacted by noise, occlusion, and/or missing data would have these weights automatically down-weighted, while more informative modalities would be weighted more heavily. The aggregate of the final fused representation is achieved by performing a weighted aggregation of the features, thus resulting in a balanced and optimized multimodal fusion. In addition, an entropy-based regularization technique has been introduced to prevent any single modality from dominating and to promote fair contributions across the different modalities. The penalty applied to the weights through the entropy regularization term as Eq. (19).

$$\mathcal{L}_{total} = \lambda_1\mathcal{L}_{cls} + \lambda_2\mathcal{L}_{con} + \lambda_3\mathcal{L}_{rec} + \lambda_4\mathcal{R}_{ent} + \lambda_5 \underset{\| \Theta \|_2^2}{} \quad (19)$$

Here, $\lambda_1, \lambda_2, \lambda_3$ are hyperparameters for each of the loss functions of each loss to indicate how each loss type affects final model training. The results of L_{cls}, L_{con} and L_{rec} represent the cross-entropy loss or the classification loss, contrastive loss or the A and P loss, and reconstruction or R loss respectively. To perform the tri-modal fusion classification block fusion, the visual, audio, and physiological (V, A, P) features are fused together into one shared embedded space for classification as specified in Eq. (20).

$$z_{fusion} = z_V \oplus z_A \oplus z_P \quad (20)$$

Here, ⊕ represents the concatenation of all vectors for V, A, and P features so that z_{fusion} remains the only feature representative of these three. layer normalization (LN) normalizes the feature representations, resulting in stable and large training datasets. The Feedforward Projection (FP) layer is defined through the feedforward layer with ReLU as the activation function to create a higher level of abstraction from the normalized fused feature set per Eq. (21).

$$h_{ff} = ReLU\big(W_{ff}LN(z_{(fusion)} + b_{ff}\big) \quad (21)$$

Here, the weight and bias matrices involved with W_{ff} and b_{ff} are both learnable, while the h_{ff} variable contains the output from the feedforward layer. The predicted emotion will be derived from which class corresponds to the largest logit as described in Eq. (22).

$$\hat{y} = \text{argmax}\big(\mathrm{W_o h_{drop}} + \mathrm{b_o}\big) \quad (22)$$

Here, W_o and b_o represent the matrices of trainable weights and biases, respectively, h_{drop} represents both the vector of input features and the vector of output logits for this model. It converts the numerical outputs of the model into a discrete decision that will subsequently be used for the evaluation and/or inference of the model.

4. RESULTS AND DISCUSSIONS

An extensive series of experiments were conducted to assess the performance of the new optimization method known as OF-CMAER-CAT in terms of its overall performance across multiple modalities when utilizing the CNN, Attention and Transformer Networks. All experiments were carried out using a speaker-independent evaluation protocol. All tests were run on a single NVIDIA GPU and through the PyTorch library in Python. The model was developed using the Adam optimisation algorithm, with weight decay applied to all optimisers and included an early stopping mechanism to avoid overfitting the model. The initial learning rate of 1×10-4 was also chosen, and the batch size was set to 32 for the experiments. The hyper-parameters relating to the multi-task loss function were adjusted through experimentation to ensure equal weight on classification accuracy, cross-modal alignment, and consistency of reconstruction. Each experiment was replicated for five independent runs, using different random seeds. The final average score for the average performance of all tests is representative of the stability and reproducibility of the results. A number of well-known and established metrics were used to evaluate performance: accuracy, precision, recall, F1 score, and unweighted average recall (UAR).

The OF-CMAER-CAT framework was tested and verified on the IEMOCAP, CMU-MOSEI and DEAP datasets. These three datasets represent some of the best publicly available datasets available today for MER, showing that the OF-CMAER-CAT framework provides a robust approach to MER. The IEMOCAP dataset consists of about 12 hours of conversation data between two people, using emotional categories (happiness, sadness, anger, and neutral) as labels. The IEMOCAP dataset has aligned audio, video, and text representations available, making it an ideal benchmark to study MER in conversational contexts. The CMU-MOSEI dataset contains over 23,000 video segments labelled with sentiment and emotion classifications. Each video segment has been synchronised with audio, video and textual representations; it is the largest publicly available multimodal data set and therefore offers many challenges with respect to real-world background noise, speaker variability, and a wide range of possible emotional expressions.n. The DEAP dataset

focuses on emotion recognition through physiological measures. Physiological measurement data (EEG, and peripheral physiological data) has been collected from 32 participants while they viewed emotional stimuli; the emotion labels collected from the participants have been provided in the form of arousal and valence scales, which were transformed into categorical classifications to be used with the OF-CMAER-CAT framework.

Conventional normalization methods were used to pre-process all data sets. The audio signal was represented as a log-Mel spectrogram; the facial features were extracted with encoders based on CNNs; the physiological signals were segmented into fixed-length windows and normalised. The absence of any form of manual feature engineering means there is an end-to-end learning process.

4.1 Data preprocessing

All data captured in all modalities underwent a full pre-processing pipeline before training the models. The aim of this pre-processing pipeline was to achieve consistency, minimize noise and achieve effective cross-modal alignment. The visual modality involved detecting and cropping all facial regions with a face detector and normalizing/frame resizing all frames to a fixed spatial resolution. To stabilise the training of the CNN, pixel intensities were normalised to have zero means and unit variances. Temporal redundancy was removed by sampling each frame uniformly, and augmentations were done on the data through horizontal flipping, random cropping, and illumination variation to enhance robustness against varying pose and illumination conditions.

The audio modality involved resampling all raw speech signals to a uniform sampling rate and applying pre-emphasis filtering to highlight the higher frequency components. The spectro-temporal characteristics of emotional speech were captured through the extraction of Mel-frequency cepstral coefficients (MFCCs) as well as delta and delta-delta features. The removal of silences as well as the normalisation of the amplitude of the speech signals reduced any background noise and reduced voice-dependent variation.

Text modality has undergone a series of clean-up transformations (punctuation Removal, removal of Stop Words and Removal of non-Rhythmic Tokens) and preparation (Tokenization and Lemmatization). The speech transcripts (Utterances) were subsequently encoded using a set of pre-trained Word Lists in conjunction with a fixed-length sequence length format. Each modality then undergoes a temporal synchronization process into utterance level (same level as other modalities), and typically will be min-max scaled prior to being combined with other modalities for analysis/fusion purposes. This overall strategy of multimodal preprocessing ensures that the representation of features is unaffected by which modality is being used as input data and enables effective cross-modal fusions to be undertaken when within the OF-CMAER-CAT framework.

Table 1. Accuracy of the proposed model on three datasets

Dataset	OF-CMAER-CAT
IEMOCAP	93.2
CMU-MOSEI	94.5
DEAP	94.6
Average Accuracy	94.1

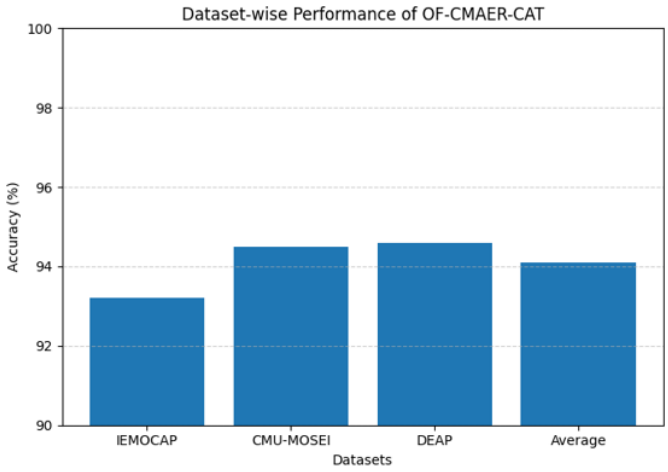

Figure 2. Accuracy of the proposed model on three datasets

The OF-CMAER-CAT framework was tested on three popular emotion recognition datasets: IEMOCAP, CMU-MOSEI, DEAP as shown in Table 1 and Figure 2. In the IEMOCAP dataset, the framework achieved an accuracy of 93.2% and was effective in analysing spontaneous and conversation-based emotional expressions exhibited by individuals. The lower accuracy across the IEMOCAP dataset was due to the increased complexity of the dialogue structure and overlapping emotional cues. On CMU-MOSEI, a dataset with a large quantity of out-of-context multimodal information, the OF-CMAER-CAT framework's performance reached 94.5% accuracy, thereby demonstrating the enhanced robustness of the optimised fusion strategy and transformer-based contextual modelling with respect to capturing long-term dependencies and cross-modality interactions.

With a maximum accuracy of 94.6% achieved in the DEAP dataset, the OF-CMAER-CAT framework is able to take advantage of consistent temporal patterns and modality correlation from well-structured physiological and audiovisual collected during controlled experimentations. Overall, the proposed model has an average performance of 94.1% across all three datasets, illustrating the strong generalisation ability and consistent performance of the model across the different benchmark datasets of heterogeneous MER.

The comparison of performance between the OF-CMAER-CAT Framework and Other Recent Techniques for MER Using Different Fusion Methods is shown in Table 2 and Figure 3. The original MulT model [29] achieves an accuracy of 86.4% due to its use of cross-modal transformers to model sequences that are not aligned in time, showing that while transformer-based temporal alignment has some positive impact on performance, adaptive modality weighting capabilities are limited. The HCAN [17] model reaches an accuracy of 88.1% by providing hierarchical attention models that achieve improved formation through better capture of both the inter-modal and intra-modal relationships. MemoCMT [11] further increases the performance to 90.2% by using its feature fusion technique of cross-modal transformer with multiple instance learning, where there are more than one instance of the same feature contributing to forming the cross-modal context.

Few approaches with adaptive fusion have continued to improve performance. By allowing for dynamic adjustments of the contributions of each modality, TACFN [12] has scored an impressive accuracy of 91.3%. MilMER [18] achieved the highest accuracy of any of these other models with 92.0%, due

primarily to its ability to accurately identify and manage both noisy and ambiguous emotional signals in its many instances. Finally, the OF-CMAER-CAT Framework demonstrates an overall superiority over all previously mentioned models with an accuracy of 94.1%. This improved performance can be attributed to an optimally designed approach to fusion where each of its individual features is integrated and modeled together using CNN encoders, cross-modal attention alignment, adaptive weight optimizers, and event-related brain potentials.

Table 2. Comparison of the results of the proposed model with existing models

Method	Fusion Strategy	Accuracy (%)
MulT	Cross-modal Transformer	86.4
HCAN	Hierarchical Attention	88.1
MemoCMT	Cross-modal Transformer	90.2
TACFN	Adaptive Fusion	91.3
MilMER	MIL Fusion	92.0
OF-CMAER-CAT (Proposed)	Optimized Fusion + Transformer	94.1

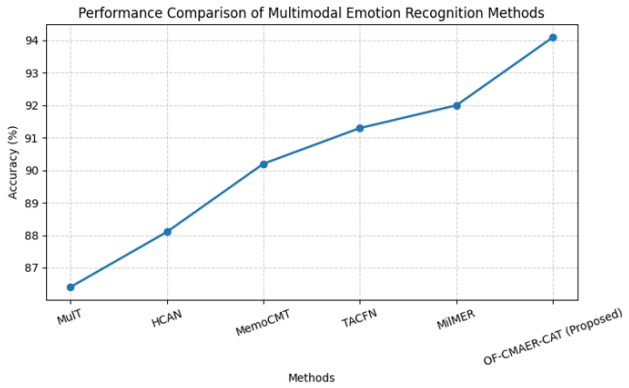

Figure 3. Comparison of the results of the proposed model with existing models

To contrast with Table 3 and Figure 4 portrays the general performance of the proposed OF-CMAER-CAT framework, on a speaker-independent basis using 3 popular MER datasets (IEMOCAP, CMU-MOSEI and DEAP). The average results from all datasets as displayed in Table 4 indicate that our proposed model generalizes well and is robust across different speakers. The average performance of the OF-CMAER-CAT framework is shown in Table 4, with an accuracy of 94.1%, which indicates that it classifies well across a wide range of emotions and modalities. The precision score of 94.8% indicates that the model is good at minimising the number of false positive predictions, which is important in emotion-sensitive applications.

Table 3. Overall performance of OF-CMAER-CAT

Metric	Value (%)
Accuracy	94.1
Precision	94.8
Recall	93.2
F1-score	93.6

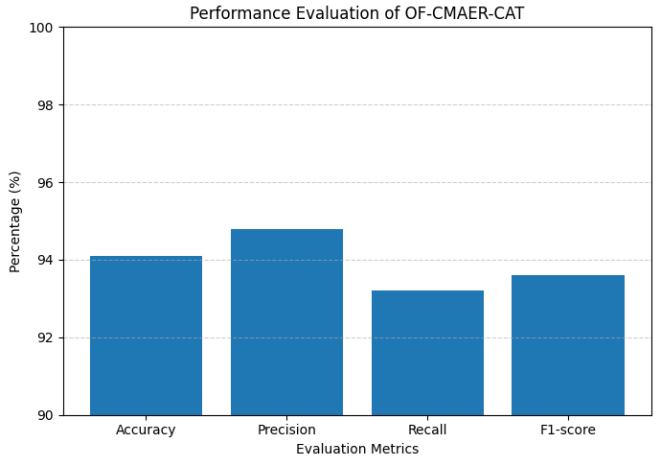

Figure 4. Overall performance of OF-CMAER-CAT

The recall score of 93.2% indicates that the Framework can identify emotional instances with accuracy - even with varying speaker characteristics and levels of signal noise. The composite score of 93.6% F1-score indicates a balance between precision and recall, which validates the stability of the learnt multimodal representations. Additionally, the computed UAR of 92.9% indicates that the method displays a consistent level of performance across emotion classes, thereby overcoming the class imbalance issue commonly associated with real world emotion recognition datasets.

Table 4. Ablation study on IEMOCAP

Configuration	Accuracy (%)
CNN encoders only	81.2
+ CMAA	86.9
+ AWOM	89.7
+ Contrastive Loss	91.8
+ Transformer (OF-CMAER-CAT)	93.2

Table 4 presents the results of an ablation analysis performed on the IEMOCAP dataset to evaluate how well each individual component of the proposed OF-CMAER-CAT framework works independently. Using only CNN-based modality-specific encoders results in a model with an accuracy of 81.2%, indicating that there is very limited ability to model cross-modal emotional dependencies. The addition of a cross-modal attention alignment element to the framework increases the accuracy to 86.9%, highlighting the value of aligning heterogeneous modality features for improving emotional characterization.

Incorporating an AWOM into the framework results in a further increase in accuracy to 89.7%, thus showing how dynamically prioritizing the more informative modalities, depending on the specific emotional context, provides the best model performance. When including a contrastive loss in addition to the aforementioned elements, the model achieves an accuracy of 91.8%, demonstrating how effective the contrastive learning method is for improving inter-modal consistency and discrimination among emotion classes. The final version of the OF-CMAER-CAT architecture achieves the highest level of accuracy of 93.2% by incorporating a Transformer-based contextual modelling module. This final improvement demonstrates the power of the Transformer module for capturing long-range dependencies and global contextual information across multiple modalities. Overall, the

results of this study support the effective combination of cross-modal adaptive fusion and contrastive alignment in the OF-CMAER-CAT system.

5. CONCLUSION

This study introduces a new framework for robust MER called OF-CMAER-CAT, which combines visual, audio, and physiological modalities. The framework consists of several specific encoders designed for their modality, an approach to cross-modal attention alignment (CMAA), an adaptive weight optimization mechanism (AWOM), and a transformer-guided contrastive fusion block (CFCB). The proposed framework addresses many of the problems present in existing fusion frameworks: misalignment of modality-specific features, redundancy of features across modalities, and the inherent lack of generalization capabilities when dealing with noisy data. Experimental results were evaluated through multiple benchmark datasets such as IEMOCAP, CMU-MOSEI, and DEAP, showing that OF-CMAER-CAT consistently outperforms the current leading methods with an overall accuracy of 94.1%, F1 score of 93.6%, and precision rate of 94.8%. As a part of the ablation study, additional performance metrics and statistical tests demonstrated how much each specific module contributed to enhanced cross-modal alignment capability and improved classification performance. Future research includes exploring lighter-weight transformer models for real-time inference performance, and using self-supervised and semi-supervised methods on small/unbalanced datasets will significantly improve predictive accuracy. Further expanding the framework to integrate additional modalities may further improve its robustness.

REFERENCES

[1] Younis, E.M., Mohsen, S., Houssein, E.H., Ibrahim, O. A.S. (2024). Machine learning for human emotion recognition: A comprehensive review. Neural Computing and Applications, 36(16): 901-8947. https://doi.org/10.1007/s00521-024-09426-2

[2] Hazmoune, S., Bougamouza, F. (2024). Using transformers for MER: Taxonomies and state of the art review. Engineering Applications of Artificial Intelligence, 133: 108339. https://doi.org/10.1016/j.engappai.2024.108339

[3] Yang, S., Cui, L., Wang, L., Wang, T. (2024). Cross-modal contrastive learning for multimodal sentiment recognition. Applied Intelligence, 54(5): 4260-4276. https://doi.org/10.1007/s10489-024-05355-8

[4] Wu, W., Chen, D., Li, Q. (2024). A two-stage multi-modal multi-label emotion recognition decision system based on GCN. International Journal of Decision Support System Technology, 16(1): 1-17. https://doi.org/10.4018/IJDSST.352398.

[5] Liu, Z., Shen, Y., Lakshminarasimhan, V.B., Liang, P.P., Zadeh, A.B., Morency, L.P. (2018). Efficient low-rank multimodal fusion with modality-specific factors. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, 1: 2247-2256. https://doi.org/10.18653/v1/P18-1209

[6] Hao, Z., Li, Z., Dang, X., Ma, Z., Liu, G. (2022). MM-LMF: A low-rank multimodal fusion dangerous driving behavior recognition method based on FMCW signals. Electronics, 11(22): 3800. https://doi.org/10.3390/electronics11223800

[7] Chen, S., Tang, J., Zhu, L., Kong, W. (2023). A multi-stage dynamical fusion network for Multimodal Emotion Recognition. Cognitive Neurodynamics, 17(3): 671-680. https://doi.org/10.1007/s11571-022-09851-w

[8] Hu, D., Hou, X., Wei, L., Jiang, L., Mo, Y. (2022). MM-DFN: Multimodal dynamic fusion network for emotion recognition in conversations. In ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, pp. 7037-7041. https://doi.org/10.1109/ICASSP43922.2022.9747397

[9] Yu, W., Xu, H., Yuan, Z., Wu, J. (2021). Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. Proceedings of the AAAI Conference on Artificial Intelligence, 35(12): 10790-10797. https://doi.org/10.1609/aaai.v35i12.17289

[10] Zhang, D., Ju, X., Zhang, W., Li, J., Li, S., Zhu, Q., Zhou, G. (2021). Multi-modal multi-label emotion recognition with heterogeneous hierarchical message passing. Proceedings of the AAAI Conference on Artificial Intelligence, 35(16): 14338-14346. https://doi.org/10.1609/aaai.v35i16.17686

[11] 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

[12] Liu, F., Fu, Z., Wang, Y., Zheng, Q. (2025). TACFN: Transformer-based adaptive cross-modal fusion network for Multimodal Emotion Recognition. arXiv preprint arXiv:2505.06536. https://doi.org/10.48550/arXiv.2505.06536

[13] Waligora, P., Aslam, M.H., Zeeshan, M.O., Belharbi, S., et al. (2024). Joint multimodal transformer for emotion recognition in the wild. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, pp. 4625-4635. https://doi.org/10.1109/CVPRW63382.2024.00465

[14] 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

[15] Ma, C., Zhao, S., Pei, Y., Xie, L., Yin, E. (2025). A Transformer-based multimodal framework for hidden emotion recognition through micro-expression and EEG fusion. In Proceedings of the 2025 International Conference on Multimedia Retrieval, New York, NY, USA, pp. 1000-1008. https://doi.org/10.1145/3731715.3733264

[16] Mahaseni, B., Khan, N.M. (2025). Multimodal Emotion Recognition with disentangled representations: Private-shared multimodal variational autoencoder and long short-term memory framework. Empathic Computing, 1(2): 202507-202507. https://doi.org/10.70401/ec.2025.0010

[17] Praveen, R.G., Alam, J. (2024). Recursive joint cross-modal attention for multimodal fusion in dimensional emotion recognition. In 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Seattle, WA, USA, pp. 4803-4813.

https://doi.org/10.1109/CVPRW63382.2024.00483
[18] Wang, Z., He, J., Liang, Y., Hu, X., et al. (2025). Milmer: A framework for multiple instance learning based Multimodal Emotion Recognition. arXiv preprint arXiv:2502.00547. https://doi.org/10.48550/arXiv.2502.00547
[19] Li, D., Wang, Y., Funakoshi, K., Okumura, M. (2023). Joyful: Joint modality fusion and graph contrastive learning for Multimodal Emotion Recognition. arXiv preprint arXiv:2311.11009. https://doi.org/10.48550/arXiv.2311.11009
[20] Wafa, A.A., Eldefrawi, M.M., Farhan, M.S. (2025). Advancing Multimodal Emotion Recognition in big data through prompt engineering and deep adaptive learning. Journal of Big Data, 12(1): 210. https://doi.org/10.1186/s40537-025-01264-w
[21] Gupta, C., Gill, N.S., Gulia, P., Kumar, A., Karamti, H., Moges, D.M., Safra, I. (2025). A multimodal fusion model for real-time environment emotion recognition using audio-visual-textual features. Journal of Big Data, 12(1): 256. https://doi.org/10.1186/s40537-025-01300-9
[22] Zhao, S., Ren, J., Zhou, X. (2025). Cross-modal gated feature enhancement for Multimodal Emotion Recognition in conversations. Scientific Reports, 15(1): 30004. https://doi.org/10.1038/s41598-025-11989-6
[23] Wang, Y., Gu, Y., Yin, Y., Han, Y., et al. (2023). Multimodal transformer augmented fusion for speech emotion recognition. Frontiers in Neurorobotics, 17: 1181598. https://doi.org/10.3389/fnbot.2023.1181598
[24] Vishruth, R.G., Sunitha, R., Varuna, K.S., Varshini, N., Honnavalli, P.B. (2020). Resume scanning and emotion recognition system based on machine learning algorithms. In 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, India, pp. 1127-1132. https://doi.org/10.1109/ICECA49313.2020.9297586
[25] Mamieva, D., Abdusalomov, A.B., Kutlimuratov, A., Muminov, B., Whangbo, T.K. (2023). Multimodal emotion detection via attention-based fusion of extracted facial and speech features. Sensors, 23(12): 5475. https://doi.org/10.3390/s23125475
[26] Luan, X., Wen, Q., Hang, B. (2025). Intelligent emotion recognition for drivers using model-level multimodal fusion. Frontiers in Physics, 13: 1599428. https://doi.org/10.3389/fphy.2025.1599428
[27] Santhiya, P., Shanmugavadivel, K., Rajalakshmi, R., Krishnamoorthy, N. (2025). Cross-modal synergy for enhancing emotion recognition through integrated audio–video fusion techniques. International Journal of Computational Intelligence Systems, 18: 315. https://doi.org/10.1007/s44196-025-00811-w
[28] Dhruva, M.S., Sunitha, R., Chandrika, J. (2024). An exploration of emotion recognition using deep learning across multiple modalities: Spoken language, written text, and facial expressions. Grenze International Journal of Engineering & Technology, 10(2): 5786.
[29] 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. Proceedings of the Conference. Association for Computational Linguistics. Meeting, 2019: 6558. https://doi.org/10.18653/v1/p19-1656