Empathic Computing

Science Exploration

Open Access Research Article

Multimodal emotion recognition with disentangled representations: private-shared multimodal variational autoencoder and long short-term memory framework

Behzad Mahaseni*, Naimul Mefraz Khan

Department of Electrical and Computer Engineering, Toronto Metropolitan University, Toronto, ON M5B 2R2, Canada.

***Correspondence to:** Behzad Mahaseni, Department of Electrical and Computer Engineering, Toronto Metropolitan University, Toronto, ON M5B 2R2, Canada. E-mail: behzad.mahaseni@torontomu.ca

Received: March 27, 2025 **Accepted:** June 27, 2025 **Published:** June 29, 2025

Cite this article: Mahaseni B, Khan NM. Multimodal emotion recognition with disentangled representations: private-shared multimodal variational autoencoder and long short-term memory framework. Empath Comput. 2025;1:202507. https://doi.org/10.70401/ec.2025.0010

Abstract

Aims: This study proposes a multimodal emotion recognition framework that combines a private-shared disentangled multimodal variational autoencoder (DMMVAE) with a long short-term memory (LSTM) network, herein referred to as DMMVAE-LSTM. The primary objective is to improve the robustness and generalizability of emotion recognition by effectively leveraging the complementary features of electroencephalogram (EEG) signals and facial expression data.

Methods: We first trained a variational autoencoder using a ResNet-101 architecture on a large-scale facial dataset to develop a robust and generalizable facial feature extractor. This pre-trained model was then integrated into the DMMVAE framework, together with a convolutional neural network-based encoder and decoder for EEG data. The DMMVAE model was trained to disentangle shared and modality-specific latent representations across both EEG and facial data. Following this, the outputs of the encoders were concatenated and fed into a LSTM classifier for emotion recognition.

Results: Two sets of experiments were conducted. First, we trained and evaluated our model on the full dataset, comparing its performance with state-of-the-art methods and a baseline LSTM model employing a late fusion strategy to combine EEG and facial features. Second, to assess robustness, we tested the DMMVAE-LSTM framework under data-limited and modality dropout conditions by training with partial data and simulating missing modalities. The results demonstrate that the DMMVAE-LSTM framework consistently outperforms the baseline, especially in scenarios with limited data, indicating its capacity to learn structured and resilient latent representations.

Conclusion: Our findings underscore the benefits of multimodal generative modeling for emotion recognition, particularly in enhancing classification performance when training data are scarce or partially missing. By effectively learning both shared and private representations, DMMVAE-LSTM framework facilitates more reliable emotion classification and presents a promising solution for real-world applications where acquiring large labeled datasets is challenging.

Keywords: Multimodal emotion recognition, multimodal variational autoencoder, long short-term memory, electroencephalogram, facial expressions

1. Introduction

Emotion is a fundamental mental state that plays a vital role in various aspects of daily life, including interpersonal interactions, decision-making, learning, and working[1,2]. Emotion recognition has attracted considerable attention and remains a challenging research area due to its broad applications across domains such as healthcare[3], education[4,5], autonomous driving[6], and gaming[7]. Recent advances in deep learning and machine learning have significantly expanded the scope and effectiveness of emotion recognition technologies. There are two primary approaches for identifying human emotions: analyzing behavioral data such as facial expressions, speech, and gestures and examining physiological signals, including electroencephalogram (EEG), electrocardiogram (ECG), galvanic skin response (GSR), and heart rate. Early research predominantly focused on unimodal techniques that treated behavioral and physiological modalities separately. While these individual modalities provide valuable insights, they also present inherent limitations. For example, behavioral data are susceptible to reliability issues due to "social masking", where individuals may consciously or unconsciously conceal their true emotions. Moreover, using behavioral data like facial expressions

and speech raises substantial privacy concerns. These factors can result in misleading behavioral cues and complicate accurate emotion inference based solely on such data. Conversely, physiological signals tend to be more difficult to consciously manipulate, offering a more objective and reliable basis for emotion recognition. However, acquiring these signals often requires specialized, medical grade sensing equipment, which can be costly. Additionally, unimodal emotion recognition systems face challenges in capturing the complexity of emotional states because a single modality may fail to account for individual differences, be vulnerable to noise, and exhibit inherent variability within the data source. Human emotions are complex, involving subjective feelings and intertwined physiological and behavioral responses elicited by external stimuli, which are difficult to fully characterize using only one modality[8,9].

These considerations underscore the necessity of treating emotion recognition as a multimodal problem. Multimodal emotion recognition systems tend to be more robust and reliable across diverse real world scenarios by allowing complementary modalities to offset each other's limitations. They also help reduce ambiguity by integrating multiple signals. Despite their advantages, multimodal systems face challenges related to data integration, feature extraction, and designing effective high dimensional fusion models. Furthermore, such systems are often computationally demanding. The multimodal variational autoencoder (MMVAE)[10], a recent extension of the variational autoencoder (VAE)[11], is designed to learn a shared latent representation from multiple modalities (e.g., images, text, audio) that encode common underlying information. However, a significant challenge in applying these models to multimodal time-series data lies in their limited ability to capture long temporal dependencies. To overcome this, we propose extending the MMVAE architecture by incorporating a recurrent neural network with memory. Specifically, we build upon the private-shared disentangled MMVAE[12] by adding a long short-term memory (LSTM) network to enhance temporal context modeling, as illustrated in Figure 1. The private-shared disentangled multimodal variational autoencoder (DMMVAE) partitions the latent space into shared representations that encode common features across modalities and private representations that capture modality-specific characteristics. We argue that integrating LSTM is particularly beneficial for complex emotion recognition tasks for two main reasons. First, the private-shared DMMVAE allows the network to learn cross-modal representations between images and EEG signals via shared latent variables z_s, while maintaining the separation of modality-specific latent variables $\{z_{pi}\}$. Empirically, we observed that since these representations are trained on short data snippets, the private latent variables $\{z_{pi}\}$ tend to be less accurate and more susceptible to noise. The LSTM's hidden layer mitigates this by effectively averaging the private latent variables, thereby producing a more stable latent representation. Second, when one modality is missing, memory-less classifiers typically yield noisy and uncertain predictions. The LSTM's memory cell enables the model to learn coherent temporal representations over time, improving robustness in such scenarios. Throughout this paper, we refer to the proposed model as DMMVAE-LSTM, denoting the private-shared Disentangled MMVAE augmented with an LSTM head.

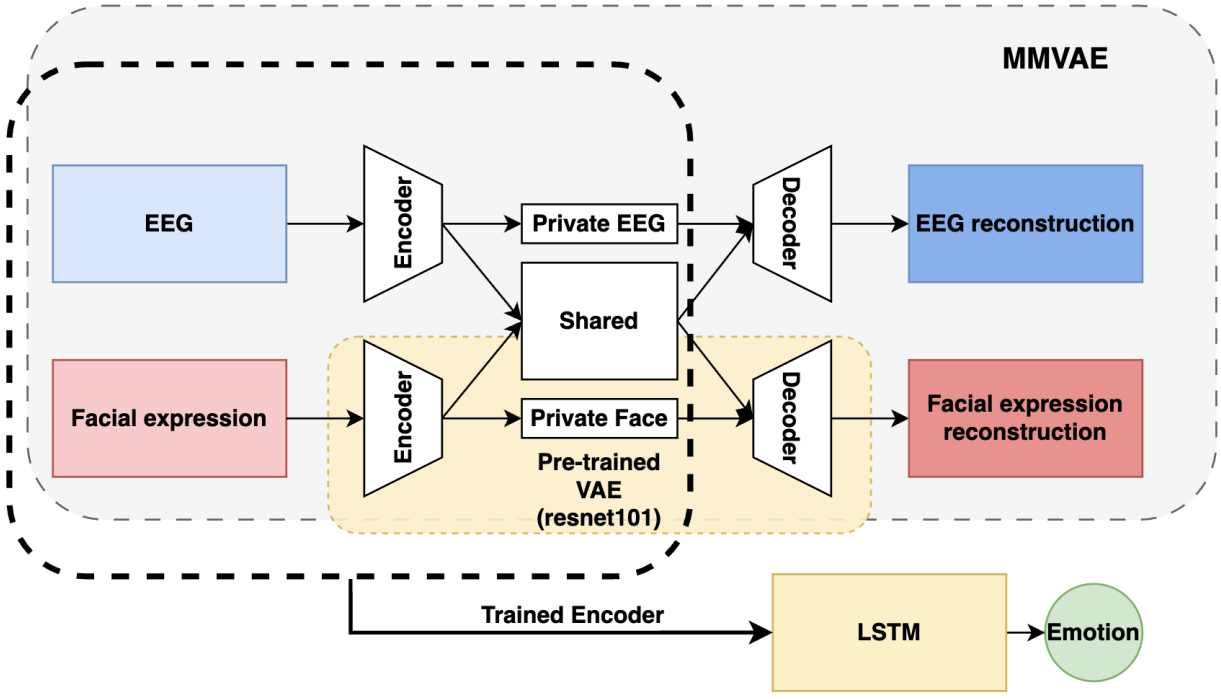

Figure 1. Overview of the DMMVAE-LSTM framework for multimodal emotion recognition. The MMVAE component extracts shared and private latent representations from EEG and facial expressions, with the facial encoder/decoder initialized using a pre-trained ResNet-101-based VAE. The trained encoder outputs are passed to an LSTM to model temporal dependencies and classify emotions. DMMVAE: disentangled multimodal variational autoencoder; LSTM: long short-term memory; MMVAE: multimodal variational autoencoder; EEG: electroencephalogram; VAE: variational autoencoder.

To evaluate our model, we conducted two sets of experiments. In the first set, the model was trained on the complete dataset to ensure it learned a comprehensive representation of emotions. We compared the performance of our DMMVAE-LSTM model with baseline methods, including a standard LSTM model that employs separate LSTM networks for each modality, followed by late fusion at the final layer. In the second set, we tested the model's robustness under data-limited conditions by training it on only 50% and 75% of the available data. This experiment aimed to assess how well the DMMVAE framework generalizes when less training data are accessible. As a generative model, DMMVAE learns a structured latent space that captures underlying patterns more effectively than purely discriminative models like LSTM, resulting in improved generalization when training data are scarce. Our results demonstrate that the proposed DMMVAE-LSTM consistently outperforms traditional LSTM-based approaches, especially in data-limited scenarios. The capacity of DMMVAE to leverage both shared and private latent representations enables it to maintain classification performance despite reduced training data, making it a promising approach for real-world applications where large-scale labeled datasets are difficult to obtain. Additionally, we performed two ablation studies to further analyze the proposed model. First, we investigated the effect of incorporating a pre-trained variational autoencoder within the MMVAE framework to evaluate how prior facial feature learning impacts multimodal emotion recognition. Second, we assessed the model's robustness under modality dropout conditions. These studies provide valuable insights into the contributions of each component to the overall performance of our approach.

To summarize, our main contributions are as follows:

1) We propose a novel private-shared DMMVAE framework for multimodal emotion recognition that effectively learns both shared and modality-specific representations.

2) We integrate sequential modeling with a generative framework by combining LSTM with DMMVAE to capture temporal dependencies in emotional states effectively.

3) We trained a ResNet-101-based variational autoencoder on a diverse set of large-scale facial datasets to enhance the robustness and generalizability of facial feature extraction.

4) We evaluated the proposed approach under limited data conditions and modality dropout scenarios, demonstrating its robustness and generalizability.

2. Related Work

This section provides an overview of prior research on emotion recognition, along with a brief summary of related deep learning architectures employed in this domain.

2.1 Emotion recognition

Psychological studies have shown that human emotions comprise complex interactions of subjective feelings, physiological reactions, and behavioral responses triggered by internal or external stimuli[13]. Early research in emotion recognition primarily focused on unimodal approaches, treating each modality independently. A wide range of behavioral and physiological signals has been explored, including facial expressions[14-18], speech[19-21], gestures[22], ECG[23,24], and EEG[25-27], each exhibiting distinct statistical characteristics. Among these, EEG signals and facial expressions are the most widely utilized modalities for emotion analysis.

EEG is a non-invasive technique that records cortical electrical activity via electrodes placed on the scalp. Its sensitivity to subtle emotional changes makes EEG a valuable modality for emotion recognition. Feature extraction methods for EEG-based emotion recognition often involve signal decomposition techniques such as Fourier transform or wavelet transform, enabling analysis of frequency bands linked to specific emotional states. Additionally, statistical features (e.g., mean, variance) and entropy-based measures are frequently used to characterize EEG signals. In contrast, the face is the most intuitive medium for individuals to express emotions. Facial expression analysis leverages techniques such as facial landmark detection, geometric feature extraction, and texture-based methods, focusing on face region localization and extraction of geometric and appearance features. More recently, deep learning methods including convolutional neural networks (CNNs)[27,28] and variational autoencoders (VAEs)[29-32] have achieved remarkable success in automatically extracting robust, hierarchical features from both facial and EEG data. These approaches have advanced the field by enabling real-time, high-accuracy facial emotion recognition.

2.2 Multimodal emotion recognition

As noted, behavioral data and physiological signals are the primary sources for emotion identification. However, behavioral data often suffer from reliability issues because individuals may involuntarily or consciously mask their true emotions, a phenomenon known as social masking[33]. Furthermore, acquiring physiological signals requires specialized equipment, which can be invasive, costly, and demands professional expertise during data collection[13]. Additionally, the inherent complexity of emotions and inter-individual differences in physiological responses can lead to reduced classification performance when relying on a single physiological modality[34].

Due to the limitations of unimodal emotion recognition, researchers have explored emotion classification using multiple modalities.

Multimodal emotion recognition involves analyzing emotional states based on a combination of behavioral cues and physiological responses. It has proven to be more effective than unimodal approaches, as different modalities can complement and compensate for each other's weaknesses. Furthermore, emotions are complex and can manifest differently across individuals. By incorporating multiple sources of information, multimodal systems can reduce ambiguity and improve recognition accuracy. Several studies have proposed different strategies for multimodal emotion recognition. Tseng *et al.*[35] and Liu *et al.*[36] developed frameworks that analyze and recognize emotions by combining facial expressions and body gesture data. Dessai and Virani[37] introduced an emotion classification method that integrates features from ECG and GSR signals using early fusion techniques. They evaluated several machine learning classifiers and found that the k-nearest neighbor classifier achieved the highest accuracy, particularly with GSR data. Lopez *et al.*[38] proposed a hypercomplex multimodal neural network featuring a novel fusion module that applies parameterized hypercomplex multiplications to enhance emotion recognition. Their model, which utilizes EEG, ECG, GSR, and eye-tracking data from the MAHNOB-HCI dataset, outperformed state-of-the-art models by effectively capturing cross-modal correlations through hypercomplex algebra. Pan *et al.*[39] developed a deep learning-based framework called Deep-Emotion, which combines an improved GhostNet for facial expressions, a lightweight fully convolutional network for speech signals, and a tree-structured LSTM model for EEG signals. Yin *et al.*[40] introduced the token-disentangling mutual transformer, a model designed to separate and integrate inter-modality emotion consistency and intra-modality heterogeneity. Their architecture includes a token separation encoder and a token mutual transformer with bi-directional query learning, achieving state-of-the-art results on the CMU-MOSI, CMU-MOSEI, and CH-SIMS datasets. Ali and Hughes[41] proposed a unified biosensor-vision multimodal transformer model, which combines a two-dimensional representation of ECG or photoplethysmogram (PPG) signals with facial data. They also evaluated the performance of unimodal and multimodal approaches using three image-based representations of the ECG/PPG signals. Fu *et al.*[42] introduced a multimodal feature fusion neural network that combines EEG and eye movement data for emotion classification. The model uses a dual-branch feature extraction module to capture spatial-temporal features from both modalities, followed by a multi-scale feature fusion module with cross-channel soft attention to enhance feature selection. Cheng *et al.*[43] presented a transformer autoencoder (TAE) designed to handle missing data in multimodal emotion recognition. The TAE model includes a modality-specific hybrid transformer encoder for extracting local and global context, an inter-modality transformer encoder for learning cross-modal correlations, and a convolutional decoder for feature reconstruction and classification. The DEAP and SEED-IV datasets were used to evaluate the model's robustness under various levels of missing data. Wu *et al.*[44] and Wang *et al.*[45] developed self-supervised learning frameworks for multimodal emotion recognition based on transformer architectures. Wu *et al.*[44] proposed a framework that employs temporal convolution-based modality-specific encoders and a transformer-based shared encoder to learn intra- and inter-modal correlations from physiological signals. Wang *et al.*[45] introduced a fusion method based on tensor decomposition and self-supervised multi-task learning, leveraging transformers to extract features from multiple modalities, including text, audio, and visual data.

A number of studies have focused on the integration of facial video and EEG signals for emotion classification[46-51]. Among these, Wang *et al.*[49], Wu and Li[50], and Zhu *et al.*[51] are most closely related to our work in terms of datasets and modality combinations. Wang *et al.*[49] proposed a deep learning-based multimodal model that integrates EEG signals and facial expressions using the DEAP and MAHNOB-HCI datasets. They applied CNNs with both local and global convolution kernels to extract spatial features from EEG signals, and used a pre-trained CNN with an attention mechanism to enhance facial expression features. Similarly, Zhu *et al.*[51] proposed a decision-level fusion approach that combines EEG signals, peripheral physiological signals, and facial expressions for emotion recognition within the valence-arousal space. The study utilizes a 1D CNN to extract features from 32 EEG channels, a 3D CNN to process facial expression video data, and a neural network-based classifier for peripheral physiological signals. Evaluation using the DEAP dataset demonstrates that the multimodal fusion strategy outperforms both unimodal and bimodal approaches in emotion recognition accuracy. Wu and Li[50] developed a multi-level CNN model for facial expression-based emotion recognition and a stacked bidirectional LSTM model for EEG-based emotion recognition. Their method was evaluated on the DEAP dataset for multimodal emotion classification. The decision-level fusion of the two modalities was achieved using dempster-shafer evidence theory, which effectively integrates the classification outputs from both models to enhance overall performance.

2.3 Multimodal approaches

Multimodal systems are more expressive, efficient, unambiguous, and reliable. Moreover, models must understand the complex intra-modal and cross-modal interactions to accurately predict emotions, and a trained model must be robust to unexpected missing or noisy modalities during testing. A fundamental challenge in multimodal approaches is the fusion of different modalities[52].

Fusion approaches in multimodal neural networks refer to techniques that combine information from multiple modalities such as text, images, audio, or other data types to improve the performance of machine learning models. Early studies in the research community explored various multimodal data fusion techniques, including early, late, and intermediate fusion[53,54]. Recently, methods have tended to be model based, involving the learning of joint representations. The MMVAE is the most representative deep learning model for multimodal data fusion, based on the standard variational autoencoder[55]. MMVAEs aim to address cross-modality and shared-modality representation learning problems in data fusion. Over the past few years, numerous studies have investigated multimodal integration based on MMVAE[56-65].

3. Proposed Method

This section introduces our proposed framework, which integrates the DMMVAE-LSTM networks for multimodal emotion recognition. The DMMVA extends the MMVAE by learning both shared and modality-specific latent representations from EEG signals and facial expressions, as illustrated in Figure 1.

3.1 MMVAE

VAE: The VAE is a variant of the autoencoder architecture, designed to encode input data into a probabilistic latent space and reconstruct the input from these latent variables[11,66]. A key property of the VAE is that it constrains the latent representation to follow a Gaussian distribution with zero mean and unit variance. By enforcing this distributional prior, random samples can be drawn from the latent manifold, enabling the generation of novel examples. Given an observed data point x, the VAE introduces a latent variable z that captures meaningful variations in the data. The joint distribution over x and z is defined as:

$$p_\theta(x, z) = p(z)p_\theta(x \mid z) \tag{1}$$

where $p(z)$ is the prior distribution of the latent variable, typically modeled as a standard Gaussian distribution $N(0,I)$, and $p_\theta(x|z)$ represents the likelihood of the observed data given the latent variable, parameterized by a decoder network. This joint distribution is generally intractable to compute analytically due to the complexity of real-world data. To overcome this, VAEs introduce an inference model $q_\varphi(z|x)$, which is optimized to approximate the true posterior distribution $p(z|x)$ by minimizing the Kullback-Leibler (KL) divergence between them, defined as follows:

$$L(x; \theta, \phi) = -E_{q_\phi(z|x)}\left[\log p_\theta(x \mid z)\right] + D_{\mathrm{KL}}\left(q_\phi(z \mid x) \,\|\, p(z)\right) \tag{2}$$

where:

• The reconstruction loss $E_{q\varphi}(z|x)[\log p_\theta(x|z)]$ ensures that the decoder accurately reconstructs the input data.

• The KL divergence term $\mathrm{KL}(q_\varphi(z|x)||p(z))$ regularizes the latent space by encouraging the approximate posterior to remain close to the prior distribution.

MMVAE: The MMVAE extends the VAE framework to handle multiple data modalities, such as EEG and facial data, by learning a joint latent representation. The MMVAE models the joint distribution of multiple modalities $x = \{x_1, x_2, \ldots, x_N\}$ and the latent variable z as follows:

$$p_\theta(x, z) = p(z)\prod_{i=1}^{N} p_\theta(x_i \mid z) \tag{3}$$

where $p_\theta(x_i|z)$ is the likelihood of modality i given the shared latent variable z.

To approximate the posterior distribution $p_\theta(z|x)$, the MMVAE employs a product-of-experts approach:

$$q_\phi(z \mid x) \propto p(z)\prod_{i=1}^{N} q_\phi(z \mid x_i) \tag{4}$$

where $q_\varphi(z|x_i)$ denotes the approximate posterior for each modality. Assuming Gaussian distributions, the posterior mean and variance of the shared latent variable z_s are computed as:

$$\Sigma_s^{-1} = \sum_{i=1}^{N} \Sigma_i^{-1},\ \mu_s = \Sigma_s \sum_{i=1}^{N} \Sigma_i^{-1} \mu_i \tag{5}$$

where $\sum_i$ and μ_i represent the covariance and mean of the posterior $q_\varphi(z|x_i)$ for each modality. Accordingly, the evidence lower bound can be formulated as follows:

$$L(X; \theta, \phi) \triangleq E_{q_\phi(z|x)}\left[\sum_{x_i \in X} \log p_\theta(x_i \mid z)\right] - KL\left(q_\phi(z \mid X) \,\|\, p(z)\right) \tag{6}$$

Private-shared DMMVAE, proposed by Lee and Pavlovic[12], is designed to enhance multimodal representation learning by explicitly disentangling shared and modality-specific latent features. Traditional multimodal VAEs primarily focus on extracting a shared latent

representation that captures commonalities between modalities. However, this approach often overlooks modality-specific variations, which are essential for effective multimodal learning. To address this, the private-shared DMMVAE introduces separate latent spaces for shared and private representations, ensuring that both common and unique aspects of each modality are effectively captured.

The objective of this model is to achieve improved latent space disentanglement by structuring the latent representation into the following components:

• **A shared latent space** (z_s), which captures features common across all modalities.

• **Private latent spaces** (z_{pEEG} and z_{pface}), which encode modality-specific features for EEG and facial expression data, respectively.

The latent representation for each modality *i* is factorized as:

$$z_i = \left[z_s, z_{p_i}\right] \tag{7}$$

where z_s is shared across all modalities, and z_{pi} is specific to modality *i*. This disentanglement representation enables the model to learn generalized emotional features while preserving distinct patterns unique to each data source.

To facilitate effective learning, the training objective of the DMMVAE-LSTM comprises the following loss components:

• **Reconstruction Loss (Lrecon):** This loss ensures that the model accurately reconstructs the input data while preserving modality-specific information:

$$L\text{recon} = \sum_{i=1}^{N} Eq\phi\left(z \mid x_i\right)\left[\log p_\theta\left(x_i \mid z_s, z_{p_i}\right)\right] \tag{8}$$

It helps maintain the integrity of each modality's input features and supports effective multimodal representation learning.

• **Cross-Reconstruction Loss (Lcross):** This loss encourages the shared latent space to be meaningful across different modalities by enabling one modality to reconstruct the data of another:

$$L\text{cross} = \sum_{i \neq j} Eq\phi\left(z \mid x_i\right)\left[\log p_\theta\left(x_j \mid z_s\right)\right] \tag{9}$$

This mechanism promotes the extraction of common emotional features, thereby improving the robustness and generalizability of the shared latent representation across modalities.

• **KL Divergence Loss (LKL):** The KL divergence loss constrains the learned latent distributions to remain close to a predefined prior, which helps prevent overfitting and promotes a smooth latent space.

$$L\text{KL} = \text{KL}\left(q_\phi\left(z_s \mid x\right) \| \, p\left(z_s\right)\right) + \sum_{i=1}^{N} \text{KL}\left(q_\phi\left(z_{p_i} \mid x_i\right) \| \, p\left(z_{p_i}\right)\right) \tag{10}$$

This regularization supports the formation of a structured and meaningful latent space by balancing shared and modality-specific information.

The overall loss function combines the above components with weighting factors α and β to regulate the contributions of the cross-reconstruction and KL divergence terms:

$$L\text{total} = L\text{recon} + \alpha L\text{cross} + \beta L_{\text{KL}} \tag{11}$$

This formulation ensures the model to learn both shared and private representations effectively while maintaining a balance between reconstruction fidelity and latent space regularization.

LSTM networks[67] are a type of recurrent neural network specifically designed to capture long-range dependencies in sequential data. To model the temporal dynamics of emotional expression in both EEG and facial image sequences, we integrate an LSTM network into our proposed framework, as illustrated in Figure 2. The LSTM receives fused latent representations from both modalities, allowing for sequential processing of emotion-related features over time. Let x_t denote the input at time step *t*. The LSTM maintains a memory cell c_t and uses three gates (input, forget, and output) to regulate the flow of information and compute the hidden state h_t. Its internal operations are defined as follows:

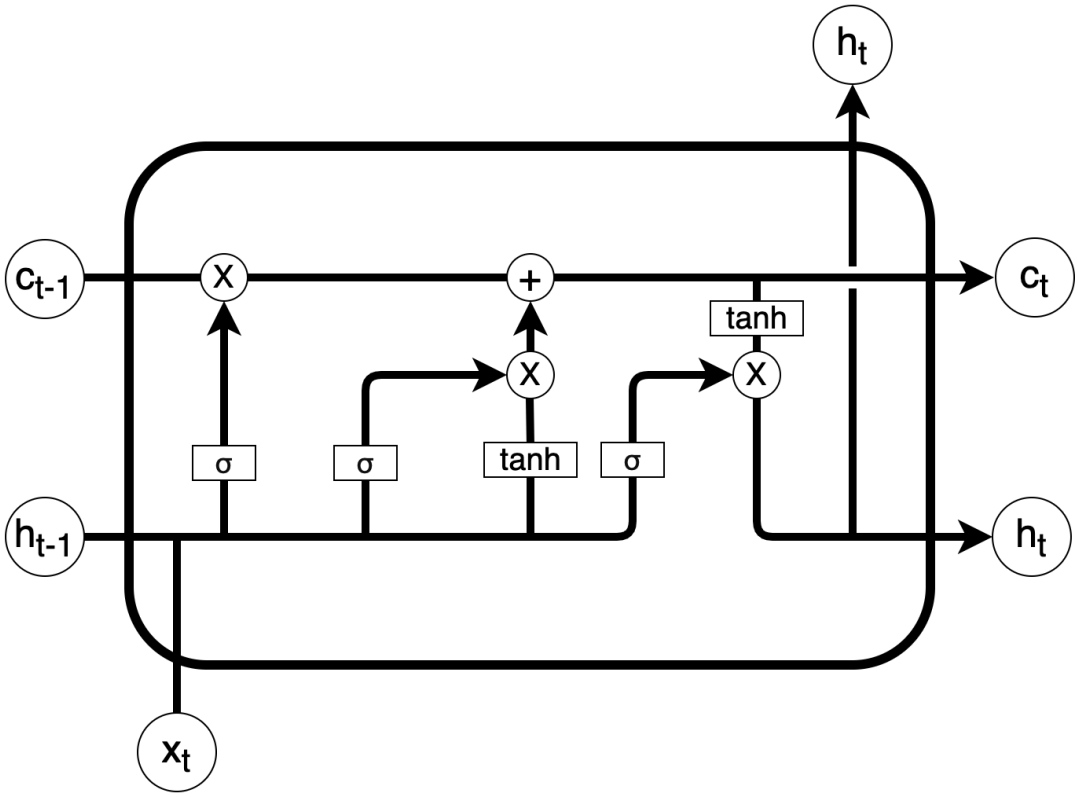

Figure 2. Illustration of an LSTM unit. LSTM: long short-term memory.

$$\text{forget gate:}\quad f_t = \sigma\left(W_f\left[h_{t-1}, x_t\right] + b_f\right) \tag{12}$$

$$\text{input gate:}\quad i_t = \sigma\left(W_i\left[h_{t-1}, x_t\right] + b_i\right) \tag{13}$$

$$\text{cell input:}\quad \tilde{c}_t = \tanh\left(W_c\left[h_{t-1}, x_t\right] + b_c\right) \tag{14}$$

$$\text{cell state:}\quad c_t = f_t \cdot c_{t-1} + i_t \cdot \tilde{c}_t \tag{15}$$

$$\text{output gate:}\quad o_t = \sigma\left(W_o\left[h_{t-1}, x_t\right] + b_o\right) \tag{16}$$

$$\text{hidden state:}\quad h_t = o_t \cdot \tanh\left(c_t\right) \tag{17}$$

where f_t, i_t, and o_t are the forget, input, and output gates respectively. These mechanisms allow the LSTM to selectively retain relevant temporal features and discard irrelevant information.

4. Training Procedure

The training pipeline consists of three stages: (1) pre-training a VAE with a ResNet-101 backbone for facial image reconstruction, (2) training a DMMVAE for reconstructing both EEG signals and facial images, and (3) using the trained DMMVAE encoder to extract latent representations for LSTM-based emotion classification. To ensure consistency, the same set of hyperparameters was applied across all three stages. These values were selected based on commonly used configurations in related studies and refined through preliminary initial empirical testing. A batch size of 32 was selected, as it offers a favorable balance between convergence speed and memory efficiency on modern GPUs. The learning rate was set to 0.001, a standard initial value that demonstrated stable convergence when combined with the Adam optimizer. The DMMVAE module, which includes the ResNet-101-based facial encoder, contains 78,785,859 parameters. The LSTM classifier comprises an additional 10,244,609 parameters, resulting in a total of 89,526,318 trainable

parameters. Although the total parameter count is relatively high, it reflects the model's capacity to learn rich, disentangled, and temporally informed multimodal representations. This capacity is essential for achieving robust emotion recognition in real-world scenarios.

4.1 VAE pre-training with ResNet-101

Both the DEAP and MAHNOB datasets include a limited number of subjects, which makes it challenging to directly train the DMMVAE on these datasets while ensuring generalization and robust facial feature extraction. To address this limitation, we pretrained a VAE with a ResNet-101 backbone (Figure 3a) using a more diverse dataset collected from multiple sources[68-72]. The ResNet-101 encoder was chosen for its strong feature extraction capabilities, enabling the VAE to learn high-level facial representations that capture essential structural and expression-related features for more generalizable representation learning.

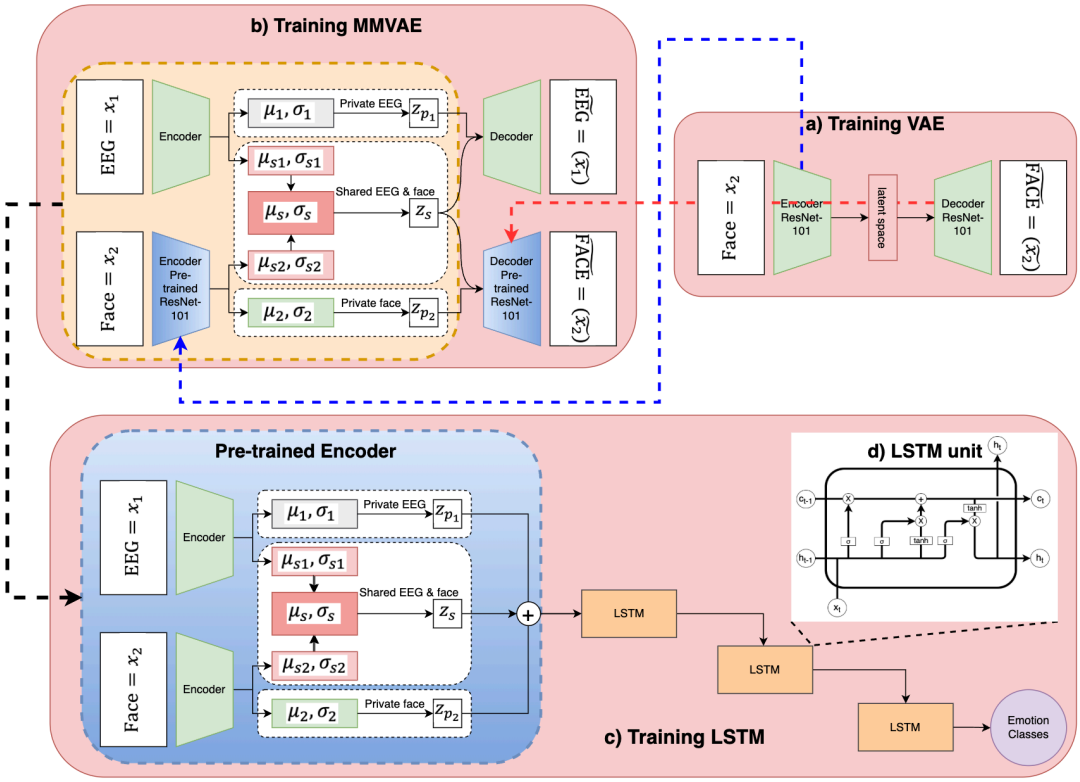

Figure 3. Overview of the proposed DMMVAE-LSTM training pipeline for multimodal emotion recognition. (a) A VAE with a ResNet-101 backbone is first trained on a large facial dataset to learn robust facial representations; (b) The DMMVAE framework is trained using both EEG and facial data, with separate private latent spaces for each modality and a shared latent space for cross-modal feature learning. The pre-trained ResNet-101 VAE is integrated into the facial encoder-decoder; (c) The trained encoder outputs from DMMVAE are concatenated and passed to an LSTM network for emotion classification, leveraging temporal dependencies in multimodal data; (d) An LSTM unit. DMMVAE: disentangled multimodal variational autoencoder; LSTM: long short-term memory; EEG: electroencephalogram; VAE: variational autoencoder.

The Chicago face database[68] contains high-quality facial images with norming data and is widely used in computational and psychological studies on face perception. The FERET database[69], developed to support face recognition research, includes facial images with variations in pose, lighting, and expression, and serves as a benchmark for training and evaluation. The FEEDTUM database[70] focuses on dynamic and spontaneous facial expressions, providing video sequences of genuine emotional displays commonly used in emotion recognition and facial dynamics analysis. The MUCT landmarked face database[71] contains facial images with manually annotated landmarks, supporting tasks such as facial alignment, recognition, and landmark detection, and includes diverse facial shapes and expressions. The person face dataset[72] contains synthetic, GAN-generated facial images, offering an unlimited supply of photorealistic and privacy-compliant faces that enhance model robustness through increased diversity. By leveraging these diverse datasets, the pretraining process improves the generalization and robustness of the facial encoder-decoder within the DMMVAE framework.

For training, we first applied multi-task cascaded convolutional networks (MTCNN) for face detection and alignment to ensure consistency across samples. The detected faces were then converted to RGB format, resized to 224 × 224 pixels, and normalized based on dataset statistics before being fed into the VAE.

The dataset was split into 80% for training and 20% for validation. Training was conducted for 10 epochs, as additional epochs yielded

only marginal improvements while increasing computational cost. These settings enabled the model to reconstruct both EEG and facial modalities effectively while minimizing the risk of overfitting or divergence. Since our primary objective is to learn a meaningful latent space, we used $\beta = 1$, corresponding to a vanilla VAE, to ensure a balance reconstruction fidelity and latent space regularization. Although higher β values (i.e., $\beta > 1$) can improve the disentanglement of latent features, they often reduce reconstruction accuracy, which is essential for effective feature extraction in our case[73]. The pretrained VAE encoder and decoder were integrated into the facial encoder-decoder within the DMMVAE framework. This pretraining step significantly reduced reconstruction loss compared to training solely on the DEAP and MAHNOB datasets, resulting in more stable and generalizable facial feature representations.

4.2 DMMVAE training

In this study, we incorporated both EEG signals and facial expression. Each modality was processed using a separate encoder-decoder networks to ensure that modality-specific features were effectively extracted and reconstructed before being integrated into a shared latent representation. The EEG encoder-decoder adopts a 1D-CNN architecture to process raw EEG signals. The input consists of 32 channels by 128 time points, representing one second of EEG data sampled at 128 Hz. The encoder compresses the EEG signals into a lower-dimensional latent representation that captures essential temporal patterns relevant to emotion recognition. The decoder reconstructs the EEG signals from this latent representation using a two-step transformation. First, a fully connected layer maps the latent space to a 512-dimensional vector, followed by a ReLU activation function to introduce nonlinearity. Second, the transformed output is reshaped to the original EEG format (32 × 128), ensuring that the reconstructed signals closely match the input. For the facial modality, we employ a pretrained ResNet-101-based VAE to extract and reconstruct facial features. Instead of using every frame in the video sequence, we extract the middle frame of each one-second interval to reduce redundancy while maintaining temporal consistency. Similar to the VAE training stage, we first apply MTCNN to detect and extract facial regions. The detected faces are then aligned and cropped before being processed by the ResNet-101-based encoder. This encoder captures high-level facial representations and encodes them into a low-dimensional latent space that retains both structural and expression-related features. The decoder then reconstructs facial images from these latent representations. After extracting EEG and facial features, the latent representations from both modalities are integrated into the DMMVAE framework. The latent space is structured into one shared latent space (z_s), which captures features common to both modalities, and two private latent spaces (z_{pEEG} and z_{pface}), which preserve modality-specific information. The shared latent space facilitates the learning of generalized representations across modalities, whereas the private latent spaces maintain domain-specific characteristics. This disentangled structure enhances the interpretability and robustness of multimodal feature learning, contributing to more reliable emotion recognition.

The DMMVAE was trained to reconstruct both EEG and facial inputs using the following learning objective:

$$\begin{aligned}&\sum_i E_{p(x_i)}\Big[\lambda_i E_{q_\phi(zp_i|x_i)q_\phi(z_s|x)}\Big[\log p_\theta\left(x_i \mid zp_i, z_s\right)\Big]\\&- KL\left(q_\phi\left(zp_i \mid x_i\right) \| \, p\left(zp_i\right)\right) - KL\left(q_\phi\left(z_s \mid x\right) \| \, p\left(z_s\right)\right)\Big]\\&+ \sum_j \Big[\lambda_i E_{q_\phi(zp_i|x_i)q_\phi(z_s|x_j)}\Big[\log p_\theta\left(x_i \mid zp_i, z_s\right)\Big]\\&- KL\left(q_\phi\left(zp_i \mid x_i\right) \| \, p\left(zp_i\right)\right) - KL\left(q_\phi\left(z_s \mid x_j\right) \| \, p\left(z_s\right)\right)\Big]\end{aligned} \tag{18}$$

where: balances the reconstruction loss across different modalities.

The first term ensures accurate reconstruction of each modality while compensating for the divergence from the prior distribution. The second term evaluates cross-modal reconstruction, which involves reconstructing one modality from the other. This objective function enables the shared latent space to encode meaningful common features across modalities while preserving the structural integrity of each input modality. The KL divergence regularization mitigates overfitting and encourages the latent representations to remain close to a standard normal distribution. The DMMVAE model was trained using the same settings as the VAE, including a batch size of 32, a learning rate of 0.001, and 10 training epochs. The dataset was split into 80% for training and 20% for validation. After training, the encoder-decoder of the DMMVAE was used to extract latent representations for emotion classification. This process is illustrated in Figure 3b.

4.3 DMMVA-LSTM-based emotion classification

After training the DMMVAE, we extracted both shared and private latent representations from the EEG and facial modalities. These representations were concatenated and fed into an LSTM network (Figure 3c,d) for emotion classification:

$$z_{\text{concat}} = \left[z_s, zp_{\text{EEG}}, zp_{\text{face}}\right] \tag{19}$$

This concatenation ensures that modality-invariant and modality-specific features contribute to the final classification. By combining the shared and private latent spaces, the model retains generalizable features as well as fine-grained details unique to EEG and facial data. Emotional states evolve over time and are reflected in the sequential structure of EEG recordings and facial image streams, making temporal modelling essential for accurate recognition. EEG captures subtle, continuous neurophysiological changes, while facial expressions change dynamically across video frames. The LSTM processes the sequence of concatenated latent features and selectively preserves relevant information through its gating mechanisms.

We used a two-layer LSTM with 128 hidden units to capture these temporal dependencies in the emotion data. The LSTM classifier was trained using binary cross-entropy loss over five epochs, with a dropout rate of 0.3 applied to prevent overfitting. This LSTM-based classifier enables the model to learn and generalize emotion patterns across time. The final hidden state h_t is passed through a fully connected layer with softmax activation for binary emotion classification:

$$\hat{y} = \text{softmax}\left(W \cdot h_t + b\right) \tag{20}$$

5. Experimental Results and Discussion

In this section, we provide a detailed description of the accuracy of emotion classification. Given the significant impact of emotion recognition across various domains, we present both classification results and an analysis of classification robustness to better understand the model's performance.

5.1 Dataset

We used the publicly available DEAP[74] and MAHNOB[75] datasets to evaluate the proposed multimodal emotion recognition framework. Both datasets are widely used multimodal human emotion datasets designated exclusively for academic research, with each participant having signed a consent form allowing their data to be used for research purposes and imagery to be published.

The DEAP dataset contains electroencephalogram (EEG) and peripheral physiological signals from 32 participants and frontal face videos from 22 participants, all exposed to 40 one-minute emotional stimuli videos. The EEG signals underwent several preprocessing steps performed by the data providers to enhance signal quality and ensure consistency, including downsampling to 128 Hz, electrooculographic (EOG) artifact removal, and band-pass filtering between 4 and 45 Hz. Additionally, signals were re-referenced using a common average reference to reduce noise and improve signal quality. The MAHNOB-HCI dataset contains EEG, video, audio, gaze, and peripheral physiological data from 30 subjects, with stimulant video durations ranging between 35 and 117 seconds. Similar to DEAP, preprocessing steps were applied by the providers to ensure data quality and usability, including downsampling to 128 Hz, artifact removal (e.g., EOG and muscle artifacts), and band-pass filtering between 4 and 45 Hz to preserve relevant frequency components while minimizing noise. Finally, signal normalization and re-referencing were applied to enhance consistency across recordings, ensuring suitability for emotion recognition studies. A summary of these datasets is shown in Table 1.

Table 1. Statistics of the datasets used in this work.

Criterion	DEAP	MAHNOB-HCI
Dataset size	22 subjects × 40 trials	30 subjects × 20 trials
Stimulus duration	60 s	49.7-117 s
Modality	Vision, EEG, Peripheral	Vision, EEG, Peripheral
Used bio signals	EEG, EOG, EMG, GSR, Respiration belt, Plethysmograph, Temperature	EEG, ECG, GSR, Respiration belt,Temperature
Labels	Continuous valence and arousalfrom 1 to 9	Discrete valence and arousalof integers from 1 to 9

EEG: electroencephalogram; EOG: electrooculographic; EMG: electromyography; GSR: alvanic skin response.

In both DEAP and MAHNOB-HCI, emotions are rated on a scale from 1 to 9, based on Russell's arousal-valence space[76]. For emotion classification, we used a binary classification approach by grouping ratings into low (1-5) and high (5-9) levels along the arousal-valence dimensions. For DEAP, only data from participants 1 to 22 were used due to the availability of facial videos. Additionally, nine trials were removed where objects or hands obscured the participant's face during part of the trial. The recorded trials consist of 60-second segments preceded by a 3-second pre-trial baseline. Since the baseline period is not part of the emotional response, the first 3 seconds were excluded, and the remaining 60 seconds were used for analysis. For MAHNOB-HCI, 526 trials with complete frontal video and physiological signals were selected. Due to variable video lengths (approximately 30 to 117 seconds), the analysis was standardized by selecting the middle 30 seconds of each trial and extracting EEG data from the corresponding 30-second segment to ensure alignment between EEG and facial expressions.

Each trial was divided into 1-second segments for further processing to facilitate fair comparison with previous studies. The middle frame from each video segment was extracted to synchronize temporally with the corresponding EEG data for each second, ensuring aligned multimodal inputs.

5.2 Comparison with baselines and previous work

In this section, we compare our results with previous state of the art methods and baseline models to evaluate the effectiveness of our proposed DMMVAE LSTM model. Before assessing classification performance, we first examine the reconstruction loss of the DMMVAE component when trained with and without a pre trained VAE on the DEAP dataset. This analysis helps determine the impact of pre training on the quality of learned representations. The reconstruction loss for the DMMVAE with a pre trained VAE was 0.0284, whereas without pre training it was 0.1031, indicating that pre training significantly improves reconstruction quality. This suggests that the pre trained VAE effectively captures meaningful image representations, benefiting downstream multimodal learning. Following this preliminary analysis, we evaluated classification performance by comparing our model against previous state of the art approaches and two baseline models. Both datasets were split into training and test sets with an 80 percent 20 percent ratio. For both baseline and proposed experiments, we performed 5 fold cross validation and reported the average testing accuracy as the performance metric. The first baseline was a standard LSTM based approach where separate LSTMs were trained independently for each modality one for EEG and one for facial features. The last hidden state from each LSTM was extracted, and a late fusion strategy concatenated these outputs before passing them through a fully connected layer for final classification. Unlike our DMMVAE model, which explicitly disentangles shared and modality specific features, this baseline relied solely on sequential modeling without a structured latent representation for multimodal fusion. The second baseline employed the same DMMVAE plus LSTM architecture on the DEAP dataset; however, the DMMVAE component did not use a pre trained VAE for the image encoder and decoder. This setup isolated the effect of pre training on multimodal representation and allowed us to better understand how structured latent space learning impacts classification performance.

Table 2 summarizes the mean classification accuracies and standard deviations for valence and arousal across different models on the DEAP and MAHNOB HCI datasets. On DEAP, our model achieved 96.82 percent accuracy in valence classification, outperforming both baselines and prior methods, while its 96.90 percent accuracy in arousal classification was comparable to previous results[49]. On MAHNOB HCI, our model surpassed all previous valence and arousal classification approaches, demonstrating notable improvements in emotion recognition. The baseline LSTM model, which lacks the DMMVAE framework, showed consistently lower classification performance on both datasets. This finding suggests that learning shared and modality specific latent representations is beneficial for capturing cross modal dependencies while preserving unique characteristics of each modality. In addition, a paired t-test was performed on the 5-fold cross-validation results, comparing the performance of DMMVAE-LSTM and the LSTM model. The results indicate that DMMVAE significantly outperforms the LSTM model (Average $p\text{-}value \approx 0.0056$ for DEAP and $p\text{-}value \approx 0.0119$ for MAHNOB). These results indicate that the observed performance gains are unlikely due to random chance. Overall, our findings reinforce the effectiveness of the DMMVAE framework in enhancing multimodal emotion recognition, particularly for valence classification, where it achieves state of the art performance on both datasets.

Table 2. Comparison of Valence and Arousal Accuracy (Mean ± SD) and Average F1 Score for Different Datasets.

Dataset	Method	Accuracy (Mean ± SD)	Average F1 Score
DEAP	CNN-based[49]	Valence: 96.63%Arousal: **97.15%**	--
	LSTM-based[50]	Valence: 94.94%Arousal: 95.30%	--
	CNN-based[51]	Valence: 77.08%Arousal: 70.01%	--
	Ours (LSTM)	Valence: 89.34% ± 2.8Arousal: 90.67% ± 2.9	0.88300.8880
	Ours (DMMVAE-LSTM) without Pre-trained VAE	Valence: 92.79% ± 2.1Arousal: 93.09% ± 1.9	0.92150.9176
	Ours (DMMVAE-LSTM) with Pre-trained VAE	Valence: **96.82%** ± **1.8**Arousal: 96.90% ± 1.7	**0.96400.9623**
MAHNOB-HCI	CNN-based[49]	Valence: 96.69%Arousal: 96.25%	--
	Ours (LSTM)	Valence: 92.12% ± 2.7Arousal: 91.47% ± 3	0.91580.9123
	Ours (DMMVAE-LSTM) without Pre-trained	Valence: 93.92% ± 1.8Arousal: 93.18% ± 1.9	0.93390.9294
	Ours (DMMVAE-LSTM) with Pre-trained	Valence: **97.12%** ± **1.5**Arousal: **96.94%** ± **1.7**	**0.96880.9621**

CNN: convolutional neural network; LSTM: long short-term memory; DMMVAE: disentangled multimodal variational autoencoder; SD: standard deviation.

5.3 Robustness analysis with limited data

To further evaluate the effectiveness of our private-shared disentangled MMVAE combined with LSTM model, we conducted additional experiments using only 50% and 75% of the available data. This was done to assess the model's performance under data-limited conditions and to compare its stability and robustness against a standard LSTM-only baseline.

We perform 5-fold cross-validation and report the average testing accuracy and F1-score as the primary metrics of model performance. As shown in Table 3, although both models experienced a decline in performance due to the reduced training data, the private-shared disentangled DMMVAE-LSTM consistently outperformed the LSTM baseline. This advantage is attributed to the generative nature of DMMVAE, which allows the model to learn a more structured and robust latent representation even with fewer training samples. Unlike traditional discriminative models such as LSTM, which rely heavily on direct feature mapping from input to output, DMMVAE exploits its variational latent space to extract meaningful representations that generalize better in data-scarce scenarios.

Table 3. Comparison of valence and arousal accuracy (Mean ± SD) and average F1 score for different datasets.

Dataset	Model	Accuracy	Average F1 Score
DEAP	LSTM	Valence: 89.34% ± 2.8Arousal: 90.67% ± 2.9	0.88300.8880
	LSTM (75%)	Valence: 82.01% ± 2.8Arousal: 83.08% ± 3.1	0.80100.7972
	LSTM (50%)	Valence: 64.14% ± 3.0Arousal: 64.27% ± 3.2	0.61190.6098
	DMMVAE-LSTM	Valence: 96.82% ± 1.8Arousal: 96.90% ± 1.7	0.96400.9623
	DMMVAE-LSTM (75%)	Valence: 91.53% ± 2.1Arousal: 91.62% ± 2.2	0.90580.9067
	DMMVAE-LSTM (50%)	Valence: 77.32% ± 2.4Arousal: 77.92% ± 2.6	0.78230.7751
MAHNOB-HCI	LSTM	Valence: 92.12% ± 2.7Arousal: 91.47% ± 3.0	0.91580.9123
	LSTM (75%)	Valence: 85.63% ± 2.9Arousal: 87.39% ± 3.1	0.84060.8543
	LSTM (50%)	Valence: 64.22% ± 3.2Arousal: 66.87% ± 3.1	0.66070.6695
	DMMVAE-LSTM	Valence: 97.12% ± 1.5Arousal: 96.94% ± 1.7	0.96880.9621
	DMMVAE-LSTM (75%)	Valence: 92.02% ± 1.8Arousal: 91.94% ± 2.1	0.91640.9127
	DMMVAE-LSTM (50%)	Valence: 77.16% ± 2.0Arousal: 76.80% ± 2.2	0.76830.7581

DMMVAE: disentangled multimodal variational autoencoder; LSTM: long short-term memory; EEG: electroencephalogram; CNN: convolutional neural network; SD: standard deviation.

Moreover, the disentangled latent representation in DMMVAE preserves cross-modal relationships and modality-specific information, mitigating the negative effects of limited data availability. By effectively capturing both shared and private feature spaces, DMMVAE can compensate for insufficient training samples by generating more informative latent embeddings, thereby enhancing classification performance. The model's ability to maintain performance under limited data conditions further demonstrates its suitability for multimodal emotion recognition tasks, making it a promising approach for applications involving small or imbalanced datasets.

5.4 Robustness analysis with data dropout

To evaluate the resilience of our proposed framework in handling missing data, we conducted an additional experiment simulating real-world challenges such as sensor failures or transmission errors, where specific modalities may be partially unavailable. This was achieved by randomly dropping 10% of the input data during training. Given the MMVAE's capability to model latent distributions and infer missing features, we expected the model to maintain stable performance despite data dropout. For this experiment, only the DEAP dataset was used, as it provides longer EEG sequences compared to the MAHNOB dataset. Longer sequences offer more contextual information, enabling the LSTM to better capture temporal dependencies and compensate for missing data. Using the DEAP dataset ensured that the model had sufficient temporal context to infer missing features, thereby reducing the negative impact of data dropout. All results reported represent the average of 5-fold cross-validation. The results in Table 4 demonstrate that the DMMVAE-LSTM model remains robust even when part of the input data is missing. This robustness arises from the private-shared disentangled MMVAE framework, which separates the latent space into shared representations (z_s) that capture common features across EEG and facial modalities, and private representations (z_{pi}) that preserve modality-specific information. While private representations are essential for capturing unique characteristics of each modality, they are also more vulnerable to noise,

particularly when trained on short data segments. The LSTM component alleviates this issue by integrating information over time, smoothing inconsistencies, and producing a more stable latent feature set for classification. Furthermore, the LSTM's memory mechanism plays a crucial role in maintaining temporal continuity by leveraging both past and present information, leading to more reliable predictions even with missing data. These findings underscore the robustness of the DMMVAE-LSTM model in real-world scenarios where missing data is common. The model's ability to reconstruct meaningful latent representations and effectively incorporate temporal dependencies reinforces its potential for multimodal emotion recognition in practical, data-limited environments.

Table 4. Comparison of valence and arousal accuracy (Mean ± SD) and average F1 score under different missing modality scenarios (DEAP dataset).

Model	Missing rate	Accuracy (Mean ± SD)	Average F1 Score
LSTM	0.0%	Valence: 89.34% ± 2.8Arousal: 90.67% ± 2.9	0.88300.8880
	10% Face	Valence: 81.22% ± 3.1Arousal: 81.78% ± 3.0	0.78590.7812
	10% EEG	Valence: 78.07% ± 3.2Arousal: 78.53% ± 3.2	0.75730.7490
DMMVAE-LSTM	0.0%	Valence: 96.82% ± 1.8Arousal: 96.90% ± 1.7	0.96400.9623
	10% Face	Valence: 90.14% ± 2.0Arousal: 91.48% ± 2.1	0.89140.9117
	10% EEG	Valence: 87.16% ± 2.2Arousal: 87.93% ± 2.4	0.86100.8578

DMMVAE: disentangled multimodal variational autoencoder; LSTM: long short-term memory; EEG: electroencephalogram; SD: standard deviation.

5.5 Visual analysis of learned representations

To gain deeper insight into what the model learns during training, we visualize feature activation map from the ResNet-101-based encoder within the facial VAE. Figure 4 presents sample activation maps extracted from intermediate and deeper convolutional layers for selected samples from the DEAP and MAHNOB-HCI datasets. As shown, the convolutional layers respond to a variety of visual features, including facial edges, illumination patterns, and regions typically associated with affective expressions such as the eyes and mouth. This hierarchical progression of feature abstraction indicates that the encoder effectively captures rich facial representations essential for downstream multimodal emotion recognition tasks. These visualizations substantiate the claim that the pretrained VAE generates semantically meaningful embeddings that generalize across subjects and contribute to improve classification performance, especially when combined with EEG features within the DMMVAE-LSTM framework.

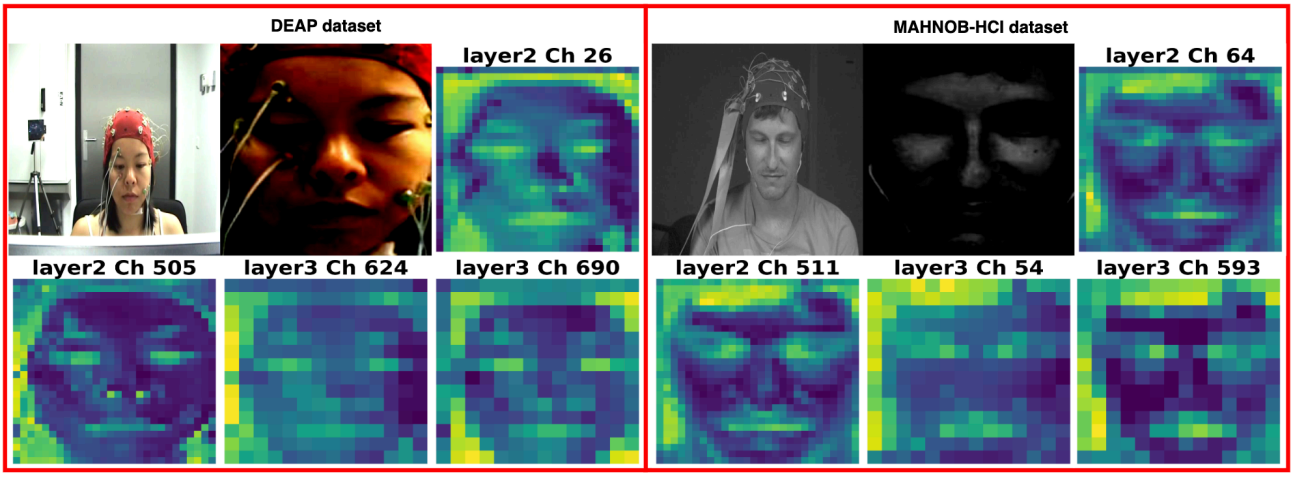

Figure 4. Visualization of four activation maps from intermediate and deep layers of the ResNet-101-based encoder used in the facial VAE. Each row corresponds to a sample subject from the DEAP and MAHNOB-HCI datasets. VAE: variational autoencoder.

6. Conclusion

In this study, we proposed a novel framework for multimodal emotion recognition by integrating the private-shared disentangled MMVAE with LSTM networks. Our approach effectively disentangles shared and modality-specific latent representations, enabling a more comprehensive understanding of emotional states from EEG signals and facial expressions. By leveraging the strengths of both generative and sequential models, the framework captures the spatial and temporal dependencies inherent in multimodal emotion

recognition. Experimental results demonstrate that our model competitive performance compared to state-of-the-art unimodal and multimodal methods in valence and arousal classification tasks across various experimental settings. We validated the model's robustness through two experimental scenarios: first, using the full DEAP and MAHNOB-HCI datasets; second, using 50% and 75% of the data, where the proposed approach consistently maintained strong classification performance. This highlights the generative capabilities of DMMVAE in scenarios with limited training data. Furthermore, we evaluated the impact of modality dropout by introducing random data omission during training. The results showed that DMMVAE-LSTM remains robust despite modality dropout, underscoring the benefits of its disentangled latent space and sequential learning mechanisms. Our study also emphasizes the advantages of multimodal emotion recognition by demonstrating how different modalities compensate each other's limitations, resulting in improved generalization and robustness. The framework's ability to sustain high accuracy even with reduced data suggests its potential applicability in real-world environments where data collection is constrained. To further explore the significance of facial feature extraction, we conducted an ablation study comparing DMMVAE-LSTM models trained with and without a pretrained VAE. The findings confirmed that incorporating a pretrained VAE significantly enhances emotion classification performance by providing a more structured and informative latent space for facial representations.

In addition, our study underscores the advantages of multimodal emotion recognition by demonstrating how different modalities can compensate for each other's limitations, leading to improved generalization and robustness. The ability of our framework to maintain high accuracy even under reduced data conditions suggests its strong potential for real-world applications where data collection may be limited or inconsistent.

Nevertheless, several open challenges remain that may affect the generalizability and practical deployment of the proposed framework. While multimodal emotion recognition benefits from leveraging complementary information across modalities, in practical scenarios it is not guaranteed that data from all modalities will be consistently available. Future research should focus on enhancing the model's robustness and ensuring reliable performance in real-world environments through adaptive learning strategies, such as test-time adaptation. Another limitation concerns individual variability, which can impact model performance across different users. Domain adaptation techniques could help mitigate this issue by enabling the model to generalize across diverse populations. Additionally, future work could explore the integration of further physiological and behavioral modalities, such as heart rate, to construct a more comprehensive and nuanced representation of emotional states.

Finally, the current evaluation was limited to the DEAP and MAHNOB-HCI datasets. Although these datasets are widely used in the field, they are relatively constrained in terms of participant diversity and data scale. Evaluating the proposed model on additional, more diverse emotion recognition datasets would further validate its generalizability and robustness across broader real-world conditions.

Declarations

Authors contribution

Mahaseni B: Performed the experiments, data analysis, result interpretation, manuscript preparation.
Khan NM: Designed the study, data analysis, result interpretation, manuscript preparation.
All authors reviewed and approved the final manuscript.

Conflicts of interest

All authors declared that there are no conflicts of interest.

Ethical approval

The authors used the publicly available DEAP and MAHNOB datasets to evaluate the proposed multimodal emotion recognition framework. Both datasets are widely used multimodal human emotion corpora designated exclusively for academic research.

Consent to participate

All participants in the DEAP and MAHNOB studies provided written informed consent to participate in the original data collection.

Consent for publication

Not applicable.

Availability of data and materials

The datasets used and analyzed in this study are publicly available. The DEAP dataset is available at http://www.eecs.qmul.ac.uk/mmv/datasets/deap/. The MAHNOB-HCI dataset can be obtained upon request via the official project page at https://mahnob-db.eu.

Funding

This research was funded by Natural Sciences and Engineering Research Council of Canada (RGPIN-2020-05471 and Alliance 577563-2022).

References

1. Bower GH. Mood and memory. Am Psychol. 1981;36(2):129-148. [DOI]
2. Izard CE. Basic emotions, relations among emotions, and emotion-cognition relations. Psychol Rev. 1992;99(3):561-565. [DOI]
3. De Prisco M, Oliva V, Fico G, Montejo L, Possidente C, Bracco L, et al. Differences in facial emotion recognition between bipolar disorder and other clinical populations: A systematic review and meta-analysis. Prog Neuropsychopharmacol Biol Psychiatry. 2023;127:110847. [DOI]
4. Lasri I, Riadsolh A, Elbelkacemi M. Facial emotion recognition of deaf and hard-of-hearing students for engagement detection using deep learning. Educ Inf Technol. 2023;28(4):4069-4092. [DOI]
5. Gupta S, Kumar P, Tekchandani RK. Facial emotion recognition based real-time learner engagement detection system in online learning context using deep learning models. Multimed Tools Appl. 2023;82(8):11365-11394. [DOI]
6. Jain DK, Dutta AK, Verdú E, Alsubai S, Sait ARW. An automated hyperparameter tuned deep learning model enabled facial emotion recognition for autonomous vehicle drivers. Image Vis Comput. 2023;133:104659. [DOI]
7. Dehghani F, Zaman L. Facial emotion recognition in VR games. In: 2023 IEEE Conference on Games (CoG); 2023 Aug 21-24; Boston, USA. New York: IEEE; 2023. p. 1-4. [DOI]
8. Tyng CM, Amin HU, Saad MN, Mailk AS. The influences of emotion on learning and memory. Front Psychol. 2017;8:235933. [DOI]
9. LeDoux JE. Emotion circuits in the brain. Annu Rev Neurosci. 2000;23:155-184. [DOI]
10. Wu M, Goodman N. Multimodal generative models for scalable weakly-supervised learning. In: Bengio S, Wallach HM, Larochelle H, Grauman K, Cesa-Bianchi N, editors. Proceedings of the 32nd International Conference on Neural Information Processing Systems; 2018 Dec 3-8; Montréal, Canada. USA: Curran Associates Inc.; 2018. p. 5580-5590. Available from: https://proceedings.neurips.cc/paper_files/paper/2018/file/1102a326d5f7c9e04fc3c89d0ede88c9-Paper.pdf
11. Kingma DP, Welling M. An introduction to variational autoencoders. Hanover: Foundation & Trends; 2019.
12. Lee M, Pavlovic V. Private-shared disentangled multimodal VAE for learning of latent representations. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW); 2021 Jun 19-25; Nashville, USA. New York: IEEE; 2021. p. 1692-1700. [DOI]
13. Yang K, Wang C, Gu Y, Sarsenbayeva Z, Tag B, Dingler T. Behavioral and physiological signals-based deep multimodal approach for mobile emotion recognition. IEEE Trans Affective Comput. 2023;14:1082-1097. [DOI]
14. Li Y, Wang M, Gong M, Liu Y, Liu L. Fer-former: Multimodal transformer for facial expression recognition. IEEE Trans Multimedia. 2025;27:2412-2422. [DOI]
15. Singh R, Saurav S, Kumar T, Saini R, Vohra A, Singh S. Facial expression recognition in videos using hybrid CNN & ConvLSTM. Int J Inf Tecnol. 2023;15:1819-1830. [DOI]
16. Sarvakar K, Senkamalavalli R, Raghavendra S, Kumar JS, Manjunath R, Jaiswal S. Facial emotion recognition using convolutional neural networks. Mater Today Proc. 2023;80:3560-3564. [DOI]
17. Karatay B, Beştepe D, Sailunaz K, Özyer T, Alhajj R. CNN-transformer based emotion classification from facial expressions and body gestures. Multimed Tools Appl. 2024;83(8):23129-23171. [DOI]
18. Hangaragi S, Singh T, Neelima N. Face detection and recognition using face mesh and deep neural network. Procedia Comput Sci. 2023;218:741-749. [DOI]
19. Akinpelu S, Viriri S, Adegun A. An enhanced speech emotion recognition using vision transformer. Sci Rep. 2024;14:13126. [DOI]
20. Ye J, Wen XC, Wei Y, Xu Y, Liu K, Shan H. Temporal modeling matters: A novel temporal emotional modeling approach for speech emotion recognition. In: ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2023 Jun 4-10; Rhodes Island, Greece. New York: IEEE; 2023. p. 1-5. [DOI]
21. Kumar S, Haq MA, Jain A, Jason CA, Moparthi NR, Mittal N, et al. Multilayer neural network based speech emotion recognition for smart assistance. Comput Mater Contin. 2023;74(1):1523-1540. [DOI]
22. Vaijayanthi S, Arunnehru J. Human emotion recognition from body posture with machine learning techniques. In: Singh M, Tyagi V, Gupta PK, Flusser J, Ören T, editors. Advances in Computing and Data Sciences. ICACDS 2022. Communications in Computer and Information Science; 2022 Apr 22-23; Kurnool, India. Cham: Springer; 2022. p. 231-242. [DOI]
23. Wang L, Hao J, Zhou TH. ECG multi-emotion recognition based on heart rate variability signal features mining. Sensors. 2023;23(20):8636. [DOI]
24. Fang A, Pan F, Yu W. ECG-based emotion recognition using random convolutional kernel method. Biomed Signal Process Control. 2024;91:105907. [DOI]
25. Liu S, Wang Z, An Y. EEG emotion recognition based on the attention mechanism and pre-trained convolution capsule network. Knowl

Based Syst. 2023;265:110372. [DOI]
26. Wei Y, Liu Y, Li C, Cheng J, Song R, Chen X. TC-Net: A transformer capsule network for EGG-based emotion recognition. Comput Biol Med. 2023;152:106463. [DOI]
27. Iyer A, Das SS, Teotia R, Maheshwari S, Sharma RR. CNN and LSTM based ensemble learning for human emotion recognition using EEG recordings. Multimed Tools Appl. 2023;82:4883-4896. [DOI]
28. Hassan F, Hussain SF, Qaisar SM. Fusion of multivariate EEG signals for schizophrenia detection using CNN and machine learning techniques. Inf Fusion. 2023;92:466-478. [DOI]
29. Wang Y, Guan X. Multimodal feature fusion and emotion recognition based on variational autoencoder. In: 2023 IEEE 5th International Conference on Civil Aviation Safety and Information Technology (ICCASIT); 2023 Oct 11-13; Dali, China. New York: IEEE; 2023. p. 819-823. [DOI]
30. Wang T, Zhang M, Shang L. DisVAE: Disentangled variational autoencoder for high-quality facial expression features. In: 2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG); 2023 Jan 5-8; Waikoloa Beach, USA. New York: IEEE; 2023. p. 1-8. [DOI]
31. Zhao T, Cui Y, Ji T, Luo J, Li W, Jiang J, et al. VAEEG: Variational auto-encoder for extracting EEG representation. NeuroImage. 2024;304:120946. [DOI]
32. Bethge D, Hallgarten P, Grosse-Puppendahl T, Kari M, Chuang LL, Özdenizci O. EEG2Vec: Learning affective EEG representations via variational autoencoders. In: 2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC); 2022 Oct 9-12; Prague, Czech Republic. New York: IEEE; 2022. p. 3150-3157. [DOI]
33. Gunes H, Pantic M. Automatic, dimensional and continuous emotion recognition. Int J Synth Emot. 2010;1(1):68-99. [DOI]
34. Zhang J, Yin Z, Chen P, Nichele S. Emotion recognition using multi-modal data and machine learning techniques: a tutorial and review. Inf Fusion. 2020;59:103-126. [DOI]
35. Tseng HT, Hsieh CC, Xu CH. A two-stage multimodal emotion analysis using body actions and facial features. Signal Image Video Process. 2015;19:313. [DOI]
36. Liu J, Wang Z, Nie W, Zeng J, Zhou B, Deng J, et al. Multimodal emotion recognition for children with autism spectrum disorder in social interaction. Int J Hum Comput Interact. 2023;40(8):1921-1930. [DOI]
37. Dessai A, Virani H. Multimodal and multidomain feature fusion for emotion classification based on electrocardiogram and galvanic skin response signals. Sci. 2024;6(1):10. [DOI]
38. Lopez E, Chiarantano E, Grassucci E, Comminiello D. Hypercomplex multimodal emotion recognition from EEG and peripheral physiological signals. In: 2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW); 2023 Jun 4-10; Rhodes Island, Greece. New York: IEEE; 2023. p. 1-5. [DOI]
39. Pan J, Fang W, Zhang Z, Chen B, Zhang Z, Wang S. Multimodal emotion recognition based on facial expressions, speech, and EGG. IEEE Open J Eng Med Biol. 2023;5:396-403. [DOI]
40. Yin G, Liu Y, Liu T, Zhang H, Fang F, Tang C, et al. Token-disentangling Mutual Transformer for multimodal emotion recognition. Intelligence. 2024;133:108348. [DOI]
41. Ali K, Hughes CE. A unified transformer-based network for multimodal emotion recognition. arXiv:2308.14160 [Preprint]. 2023. [DOI]
42. Fu B, Gu C, Fu M, Xia Y, Liu Y. A novel feature fusion network for multimodal emotion recognition from EEG and eye movement signals. Front Neurosci. 2023;17:1234162. [DOI]
43. Cheng C, Liu W, Fan Z, Feng L, Jia Z. A novel transformer autoencoder for multi-modal emotion recognition with incomplete data. Neural Netw. 2024;172:106111. [DOI]
44. Wu Y, Daoudi M, Amad A. Transformer-based self-supervised multimodal representation learning for wearable emotion recognition. IEEE Trans Affective Comput. 2024;15(1):157-172. [DOI]
45. Wang R, Zhu J, Wang S, Wang T, Huang J, Zhu X. Multi-modal emotion recognition using tensor decomposition fusion and self-supervised multi-tasking. Int J Multimed Info Retr. 2024;13:39. [DOI]
46. Singh P, Tripathi MK, Patil MB, Neelakantappa M. Multimodal emotion recognition model via hybrid model with improved feature level fusion on facial and EEG feature set. Multimed Tools Appl. 2025;84:1-36. [DOI]
47. Mutawa A, Hassouneh A. Multimodal real-time patient emotion recognition system using facial expressions and brain EEG signals based on machine learning and Log-Sync methods. Biomed Signal Process Control. 2024;91:105942. [DOI]
48. Chen Y, Bai Z, Cheng M, Liu Y, Zhao X, Song Y. Multimodal emotion recognition for hearing-impaired subjects by fusing EGG signals and facial expressions. In: 2023 42nd Chinese Control Conference (CCC); 2023 Jul 24-26; Tianjin, China. New York: IEEE; 2023. p. 1-6. [DOI]
49. Wang S, Qu J, Zhang Y, Zhang Y. Multimodal emotion recognition from EEG signals and facial expressions. IEEE Access. 2023;11:33061-33068. [DOI]
50. Wu Y, Li J. Multi-modal emotion identification fusing facial expression and EEG. Multimed Tools Appl. 2023;82:10901-10919. [DOI]
51. Zhu Q, Lu G, Yan J. Valence-arousal model based emotion recognition using EEG, peripheral physiological signals and facial expression. In: Proceedings of the 4th International Conference on Machine Learning and Soft Computing; 2020 Jan 17-19; Haiphong City Viet Nam. New York: Association for Computing Machinery; 2020. p. 81-85. [DOI]
52. Koromilas P, Giannakopoulos T. Deep multimodal emotion recognition on human speech: a review. Appl Sci. 2021;11(17):7962. [DOI]
53. Poria S, Cambria E, Bajpai R, Hussain A. A review of affective computing: from unimodal analysis to multimodal fusion. Inf Fusion. 2017;37:98-125. [DOI]

54. Ramachandram D, Taylor GW. Deep multimodal learning: a survey on recent advances and trends. IEEE Signal Process Mag. 2017;34(6):96-108. [DOI]
55. Kingma DP, Welling M. Auto-encoding variational bayes. arXiv:1312.6114 [Preprint]. 2013. [DOI]
56. Lawry Aguila A, Chapman J, Altmann A. Multi-modal variational autoencoders for normative modelling across multiple imaging modalities. In: Greenspan H, Madabhushi A, Mousavi P, Salcudean S, Duncan J, Syeda-Mahmood T, Taylor R, editors. Medical Image Computing and Computer Assisted Intervention - MICCAI 2023. Lecture Notes in Computer Science; 2023 Oct 8-12; Vancouver, Canada. Cham: Springer; 2023. p. 425-434. [DOI]
57. Kumar S, Qiu P, Yang B, Bani A, Payne PRO, Sotiras A. Multimodal normative modeling in alzheimer's disease with introspective variational autoencoders. bioRxiv. 2024. [DOI]
58. Martí-Juan G, Lorenzi M, Piella G; Alzheimer's Disease Neuroimaging Initiative. MC-RVAE: Multi-channel recurrent variational autoencoder for multimodal Alzheimer's disease progression modelling. Neuroimage. 2023;268:119892. [DOI]
59. Chen R, Zhou W, Hu H, Fei Z, Fei M, Zhou H. Disentangled variational auto-encoder for multimodal fusion performance analysis in multimodal sentiment analysis. Knowl Based Syst. 2024;301:112372. [DOI]
60. Shi T, Wei Y, Kender JR. An efficient and explanatory image and text clustering system with multimodal autoencoder architecture. arXiv:2408.07791 [Preprint]. 2024. [DOI]
61. Suzuki M, Matsuo Y. A survey of multimodal deep generative models. Adv Robot. 2022;36(5-6):261-278. [DOI]
62. Kutuzova S, Krause O, McCloskey D, Nielsen M, Igel C. Multimodal variational autoencoders for semi-supervised learning: In defense of product-of-experts. arXiv:2101.07240 [Preprint]. 2021. [DOI]
63. Geenjaar E, Lewis N, Fu Z, Venkatdas R, Plis S, Calhoun V. Fusing multimodal neuroimaging data with a variational autoencoder. In: 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC); 2021 Nov 1-5; Mexico. New York: IEEE; 2021. p. 3630-3633. [DOI]
64. Da Silva-Filarder M, Ancora A, Filippone M, Michiardi P. Multimodal variational autoencoders for sensor fusion and cross generation. In: 2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA); 2021 Dec 13-16; Pasadena, USA. New York: IEEE; 2021. p. 1069-1076. [DOI]
65. Xu Y, Liu X, Pan L, Mao X, Liang H, Wang H. Explainable dynamic multimodal variational autoencoder for the prediction of patients with suspected central precocious puberty. IEEE J Biomed Health Inform. 2022;26(3):1362-1373. [DOI]
66. Mathieu E, Rainforth T, Siddharth N. Disentangling disentanglement in variational autoencoders. arXiv:1812.02833 [Preprint]. 2019. [DOI]
67. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput. 1997;9(8):1735-1780. [DOI]
68. Ma DS, Correll J, Wittenbrink B. The Chicago face database: A free stimulus set of faces and norming data. Behav Res. 2015;47:1122-1135. [DOI]
69. Phillips PJ, Wechsler H, Huang J. The FERET database and evaluation procedure for face-recognition algorithms. Image Vis Comput. 1998;16(5):295-306. [DOI]
70. Wallhoff F, Schuller B, Hawellek M, Rigoll G. Efficient recognition of authentic dynamic facial expressions on the feedtum database. In: 2006 IEEE International Conference on Multimedia and Expo; 2006 Jul 9-12; Toronto, Canada. New York: IEEE; 2006. p. 493-496. [DOI]
71. Milborrow S, Morkel J, Nicolls F. The MUCT Landmarked Face Database [Internet]. South Africa: University of Cape Town; 2010. Available from: http://www.milbo.org/muct/
72. Xu J. Person Face Dataset - This Person Does Not Exist [dataset]. Kaggle; 2021. Available from: https://www.kaggle.com/datasets/almightyj/person-face-dataset-thispersondoesnotexist
73. Burgess CP, Higgins I, Pal A, Matthey L, Watters N, Desjardins G, et al. Understanding disentangling in β-VAE. arXiv:1804.03599 [Preprint]. 2018. [DOI]
74. Koelstra S, Muhl C, Soleymani M, Lee JS, Yazdani A, Ebrahimi T. Deap: A database for emotion analysis; using physiological signals. IEEE Trans Affective Comput. 2012;3(1):18-31. [DOI]
75. Soleymani M, Lichtenauer J, Pun T, Pantic M. A multimodal database for affect recognition and implicit tagging. IEEE Trans Affective Comput. 2011;3(1):42-55. [DOI]
76. Russell JA. A circumplex model of affect. J Pers Soc Psychol. 1980;39(6):1161-1178. [DOI]

Publisher's Note