frontiers | Frontiers in Neurorobotics

TYPE Methods
PUBLISHED 11 September 2024
DOI 10.3389/fnbot.2024.1442080

OPEN ACCESS

EDITED BY
Xin Luo,
Chinese Academy of Sciences (CAS), China

REVIEWED BY
Lun Hu,
Chinese Academy of Sciences (CAS), China
Jiahui Pan,
South China Normal University, China
Zhao Ren,
University of Bremen, Germany
Weiyi Yang,
University of Chinese Academy of Sciences, China

*CORRESPONDENCE
Xia Ye
✉ yex133@outlook.com

RECEIVED 01 June 2024
ACCEPTED 26 July 2024
PUBLISHED 11 September 2024

CITATION
Du Z, Ye X and Zhao P (2024) A novel signal channel attention network for multi-modal emotion recognition.
Front. Neurorobot. 18:1442080.
doi: 10.3389/fnbot.2024.1442080

A novel signal channel attention network for multi-modal emotion recognition

Ziang Du, Xia Ye* and Pujie Zhao

Xi'an Research Institute of High-Tech, Xi'an, Shaanxi, China

Physiological signal recognition is crucial in emotion recognition, and recent advancements in multi-modal fusion have enabled the integration of various physiological signals for improved recognition tasks. However, current models for emotion recognition with hyper complex multi-modal signals face limitations due to fusion methods and insufficient attention mechanisms, preventing further enhancement in classification performance. To address these challenges, we propose a new model framework named Signal Channel Attention Network (SCA-Net), which comprises three main components: an encoder, an attention fusion module, and a decoder. In the attention fusion module, we developed five types of attention mechanisms inspired by existing research and performed comparative experiments using the public dataset MAHNOB-HCI. All of these experiments demonstrate the effectiveness of the attention module we addressed for our baseline model in improving both accuracy and F1 score metrics. We also conducted ablation experiments within the most effective attention fusion module to verify the benefits of multi-modal fusion. Additionally, we adjusted the training process for different attention fusion modules by employing varying early stopping parameters to prevent model overfitting.

KEYWORDS

hypercomplex neural networks, physiological signals, attention fusion module, multi-modal fusion, emotion recognition

1 Introduction

Multi-modal signal recognition is a critical area of research within multi-modal fusion, encompassing fields such as speech signal recognition, physiological signal recognition, and radar signal recognition. The primary task is to classify multi-modal signals from the same individual, as individual signals alone often fail to capture the comprehensive features required for study. Thus, developing multi-modal signal classification models is crucial for a deeper understanding of these signals.

Using physiological signals for emotion recognition is a significant approach to studying human emotions. Since expressions and speech can conceal emotions and behavioral responses can suppress abnormal emotions, the advent of non-invasive and affordable wearable devices has propelled deep learning-based physiological emotion recognition into a prominent research area. Common physiological signals used include electroencephalogram (EEG), electrocardiogram (ECG), galvanic skin response (GSR), and eye data, etc. In the past, emotion recognition tasks frequently relied on data involving facial expressions or speech signals. Unlike these outward expressions, EEG signals offer a direct window into the brain's physiological activity, making them less prone to artifacts or manipulation and thus providing a more authentic and unbiased view of one's emotional state. ECG signals, on the other hand, directly reflect the heart's activity, which is closely tied to emotions, and are adept at capturing physiological responses to emotional shifts.

ECG typically boasts a higher signal-to-noise ratio than EEG and is more easily obtained and analyzed non-invasively. GSR is particularly sensitive to an individual's emotional arousal, such as nervousness or excitement, often marked by changes in skin conductance. Eye movement signals are closely linked to attention and interest, with emotional states deduced from the duration and direction of gaze. While each of these modalities offers unique benefits for emotion recognition, there are relatively few studies that have harnessed all four signals in tandem for this purpose. These signals serve as inputs for deep learning models in emotion classification.

With the rise of deep learning, Advanced neural networks with attention modules have also proliferated in recent years. Kalman filters can be combined with residual neural networks to obtain even better neural networks (Yang et al., 2023a). A novel deep saliency-aware bi-embedded attention network (SAD-Net) for non-periodic multivariate time series prediction and has demonstrated high performance on correlated datasets (Li et al., 2023). Multi-layer fully connected networks and lightweight graph-convolutional networks can be fused into a dual-stream graph-convolutional network fusing potential features, who can solve the key problem of linear properties and the limitations of implicitly encoded cooperative QoS signal (Bi et al., 2023). Concurrently, numerous physiological signal classification models have been developed annually. Among these, multi-modal hyper complex neural network models show great potential and have achieved notable success in physiological signal classification. These models utilize multi-modal physiological signals from various emotional states, surpassing previous methods focused solely on EEG signals. Past research has predominantly remained single-modal and lacked the capability to extract comprehensive data features using deep learning models. To address this challenge, the authors proposed a hyper complex neural network model (Lopez et al., 2023).

In current research on multi-modal neural networks, the varying significance of different information types in determining data features remains unresolved, limiting the training potential of these models. Proper deep learning neural network models can adjust the weights of each modality to optimal values, and incorporating attention mechanisms can enhance model performance. However, due to inadequate attention mechanisms, hyper complex neural networks have not yet achieved their expected performance. Therefore, increasing attention mechanisms will help neural networks better capture key information, thereby improving model effectiveness.

To this end, we have enhanced the existing multi-modal signal classification model, specifically the multi-modal hypercomplex neural network model, by incorporating five newly designed attention modules. This has led to the development of five improved models that outperform previous hypercomplex multi-modal signal models in the realm of physiological signal research. Crucially, we preserved the original model's innovative aspects. Through multiple model comparison experiments and ablation studies using the publicly available benchmark dataset MAHNOB-HCI, we verified the efficacy of our approach. Our model demonstrated significant performance improvements. Figure 1 summarizes the tasks our designed network will undertake, with the three images illustrating the different classification samples corresponding to our final three classification tasks. Additionally, this study identified shortcomings in the data processing and conclusions of previous research, which we will address in detail in the paper.

Our contributions can be summarized as follows:

- We have designed a novel framework based on previous research, named SCA-Net, which achieves higher accuracy and better prediction balance in the field of multi-modal physiological signal recognition.
- We discovered that improvements in multi-modal physiological signal processing models can be attained through two types of attention methods: channel attention and self-attention. Incorporating both types of attention resulted in varying degrees of enhancement in the model's performance.
- We conducted comparative experiments on five models using the publicly available dataset MAHNOB-HCI, both with and without data augmentation. Additionally, we performed ablation experiments on the best-performing models to validate the effectiveness of our approach.

2 Related work

In recent years, the advantages of deep learning have become increasingly evident, sparking a renewed enthusiasm for emotion recognition research. Extensive studies have already been conducted across various domains, including natural language processing, computer vision, and signal processing. Broadly speaking, these studies fall into two categories: emotion recognition based on a single modality and emotion recognition based on multiple modalities.

2.1 Emotional recognition under single modality

In the past many studies, and there have been many methods that can be used for emotion recognition under a single mode (Rayatdoost and Soleymani, 2018; Wang et al., 2019; Du et al., 2020; Maeng et al., 2020), but when these neural network frameworks carry out emotion recognition tasks, they excessively rely on extracted features, such as power spectral density (PSD) and differential entropy (DE), which ignore the ability of neural networks to recognize and extract features. Among the numerous available data, facial images can be used as an auxiliary means to preliminarily recognize emotions (Tao et al., 2020), and the speech signals received by sensors can also be used as a means of emotion recognition (Sajjad et al., 2020; Shen et al., 2024). ECG signals also have obvious features for emotion recognition when humans listen to music (Hsu et al., 2017), and later there were also many efficient single-mode networks designed for this type of data (Lv et al., 2022; Ye et al., 2023; Ju et al., 2024; Wang et al., 2024). The 3D representation of EEG signals is used for learning 3D convolutional neural networks (Salama et al., 2018), thermal imaging of facial expressions is used for emotion recognition research (Gupta and Sengupta, 2023), and so on. Continuously studying emotion recognition methods from new perspectives is

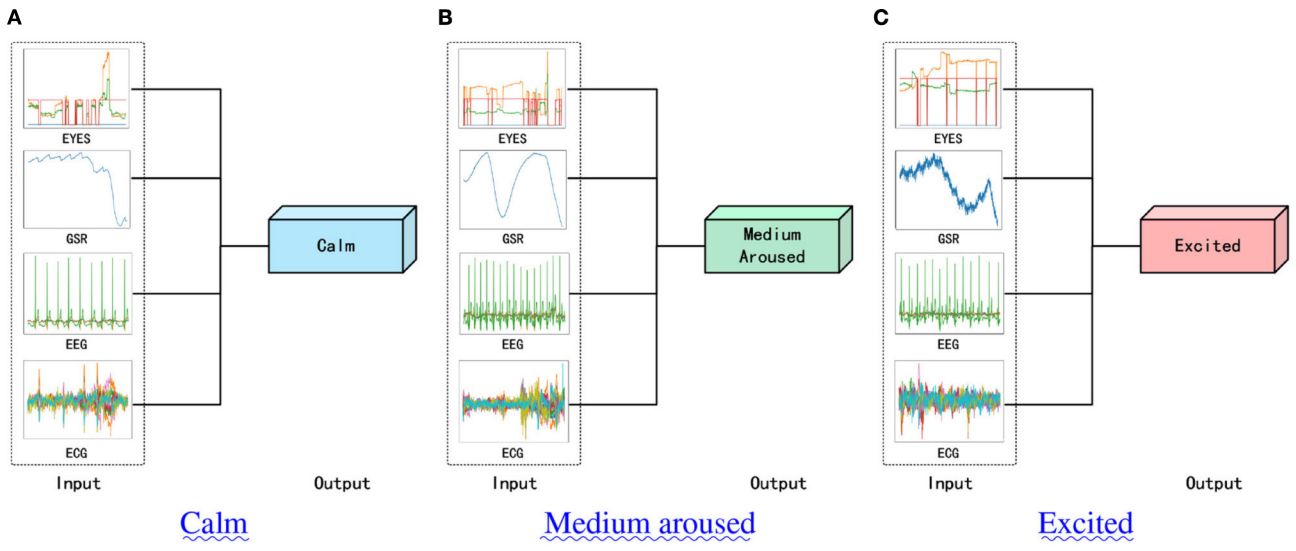

FIGURE 1
The summary of the model we designed. **(A)** shows a sample of data predicted as "Clam," **(B)** shows a sample of data predicted as "Medium surrounded," and **(C)** shows a sample of data predicted as "Excited."

worthy of recognition. It fully utilizes neural networks to learn data features, and related research has also achieved certain results. In a word, all the above researches are based on the single mode method.

2.2 Emotional recognition in multi-modal settings

In order to fully utilize the features in the data and enhance the interactivity between data features, multi-modal fusion between data is particularly important. Recently, a large number of studies have used multi-modal fusion methods, some of which rely on the features extracted by neural networks (Rayatdoost et al., 2020; Tan et al., 2020; Zhang et al., 2022), while others have undergone data transformation without using the original data (Nakisa et al., 2020; ZENG et al., 2020; Dolmans et al., 2021). At the same time, there have been many advances in the study of modal interactivity, such as the interaction between acoustic information and natural language (Sakurai and Kosaka, 2021), modal interaction between audio and video (Chang and Skarbek, 2021), and the emotional recognition of gestures and facial expressions (Avula et al., 2022). In recent studies, features of human facial data, speech data, and EEG signals have been fused at the decision level through three branches, and very significant breakthroughs have been made in utilizing such data for multi-modal emotion recognition tasks (Pan et al., 2023). Faced with the difficulty of using physiological signals to solve research gap of emotional recognition, Parameterized hypercomplex neural networks (PHNN) is proposed as an emerging family of models which operate in a hypercomplex number domain (Zhang et al., 2021; Grassucci et al., 2022). In the study of using hypercomplex theory to solve multi-modal signal emotion recognition, the focus of data feature learning is on the parameterized hypercomplex multiplication (PHM) layer in the latter half of the network (Lopez et al., 2023).

One of the key steps to improve the effectiveness of multi-modal learning is the attention network, which can simulate human behavior to classify the information we obtain, filter secondary information, and grasp the main information. The use of attention can effectively enable the model to grasp the key parts of the many features in the data (Chen et al., 2020). Secondly, fusion strategy is also a key step that affects multi-modal networks. Early fusion methods did not consider the properties of different modalities, which can easily overlook the complementary information between modalities. Later fusion methods can easily cause network complexity, and more importantly, they cannot fully utilize cross modal information (Kaliciak et al., 2014; Gadzicki et al., 2020). A fusion strategy that combines the advantages of the above two methods, namely hybrid fusion, also known as intermediate fusion, is relatively complex and requires full consideration of various complexity issues (Stahlschmidt et al., 2022). In this paper, we propose a new network model based on the theory of hyper complex multi-modal emotion recognition networks. This network model can not only grasp the weight relationship of multiple modalities through attention, but also interact the fused modal information with the PHM layer in the decoder to obtain better modal interaction information.

3 Method

3.1 Framework overview

In Figure 2, we present a streamlined and efficient architecture for emotion recognition of multi-modal signals, utilizing a novel fusion approach and attention mechanisms. This architecture, termed the multi-modal hyper complex fusion network, optimizes the integration of modal information from various physiological signals to enhance sentiment classification accuracy. The framework comprises three main components: (i) a data encoder,

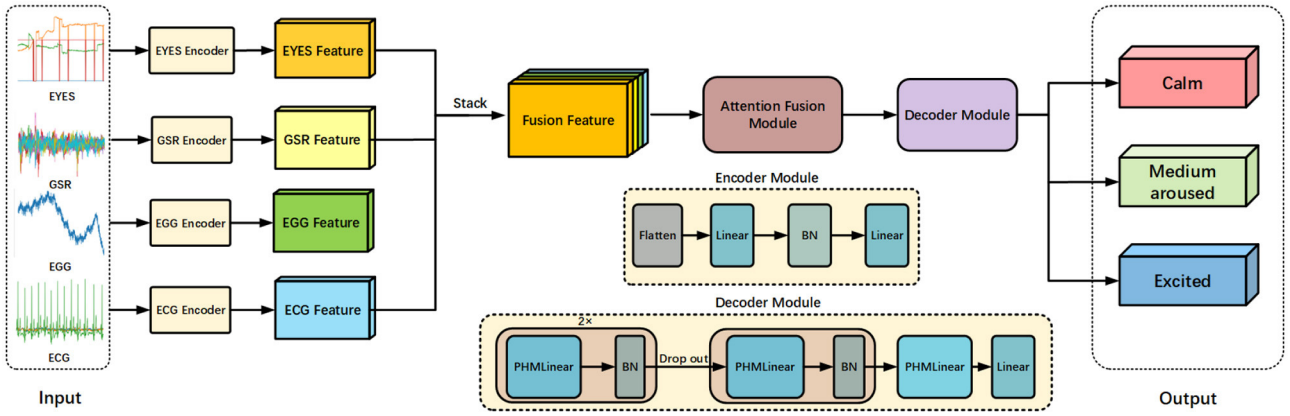

FIGURE 2
Overall framework of our proposed approach. We extract multi-modal features of signals through the following methods. The signals of four modes are used as inputs and converted into features for fusion by their respective encoders. The architecture of the encoders includes a linear layer and a batch normalization layer. The main features is emphasized in our designed attention fusion module. The fused features of the final signal are refined by a decoder containing a PHM layer, and ultimately mapped to the output.

which extracts features from the modal data of each signal while converting its represented feature dimensions into consistent ones through linear layers. (ii) Attention fusion module, which aims to multiply the fused multi-modal data with the weight matrix. Through continuous training of the model, the final model will focus on the main features when processing information. (iii) Feature decoder, which decodes the output from the attention fusion module and further transforming dimensions through the PHM layer and linear layer for comprehensive sentiment recognition of multi-modal signal data.

Firstly, we extract features from the data of four physiological signals using an encoder, comprising linear mapping and normalization operations.Among these, the EEG, ECG, and EYE encoders traverse the data through the linear layer and batch normalization layer within the module twice, whereas GSR only requires a single pass. This encoder standardizes the feature dimensions of all four modalities to 512, followed by feature-level fusion to create features of $8 \times 4 \times 512$ dimensions. Secondly, we employ five attention modules to focus on the fused main features across the four channels, assigning weights to modal features from different channels. This module is collectively referred to as the attention fusion module within the model. Finally, we utilize the PHM layer in the feature decoder to capture both local and global information from the fused features, and a linear layer to map the weighted features into dimensions, culminating in the downstream task of sentiment recognition.

3.2 Input and output modeling

Our model begins with four distinct dimensions of signal modal data, all comprising waveform data. In the preprocessing stage prior to model input, we adopted processing methods from prior studies on this dataset. Specifically, for EEG, ECG, and GSR signals, we conducted downsampling operations, reducing their frequencies from 256Hz to 128Hz. Conversely, for eye data, we retained its original frequency and applied relevant filters. Thus, the dimensions of the input data are as follows: EYES [8,600,4], GSR [8, 1,280], EEG [8, 1,280, 10], and ECG [8, 1280, 3]. By utilizing encoders tailored to each of the four signal modes, we standardized the signals into unified [8, 512] dimensional features, ensuring suitable data input for the attention fusion module.

Moreover, we preserved the linear mapping aspect of the PHM layer from the original hyper-complex multi-modal fusion model to apply to the fused features. This ensures that our model's decoder maintains the original mapping relationship with the output of the attention fusion module.

3.3 Early interaction

In previous research on hyper-complex multi-modal fusion networks, these four modal data types were fused via batch alignment. However, only the PHM layer could map the feature vectors of the four modalities in the fused data, which did not effectively capture the impact of each modality on final emotion recognition. To address this, we employed modal fusion by stacking. The fused modal features were then processed through our designed attention module, which includes five types of attention based on the fused features. We conducted experiments using these attention mechanisms, which were developed from channel attention, self-attention, and enhanced modal frameworks.

With our proposed fusion method and attention model, each modality can be assigned its respective weight before entering the PHM layer. Finally, guided by the decoder, emotion recognition is performed, establishing a novel multi-modal emotion recognition network. Additionally, we adjusted the early stopping parameter based on the model's adaptation to various attention levels to mitigate potential overfitting issues.

3.4 Attention fusion module

The attention fusion module we designed encompasses five types of attention tailored for multi-modal physiological signal research. Each type of attention network is individually used

as our attention fusion module to realize its own function in the network. Drawing inspiration from channel attention, self-attention, and their variant frameworks, our aim is to enhance the training performance of hyper-complex multi-modal neural network models. The following is an in-depth explanation of our proposed Species Attention Fusion Module, including the framework for modeling and the underlying principles of the formulas.

3.4.1 Signal channel squeeze and excitation attention

We drew inspiration from the block design in SE net (Hu et al., 2018) and identified that extracting channel features for modal weighting could enhance training performance. Consequently, we devised the Signal Channel Squeeze and Excitation Attention (SCSEA) module, comprising three pivotal steps. The module's framework diagram is illustrated in Figure 3. The initial step maximizes input feature pooling to derive channel features, followed by channel feature extraction via linear layers in the second step. Finally, the third step involves channel feature extraction from functions processed by activation functions.

Assuming x is the input signal fusion feature. $x \in R^{B*C*F}$. Where B represents batch size, C represents channel, and F represents the characteristics of the channel.After performing global average pooling, the features of x are preliminarily extracted now $x \in R^{1*1*C*1}$. Before that, there was actually a dimension extension operation that changed $x_0 \in R^{B*C*F*1}$ to $x_1 \in R^{1*1*C*1}$.Then, the feature is further transformed into $x_2 \in R^{1*1*C/r*1}$ with the aim of having the activation function act on it and then restore it to the same dimension as x_1 through a linear layer. We assume it as x_3, and the entire process of this attention can be expressed by the following formula:

$$x_1 = GAP(x_0) \tag{1}$$

$$y = f(\sigma(f(x_1))) * x_0 \tag{2}$$

where GAP is the global average pooling operation, σ is the activation function, and f is the linear mapping function.

Using the data from our research as an example, the fusion features of the four signals need to undergo dimension expansion from [8, 4, 512] to [8, 4,512, 1] before entering the attention module. This expansion aims to allow the attention module to adjust the weight of the channel dimension. Initially, the input data is represented as channel features via a max-pooling layer, altering the data dimension to [1, 1, 4, 1]. Here, each modality's data is compressed into a single dimension, making its scattered features more accessible. Next, a linear layer squeezing operation extracts the four most prominent channel features, followed by activation with a ReLU function for nonlinear mapping. Subsequently, the dimension of the weight matrix obtained from this extraction changes to [1, 1, 4, 1], and it multiplies with the original input fusion features to fulfill the attention module's final objective. Lastly, a summation in the channel dimension is performed to assist the encoder in decoding the features.

3.4.2 Efficient signal channel attention

Inspired by EC Attention (Wang et al., 2020), our approach for signal fusion features calculates the attention matrix from a different perspective, focusing on the convolution angle. For sequential data like signals, an appropriate convolutional kernel proves more effective. Building upon the methods and principles of EC Attention, we've devised an attention module tailored for our fusion signal, termed Efficient Signal Channel Attention (ESCA). Refer to Figure 4 for the module's framework diagram.

Given our input feature x_0, similar to the previous SCSEA, we perform an average pooling operation to preliminarily extract channel features, resulting in x_1. Then, through a linear mapping, it is transformed into a dimension that can interact with the convolution kernel, and then subjected to one-dimensional convolution. We have designed a convolution layer for x_1, called AC Layer, which includes using appropriate convolution kernels for convolution operations. After performing one-dimensional convolution, we obtained result x_2 and obtained the weight matrix of attention through non-linear mapping using an activation function. By multiplying it with the original input features, we obtained the purpose of our attention. This process can be expressed as a function:

$$y = \sigma(f_{AC}(GAP(x_0))) * x_0 \tag{3}$$

where GAP is the global average pooling operation, Σ is the activation function, and f_{AC} is the adaptive convolution function.

3.4.3 Signal feature dot product block

Both of the approaches mentioned above involve attention calculation at the channel level. Building on the principles of SDP Attention (Vaswani et al., 2017), we've revamped our strategy by integrating Q, K, and V as input matrices for the attention function, dubbing it Signal Feature Dot Product Block (SFDPB).

Refer to Figure 5 for the module's framework diagram. This shift stems from our team's recognition of the benefits of self-attention, particularly in capturing global features and showcasing exceptional adaptability. Given the sequential nature of signal data, besides global features, capturing its inherent characteristics poses a significant challenge. Moreover, the psychological signal data we examine also displays substantial volatility, necessitating attention functions with robust adaptability. The calculation method of this attention function is:

$$f(Q, K, V) = V * softmax(\frac{QK^T}{\sqrt{d_k}}) \tag{4}$$

where Q is the query matrix, K is the key matrix, and V is the value matrix. d_k is the dimension of the key matrix. We have maintained the same approach as the original author in calculating this attention function, as this method has been proven to be more effective. Even with the use of previous methods, we have fine tuned the internal attention of the signal fusion feature format in this study to make it effective in the network we are studying. For example, we set the d_v, d_k, and h parameters during the calculation process to 64, 64, and 8, respectively, and we add and sum the dim=1 dimension of the attention calculation result to facilitate addition at the decoder level.

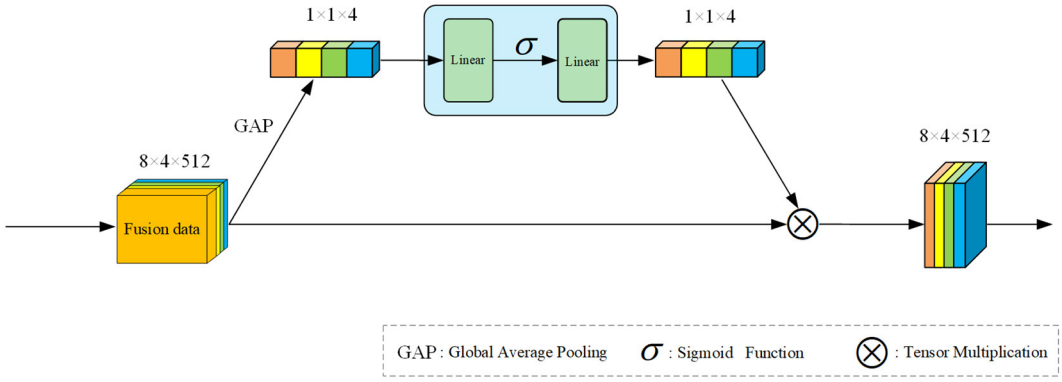

FIGURE 3
Overall framework of our proposed SCSEA. The input feature data is channel-level extracted through Global Average Pooling (GAP). Squeezing and extraction are performed in two separate linear layers, with non-linear activation applied using an activation function. The resulting attention weight matrix is then multiplied with the original feature data.

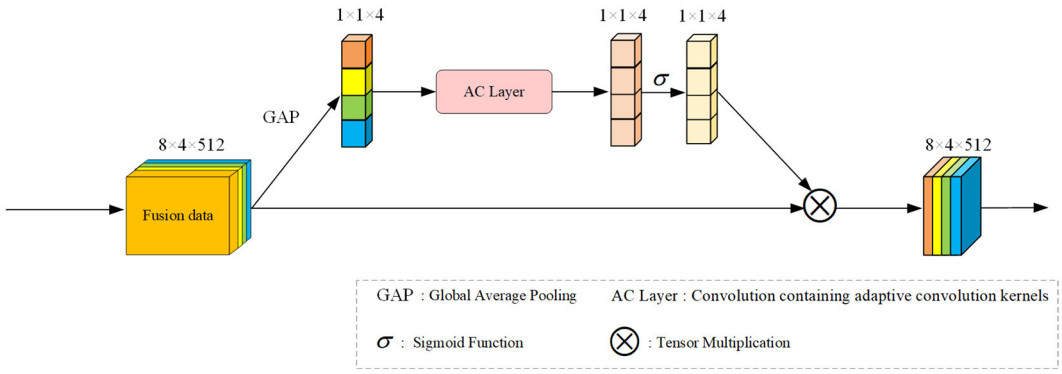

FIGURE 4
Overall framework of our proposed ESCA. We'll extract channel features from the input feature data using the GAP method. This method conducts convolutional extraction of local features within the channel in our designed AC Layer, resulting in a weight matrix under the activation function's influence. This tensor is then multiplied with the original input to generate our output.

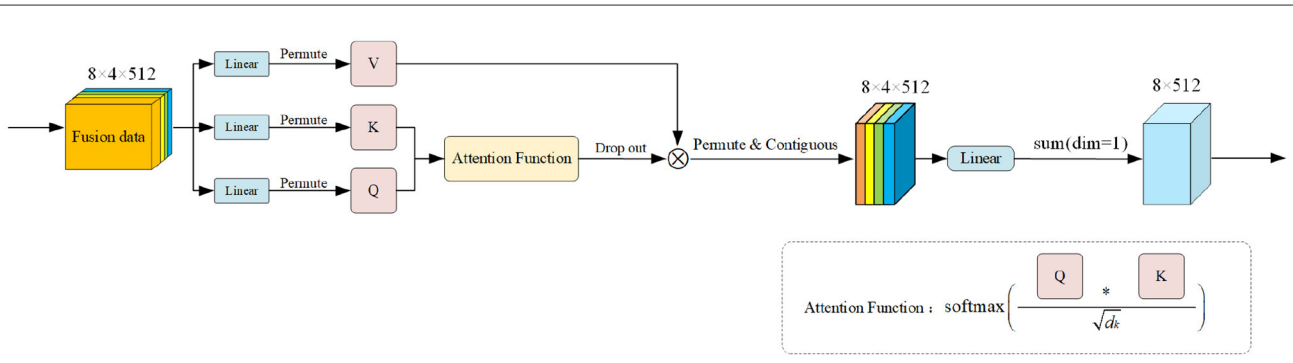

FIGURE 5
Overall framework of our proposed SFDPB.The input feature data undergoes linear mapping across three designated linear layers to derive *Q*, *K*, and *V* matrices. *Q* and *K* undergo operations and dropout within specific attention functions to bolster feature representation. The resultant matrix is then multiplied by the *V* matrix, and the data is permuted and concatenated to restore the feature dimension. Lastly, we sum up the dimensions with $dim = 1$ to yield the output result.

3.4.4 Signal attention free block

Following the previous calculation approach, the matrices *Q*, *K*, and *V* are derived through linear transformation of the original data features. In contrast to SDP attention, AFT attention employs a different attention function calculation method (Zhai et al., 2021). Inspired by the structure of AFT attention, we

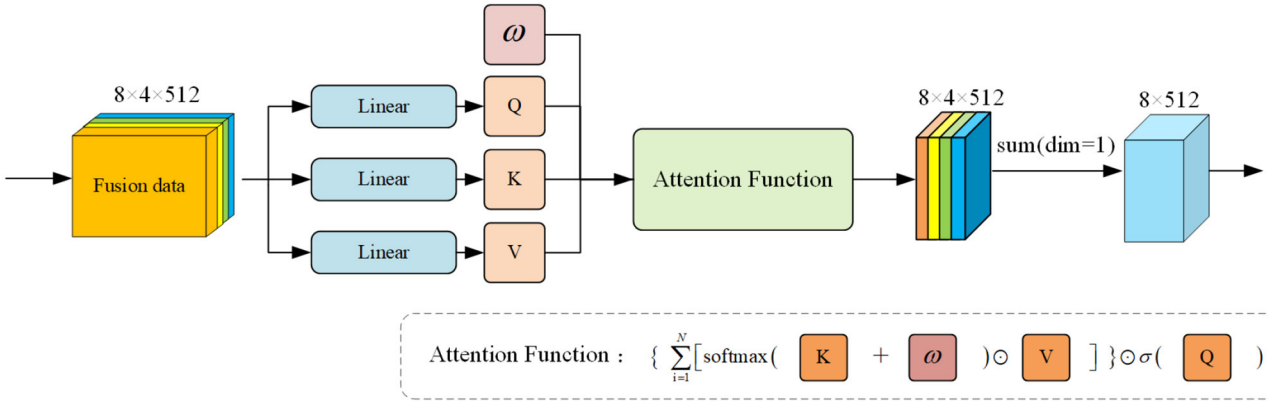

FIGURE 6
Overall framework of our proposed SAFB. The input feature data is mapped linearly across three designated linear layers to derive *Q*, *K*, and *V* matrices. Initially, a weight is initialized and operated on with *Q*, *K*, and *V* within specific attention functions to obtain preliminary weighted features. To aid the decoder's final linear mapping, we then sum the dimensions with $dim = 1$ to produce the output result.

devised our Signal Attention Free Block (SAFB), as depicted in Figure 6.

For data signals, each feature within a single channel undergoes a weighted average of AFT execution values, which are then combined with element-wise multiplication queries. This attention calculation method simplifies the weight computation, relying solely on a key and a set of learned paired positional deviations, offering the direct advantage of avoiding the need to compute and store costly attention matrices. We also drew on its advantages to produce this attention block that resonates with the model we are studying. The calculation of this attention function is as follows:

$$y_{output} = \sigma(Q) \odot \sum_{i=1}^{N} [softmax(Q + w) \odot V] \tag{5}$$

where Q is the query matrix, K is the key matrix, and V is the value matrix. d_k is the dimension of the key matrix. $\odot$ represents the product of two elements before and after and w is the weight matrix generated by initialization.

3.4.5 Multiscale signal transformer block

When we don't focus on processing signal sequence data, attention in the field of computer vision will also achieve good results. MVITv2attention is an example of using matrices Q, K, and V (Li et al., 2022). The way we obtain these three matrices in our designed attention remains consistent with the previous text, but it reduces pooling operations compared to MVITv2attention, so we will not elaborate on it here. We found that MVITv2 attention changed the calculation method of the Q, K, and V matrices. Our research retained some of its advantages in the module and designed a new attention, which is Multiscale Signal Transformer Block (MSTB). Its framework diagram is shown in Figure 7.

The expression for this attention is as follows:

$$y_c = K \odot softmax(I) \tag{6}$$

$$y_{output} = f_{dsum}(f_{dsum}(y_c) \odot V) \tag{7}$$

where Q is the query matrix, K is the key matrix, and V is the value matrix. y_c represents context score, which is an intermediate variable for us to calculate the weight matrix. $\odot$ represents the product of two elements before and after and w is the weight matrix generated by initialization. y_c is a function that sums up a specific dimension. In this module, we default to summing up in the $sum = 1$ dimension.The main function of MSTB is to capture the contextual features of each modality, sum them up in specific dimensions, and reduce the space occupation of features, which is more conducive to the decoder's feature classification.

4 Experiments

4.1 Dataset and evaluation metrics

4.1.1 Dataset

All experiments in this study were conducted using the publicly available MAHNOB-HCI dataset, consistent with the original research on hypercomplex multi-modal emotion recognition networks. This dataset encompasses diverse physiological response data alongside subjective emotional reports from participants. Specifically, it serves as a multi-modal resource for emotion recognition, comprising synchronized recordings of facial videos, audio signals, eye gaze data, and peripheral/central nervous system physiological signals from 27 participants viewing emotional video clips. Eye gaze data includes attributes like eye distance, pupil size, and gaze coordinates. For our physiological signal recognition study, we focused solely on EEG, ECG, and GSR due to their strong correlation with human emotions. Additionally, the dataset provides relevant labels including calm, moderate arousal, excitement, and valence categories (unpleasant, neutral, and pleasant). However, within our multi-modal model, our emphasis was on validating method effectiveness, thus concentrating on emotion recognition for three specific labels: Calm, Medium Aroused, and Excited.The whole dataset was processed through data processing on the basis of the original MAHNOB-HCI dataset, 80% of the data was constructed as a training set, and 20% of the data was constructed as a test set. The divided training and test sets are then put into the model for training.

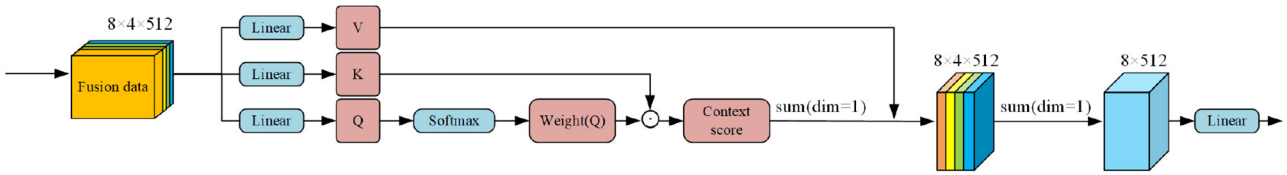

FIGURE 7
Overall framework of our proposed MSTB. The input feature data undergoes linear mapping across three specific linear layers to derive *Q*, *K*, and *V* matrices. The *Q* matrix requires non-linear mapping through an activation function to obtain its corresponding weight matrix. This weight matrix is then multiplied with the *K* matrix to obtain the context score matrix. Summation along $dim = 1$ is performed to reduce feature space occupation, followed by multiplication with the *V* matrix to yield preliminary results. Similarly, akin to the previous module, dimension summation with $dim = 1$ occurs, followed by linear layer mapping to obtain the final result.

4.1.2 Evaluation

We utilize accuracy and F1 score as evaluation metrics for our model. Accuracy indicates the percentage of correct predictions made by the model across the entire sample, while the F1 score, being the harmonic mean of accuracy and recall, offers a balanced assessment of model prediction performance. The F1 score can be calculated using the following formula:

$$Pr = \frac{TP}{TP + FP} \tag{8}$$

$$Re = \frac{TP}{TP + FN} \tag{9}$$

$$F1 = 2 * \frac{Pr * Re}{Pr + Re} \tag{10}$$

Where *TP* (True Positive) is the number of true positive classes that were correctly predicted, *FP* (False Positive) is the false positive class prediction, and *FN* (False Negative) is the false negative class prediction. *Pr* and *Re* represent precision and recall.

These metrics together provide a comprehensive evaluation of the model quality. Moreover, to address overfitting concerns, various early stopping parameters were employed for different attention modules in our study. Ablation experiments were conducted using the attention module that exhibited the best classification performance.

4.2 Implementation details

4.2.1 Architecture

The input signal comprises four modes: EYES with dimensions [8, 600, 4], GSR with dimensions [8, 1,280], EEG with dimensions [8, 1,280, 10], and ECG with dimensions [8, 1,280, 3]. Utilizing encoders corresponding to each signal mode, we convert these signals into unified features of dimensions [8, 512]. In the feature fusion module, we merge features of the same dimension at the feature level before inputting them into the attention fusion module. Our study employs five distinct types of attention fusion modules, each with its unique structure. The decoder consists of multiple $n = 4$ PHM layers and intersecting normalization layers. In each PHM layer and normalization layer within the decoder, the feature dimension is halved, culminating in the final prediction output.

TABLE 1 **The table illustrates the performance of each attention module with augment data, with model evaluation metrics categorized into accuracy and F1 score.**

Attnetion module	Early stopping	Accuracy	F1 score	Parameters
Baseline	20	34.4482	0.3196	19663747
ESCA	20	42.4749	0.3251	18506051
SFDPB	20	40.1338	0.1909	17913091
SAFB	8	38.4615	0.2308	17913091
MSTB	20	40.1338	0.1909	17913091
SCSEA	8	**43.1438**	**0.3328**	17847451

"Early stopping" represents the number of training iterations needed for the model when the training loss shows signs of increase. "Parameters" represents the number of parameters included in the model. The bold values means that the corresponding model has the best performance for the corresponding metrics.

4.2.2 Training

In our model, we employed the Adam optimizer with a fixed learning rate of 0.000000796 and zero weight decay. Training was conducted on a single Nvidia RTX4090 GPU, utilizing a total batch size of 8, with all networks operating on this GPU. Optimization was achieved using CrossEntropy Loss, aligning predicted physiological signal labels with actual emotional categories.

4.3 Comparison with previous works

From the results of this experiment, we observed variations in replicating the original hypercomplex neural network model's results. Further analysis unveiled missing and erroneous data within our study. Consequently, we purged these data points, obtaining a clean dataset suitable for training, albeit influencing our final model predictions. Notably, the original paper didn't address this issue or offer solutions based on their findings. We reproduced the initial hypercomplex neural network model (Lopez et al., 2023), which is named "Baseline" in our experiment. The most notable differences from our study are the modal fusion approach and the lack of an attentional module. Our experiments encompassed both unaugment and augment datasets, with the comparative results detailed in Table 1.

Utilizing five types of attention, the highly intricate multi-modal physiological signal sentiment classification model

TABLE 2 Results table of modal ablation experiment.

	Modal				Results	
	EYES	GSR	ECG	EGG	Accuracy	F1 score
Exp1	-	+	-	-	37.4581	0.3299
Exp2	-	+	+	-	33.1104	0.3206
Exp3	-	+	+	+	41.1371	0.2972
Exp4	+	+	+	+	**43.1438**	**0.3328**

The "+" represents the use of the corresponding modality, while the "-" corresponds to the removal of the corresponding modality. The bold values means that the corresponding model has the best performance for the corresponding metrics.

demonstrated notable enhancements, particularly with SCSEA and ESCA, resulting in substantial overall improvements in accuracy and F1 score. Additionally, the five attention modules we developed each contributed to a reduction in the model's parameter size to varying extents, with the SCSEA module demonstrating the most substantial decrease. Furthermore, as indicated in Table 1, the attention modules SFDPB, SAFB, and MSTB exerted a comparable impact on the model's parameter size, a result of their similar design principles. Although they differ slightly in computation methodology, the specifics can be referenced in the module framework diagram provided above. While the remaining three attention types exhibited accuracy improvements, their F1 scores notably declined. The reason for this phenomenon should be the following two reasons:

1) When dealing with a modest amount of data, certain discrepancies become more pronounced. This is particularly true for the F1 score calculation, which includes inverse operations, making the disparities in predictive balance particularly evident.
2) The F1 score serves as a holistic measure of both precision and recall. Incorporating the three sub-attention modules leads to a reduction in recall, consequently boosting the accuracy, which inversely affects the F1 score, causing it to decline.

Consequently, integrating these three attention types enhanced the model's prediction accuracy, albeit resulting in a more unbalanced prediction compared to the other two attention types.

4.4 Ablation study

From the preceding analysis, it's evident that SCSEA experienced the most significant improvement post-model addition. We conducted ablation experiments on multi-modal physiological signals, comparing single-mode (GSR signal), dual-mode (GSR and ECG signals), three-mode (GSR, EEG, and ECG signals), and four-mode (EYES, GSR, EEG, and ECG signals) scenarios. The Table 2 showcases the experimental results under consistent experimental conditions and parameters.

The results of the ablation experiment clearly indicate the necessity of modal synergy. Despite poorer dual-modals performance in the "exp2" experiment, the overall trend highlights that leveraging interaction among the four modalities is the most effective approach for improving emotion signal recognition. Compared to initial single-mode results, interaction among the four physiological signals yielded superior accuracy and F1 scores. Boosted by our optimal attention module from "exp4," accuracy improved by approximately 6% compared to single-mode experiments, while the F1 score rose by 0.004. However, significant enhancement of the F1 score remains limited by data diversity. Despite efforts to enrich existing data, this limitation persists.

4.5 Discussion

Our proposed SCA-Net offers three key advantages:

1) It integrates the features of the four modalities-EEG, ECG, GSR, and EYE-along the channel dimension, enabling the attention fusion module to assign weights directly to each modality. This allows the primary modality to take precedence while secondary modalities contribute differently to the final classification task in emotion recognition.
2) We have validated multiple attention methods and demonstrated that channel attention is more appropriate for our network design than self-attention, as evidenced by its superior accuracy, F1 score, and parameter count. Specifically, the SCSEA module we developed compresses and extracts features from individual channels, providing a more direct response to the core characteristics of each modality. As a result, SCSEA outperforms all other attention modules.
3) For the output of the attention fusion module, we have adapted its overall dimensionality to align with the features of the PHM layer, which in turn facilitates the PHM's decoding and pattern recognition tasks on the weighted data features.

These three complementary advantages collectively enhance the performance of SCA-Net, surpassing that of the original hypercomplex multi-modal neural network model in the realm of emotion recognition using multichannel physiological signals.

5 Conclusion

In this study, we introduce five novel attention-based hypercomplex models for sentiment recognition of physiological signals. After conducting our experimental research, we've found that SCA-net, as a multi-modal neural network, exhibits the most significant enhancement in model performance. These signal models are trained on data from four physiological signals. By incorporating an attention layer into the hypercomplex layer, which already captures feature relationships, each modality is appropriately weighted before entering the linear layer, effectively enhancing model predictive performance. However, we observed that while partial attention improves accuracy, it doesn't ensure balanced predictions. Additionally, even in attention networks with strong predictive performance, there's room to improve F1 scores. Thus, achieving a balanced prediction in the hypercomplex physiological signal emotion classification model represents a significant research milestone. In future research, we will aim to

streamline the model parameters and refine its structure, ensuring concurrent enhancements in both accuracy and F1 score. In addition, we would like our proposed attention module to be utilized in more advanced fusion networks, such as Fuzzy-Based Deep Attributed Graph Clustering (Yang et al., 2023b), in order to facilitate the accuracy of the corresponding models. The novel network architecture we proposed in our study also has the potential to be used in the field of RNA N6-methyladenosine modification site prediction and drug repositioning in the future after our improvement and refinement (Li et al., 2022, 2024).

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: https://mahnob-db.eu/hci-tagging/.

Author contributions

ZD: Data curation, Formal analysis, Validation, Writing – original draft, Writing – review & editing. YX: Conceptualization, Funding acquisition, Project administration, Writing – review & editing. PZ: Supervision, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This research was funded by the Shaanxi Province Basic Science Research Program, grant number 2024JC-YBQN-0664.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Avula, H., Ranjith, R., and Pillai, A. S. (2022). "Cnn based recognition of emotion and speech from gestures and facial expressions," in *2022 6th International Conference on Electronics, Communication and Aerospace Technology* (IEEE), 1360–1365. doi: 10.1109/ICECA55336.2022.10009316

Bi, F., He, T., Xie, Y., and Luo, X. (2023). Two-stream graph convolutional network-incorporated latent feature analysis. *IEEE Trans. Serv. Comput.* 16, 3027–3042. doi: 10.1109/TSC.2023.3241659

Chang, X., and Skarbek, W. (2021). Multi-modal residual perceptron network for audio-video emotion recognition. *Sensors* 21, 5452. doi: 10.3390/s21165452

Chen, H., Jiang, D., and Sahli, H. (2020). Transformer encoder with multi-modal multi-head attention for continuous affect recognition. *IEEE Trans. Multim.* 23, 4171–4183. doi: 10.1109/TMM.2020.3037496

Dolmans, T. C., Poel, M., van't Klooster, J.-W. J., and Veldkamp, B. P. (2021). Perceived mental workload classification using intermediate fusion multimodal deep learning. *Front. Hum. Neurosci.* 14:609096. doi: 10.3389/fnhum.2020.609096

Du, X., Ma, C., Zhang, G., Li, J., Lai, Y.-K., Zhao, G., et al. (2020). An efficient lstm network for emotion recognition from multichannel EEG signals. *IEEE Trans. Affect. Comput.* 13, 1528–1540. doi: 10.1109/TAFFC.2020.3013711

Gadzicki, K., Khamsehashari, R., and Zetzsche, C. (2020). "Early vs late fusion in multimodal convolutional neural networks," in *2020 IEEE 23rd International Conference on Information Fusion (FUSION)* (IEEE), 1–6. doi: 10.23919/FUSION45008.2020.9190246

Grassucci, E., Zhang, A., and Comminiello, D. (2022). Phnns: lightweight neural networks via parameterized hypercomplex convolutions. *IEEE Trans. Neural Netw. Learn. Syst.* 35, 8293–8305. doi: 10.1109/TNNLS.2022.3226772

Gupta, S., and Sengupta, A. (2023). "Unlocking emotions through heat: Facial emotion recognition via thermal imaging," in *2023 3rd International Conference on Emerging Frontiers in Electrical and Electronic Technologies (ICEFEET)* (IEEE), 1–5. doi: 10.1109/ICEFEET59656.2023.10452206

Hsu, Y.-L., Wang, J.-S., Chiang, W.-C., and Hung, C.-H. (2017). Automatic ECG-based emotion recognition in music listening. *IEEE Trans. Affect. Comput.* 11, 85–99. doi: 10.1109/TAFFC.2017.2781732

Hu, J., Shen, L., and Sun, G. (2018). "Squeeze-and-excitation networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7132–7141. doi: 10.1109/CVPR.2018.00745

Ju, X., Li, M., Tian, W., and Hu, D. (2024). EEG-based emotion recognition using a temporal-difference minimizing neural network. *Cogn. Neurodyn.* 18, 405–416. doi: 10.1007/s11571-023-10004-w

Kaliciak, L., Myrhaug, H., Goker, A., and Song, D. (2014). "On the duality of specific early and late fusion strategies," in *17th International Conference on Information Fusion (FUSION)* (IEEE), 1–8.

Li, G., Zhao, B., Su, X., Yang, Y., Hu, P., Zhou, X., et al. (2024). Discovering consensus regions for interpretable identification of RNA n6-methyladenosine modification sites via graph contrastive clustering. *IEEE J. Biomed. Health Inform.* 28, 2362–2372. doi: 10.1109/JBHI.2024.3357979

Li, J., Tan, F., He, C., Wang, Z., Song, H., Hu, P., et al. (2023). Saliency-aware dual embedded attention network for multivariate time-series forecasting in information technology operations. *IEEE Trans. Ind. Inform.* 20, 4206–4217. doi: 10.1109/TII.2023.3315369

Li, Y., Wu, C.-Y., Fan, H., Mangalam, K., Xiong, B., Malik, J., et al. (2022). "Mvitv2: improved multiscale vision transformers for classification and detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4804–4814. doi: 10.1109/CVPR52688.2022.00476

Lopez, E., Chiarantano, E., Grassucci, E., and Comminiello, D. (2023). "Hypercomplex multimodal emotion recognition from EEG and peripheral physiological signals," in *2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)* (IEEE), 1–5. doi: 10.1109/ICASSPW59220.2023.10193329

Lv, Z., Zhang, J., and Epota Oma, E. (2022). A novel method of emotion recognition from multi-band EEG topology maps based on erenet. *Appl. Sci.* 12:10273. doi: 10.3390/app122010273

Maeng, J.-H., Kang, D.-H., and Kim, D.-H. (2020). Deep learning method for selecting effective models and feature groups in emotion recognition using an Asian multimodal database. *Electronics* 9:1988. doi: 10.3390/electronics9121988

Nakisa, B., Rastgoo, M. N., Rakotonirainy, A., Maire, F., and Chandran, V. (2020). Automatic emotion recognition using temporal multimodal deep learning. *IEEE Access* 8, 225463–225474. doi: 10.1109/ACCESS.2020.3027026

Pan, J., Fang, W., Zhang, Z., Chen, B., Zhang, Z., and Wang, S. (2023). Multimodal emotion recognition based on facial expressions, speech, and EEG. *IEEE Open J. Eng. Med. Biol.* 5, 396–403. doi: 10.1109/OJEMB.2023.3240280

Rayatdoost, S., Rudrauf, D., and Soleymani, M. (2020). "Expression-guided EEG representation learning for emotion recognition," in *ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE), 3222–3226. doi: 10.1109/ICASSP40776.2020.9053004

Rayatdoost, S., and Soleymani, M. (2018). "Cross-corpus EEG-based emotion recognition," in *2018 IEEE 28th international workshop on machine*

learning for signal processing (MLSP) (IEEE), 1–6. doi: 10.1109/MLSP.2018.8517037

Sajjad, M., Kwon, S., et al. (2020). Clustering-based speech emotion recognition by incorporating learned features and deep BILSTM. *IEEE Access* 8, 79861–79875. doi: 10.1109/ACCESS.2020.2990405

Sakurai, M., and Kosaka, T. (2021). "Emotion recognition combining acoustic and linguistic features based on speech recognition results," in *2021 IEEE 10th Global Conference on Consumer Electronics (GCCE)* (IEEE), 824–827. doi: 10.1109/GCCE53005.2021.9621810

Salama, E. S., El-Khoribi, R. A., Shoman, M. E., and Shalaby, M. A. W. (2018). EEG-based emotion recognition using 3d convolutional neural networks. *Int. J. Adv. Comput. Sci. Applic*. 9:43. doi: 10.14569/IJACSA.2018.090843

Shen, S., Gao, Y., Liu, F., Wang, H., and Zhou, A. (2024). "Emotion neural transducer for fine-grained speech emotion recognition," in *ICASSP 2024–2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE), 10111–10115. doi: 10.1109/ICASSP48485.2024.10446974

Stahlschmidt, S. R., Ulfenborg, B., and Synnergren, J. (2022). Multimodal deep learning for biomedical data fusion: a review. *Brief. Bioinform*. 23:bbab569. doi: 10.1093/bib/bbab569

Tan, C., Ceballos, G., Kasabov, N., and Puthanmadam Subramaniyam, N. (2020). Fusion sense: emotion classification using feature fusion of multimodal data and deep learning in a brain-inspired spiking neural network. *Sensors* 20:5328. doi: 10.3390/s20185328

Tao, Y., Huo, S., and Zhou, W. (2020). "Research on communication app for deaf and mute people based on face emotion recognition technology," in *2020 IEEE 2nd International Conference on Civil Aviation Safety and Information Technology (ICCASIT)* (IEEE), 547–552. doi: 10.1109/ICCASIT50869.2020.9368771

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. *arXiv preprint arXiv:1706.03762*.

Wang, K.-Y., Ho, Y.-L., Huang, Y.-D., and Fang, W.-C. (2019). "Design of intelligent EEG system for human emotion recognition with convolutional neural network," in *2019 IEEE International Conference on Artificial Intelligence Circuits and Systems (AICAS)* (IEEE), 142–145. doi: 10.1109/AICAS.2019.8771581

Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., and Hu, Q. (2020). "ECA-net: Efficient channel attention for deep convolutional neural networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11534–11542. doi: 10.1109/CVPR42600.2020.01155

Wang, Y., Wu, Q., Wang, S., Fang, X., and Ruan, Q. (2024). Mi-EEG: Generalized model based on mutual information for EEG emotion recognition without adversarial training. *Expert Syst. Appl.* 244:122777. doi: 10.1016/j.eswa.2023.122777

Yang, W., Li, S., Li, Z., and Luo, X. (2023a). Highly accurate manipulator calibration via extended kalman filter-incorporated residual neural network. *IEEE Trans. Ind. Inform*. 19, 10831–10841. doi: 10.1109/TII.2023.3241614

Yang, Y., Su, X., Zhao, B., Li, G., Hu, P., Zhang, J., et al. (2023b). Fuzzy-based deep attributed graph clustering. *IEEE Trans. Fuzzy Syst*. 35, 1951–1964. doi: 10.1109/TFUZZ.2023.3338565

Ye, Z., Jing, Y., Wang, Q., Li, P., Liu, Z., Yan, M., et al. (2023). Emotion recognition based on convolutional gated recurrent units with attention. *Conn. Sci.* 35:2289833. doi: 10.1080/09540091.2023.2289833

Zeng, F., Lin, Y., Siriaraya, P., Choi, D., and Kuwahara, N. (2020). Emotion detection using EEG and ECG signals from wearable textile devices for elderly people. *J. Textile Eng.* 66, 109–117. doi: 10.4188/jte.66.109

Zhai, S., Talbott, W., Srivastava, N., Huang, C., Goh, H., Zhang, R., et al. (2021). An attention free transformer. *arXiv preprint arXiv:2105.14103*.

Zhang, A., Tay, Y., Zhang, S., Chan, A., Luu, A. T., Hui, S. C., et al. (2021). Beyond fully-connected layers with quaternions: Parameterization of hypercomplex multiplications with $1/n$ parameters. *arXiv preprint arXiv:2102.08597*.

Zhang, Y., Cheng, C., and Zhang, Y. (2022). Multimodal emotion recognition based on manifold learning and convolution neural network. *Multimed. Tools Appl.* 81, 33253–33268. doi: 10.1007/s11042-022-13149-8