frontiers | Frontiers in Neuroscience

TYPE Original Research
PUBLISHED 06 July 2023
DOI 10.3389/fnins.2023.1188696

OPEN ACCESS

EDITED BY
Waldemar Karwowski,
University of Central Florida, United States

REVIEWED BY
Archi Banerjee,
Jadavpur University, India
Shankha Sanyal,
Jadavpur University, India

*CORRESPONDENCE
Jian Lian
✉ lianjianlianjian@163.com

RECEIVED 17 March 2023
ACCEPTED 20 June 2023
PUBLISHED 06 July 2023

CITATION
Zhou Y and Lian J (2023) Identification of emotions evoked by music via spatial-temporal transformer in multi-channel EEG signals. *Front. Neurosci.* 17:1188696. doi: 10.3389/fnins.2023.1188696

Identification of emotions evoked by music via spatial-temporal transformer in multi-channel EEG signals

Yanan Zhou[1] and Jian Lian[2]*

[1]School of Arts, Beijing Foreign Studies University, Beijing, China, [2]School of Intelligence Engineering, Shandong Management University, Jinan, China

Introduction: Emotion plays a vital role in understanding activities and associations. Due to being non-invasive, many experts have employed EEG signals as a reliable technique for emotion recognition. Identifying emotions from multi-channel EEG signals is evolving into a crucial task for diagnosing emotional disorders in neuroscience. One challenge with automated emotion recognition in EEG signals is to extract and select the discriminating features to classify different emotions accurately.

Methods: In this study, we proposed a novel Transformer model for identifying emotions from multi-channel EEG signals. Note that we directly fed the raw EEG signal into the proposed Transformer, which aims at eliminating the issues caused by the local receptive fields in the convolutional neural networks. The presented deep learning model consists of two separate channels to address the spatial and temporal information in the EEG signals, respectively.

Results: In the experiments, we first collected the EEG recordings from 20 subjects during listening to music. Experimental results of the proposed approach for binary emotion classification (positive and negative) and ternary emotion classification (positive, negative, and neutral) indicated the accuracy of 97.3 and 97.1%, respectively. We conducted comparison experiments on the same dataset using the proposed method and state-of-the-art techniques. Moreover, we achieved a promising outcome in comparison with these approaches.

Discussion: Due to the performance of the proposed approach, it can be a potentially valuable instrument for human-computer interface system.

KEYWORDS

human computer interface, emotion classification, deep learning, electroencephalographic, machine learning

1. Introduction

Emotion plays an essential role in the enjoyment of music, which consists of a large variety of affective states consistently reported by people while listening to music (Vuilleumier and Trost, 2015). Music is one of the crucial ways to express emotions. Different music can evoke various emotional responses from listeners. Listening to music is also an easy and effective way to change moods or reduce stress (Cui et al., 2022). In recent decades, music emotion recognition has become one hotspot in neuroscience. Plenty of music emotion recognition applications have achieved promising outcomes in different areas, including automated music composition, psychotherapy, and dancing generation with music (Eerola and Vuoskoski, 2012; Cui et al., 2022).

Electroencephalography (EEG) has received considerable attention in emotion state identification due to its simplicity, inexpensiveness, and portability (Alarcao and Fonseca, 2017). Specifically, EEG has been extensively employed in evaluating the impact of music on human brain activity, e.g., human emotion (Lin et al., 2006). For instance, Sammler et al. (2007) explored whether the valence of emotions would affect EEG power spectra and heart rate differently. The experiments collected pleasant and unpleasant emotions induced by a consonant and dissonant music using EEG signals. This study showed that pleasant music is significantly related to increased EEG theta power, while unpleasant music could significantly decrease heart rate. In the work of Balasubramanian et al. (2018), the authors analyzed the emotional responses of human beings to music in EEG. Ten healthy subjects (average age is 20) participated in this study. The self-assessment manikin (SAM) test assessed the subjects' perceived emotions when listening to music. The outcome from their experiments demonstrated an increase in the theta band of the frontal midline for liked music, and an increase in the beta band was found for disliked music. The publicly available dataset DEAP was presented in the work of Koelstra et al. (2011) for analyzing the human affective states. EEG and other physiological signals from 32 participants were recorded as each watched 40 1-min excerpts of music videos. Moreover, the participants must rate the videos regarding arousal, valence, like/dislike, dominance, and familiarity. In addition, Ozel et al. (2019) presented an approach for emotion recognition using time-frequency analysis of multivariate synchrosqueezing transform in multi-channel EEG signals. The evaluation of this study employed the DEAP, and a total of eight emotional states were considered by combining arousal, valence, and dominance. Lin et al. (2010) leveraged machine learning-based methods to classify EEG dynamics based on self-assessed emotional states when subjects listen to music. A framework was presented to optimize emotion recognition in EEG signals by extracting emotion-specific features from EEG recordings and enhancing the effectiveness of the classifiers. Zheng et al. (2017) aimed to identify EEG stability in the process of emotion recognition by comparing the performance of various feature extraction, feature selection, feature smoothing, and classification methods on the DEAP dataset and the SEED (Zheng and Lu, 2015) dataset. Generally, a regularized robust learning algorithm with differential entropy features yields the optimal outcome on the DEAP and SEED datasets. Using both a private dataset and the public dataset DEAP, Thammasan et al. (2017) investigated the influence of familiarity on brain activity in EEG signals. In the experiments, the subjects were asked to determine an equal number of familiar and unfamiliar songs; the datasets' outcomes both demonstrated the significance of self-emotion assessment according to the assumption that the emotional state is subjective when listening to music. Hou and Chen (2019) conducted experiments to extract the optimal EEG features induced by music evoking various emotions, including calm, joy, sadness, and anger. The 27-dimensional features were generated from EEG signals during the feature extraction process. A feature selection method was exploited to identify the features significantly related to the emotions evoked by music. Since music emotion recognition has attracted widespread attention with the enhancement of deep learning-based artificial intelligence applications, Han et al. (2022) surveyed music emotion recognition. The evaluation metrics, emotion recognition algorithms, datasets, and features involved were provided in detail. Recently, there have been numerous applications of deep learning methods in music classification and recognition (McIntosh, 2021; Eskine, 2022; Nag et al., 2022; Daly, 2023).

Several limitations still need to be resolved in the current automatic music emotion classification algorithms in EEG signals. On the one hand, for the machine learning-based methods, a set of optimal features is required to be extracted and selected from the EEG recordings. On the other hand, deep learning-based approaches, especially convolutional neural networks (CNNs), are prone to a need for global associations between long-range sampling signals due to the local receptive field issue of CNN-based models. Bearing the analysis mentioned above in mind, we proposed a spatial-temporal transformer-based pipeline for music emotion identification. The presented approach extracts the spatial features from EEG signals from different subjects listening to the same music. The temporal features are extracted from EEG signals from the same subject listening to the same type of music. Experimental results demonstrate that the performance of the proposed approach outperforms the state-of-the-art techniques in binary and ternary music emotion classification. The proposed approach could be a valuable instrument for various applications built upon music emotion recognition.

The remaining of this following content is organized as follows: First, the data samples used in this study and the details of the proposed transformer model are described in Section 2; Section 3 gives the experimental results of the proposed approach on the collected EEG recordings and the comparison between the state-of-the-art techniques and our proposed method; the discussion of the experimental results and the proposed method are given in Section 4; Finally, the conclusion of this study is given in Section 5.

2. Materials and methods

In this section, the process of collecting the EEG recordings used in this study is introduced first. Then, the content of the proposed deep learning model is given in detail.

2.1. Dataset

A dataset was established by capturing three types of emotions, including positive, negative, and neural, from the EEG recordings. The entire workflow is illustrated in Figure 1.

Before the EEG collection process, the following preparations were performed:

- Ensure that the participants participated in this experiment on an entirely voluntary basis;

Abbreviations: EEG, electroencephalographic; CNN, convolutional neural network; GPU, graphical processing unit; TP, true positive; FN, false negative; TN, true negative; FP, false positive.

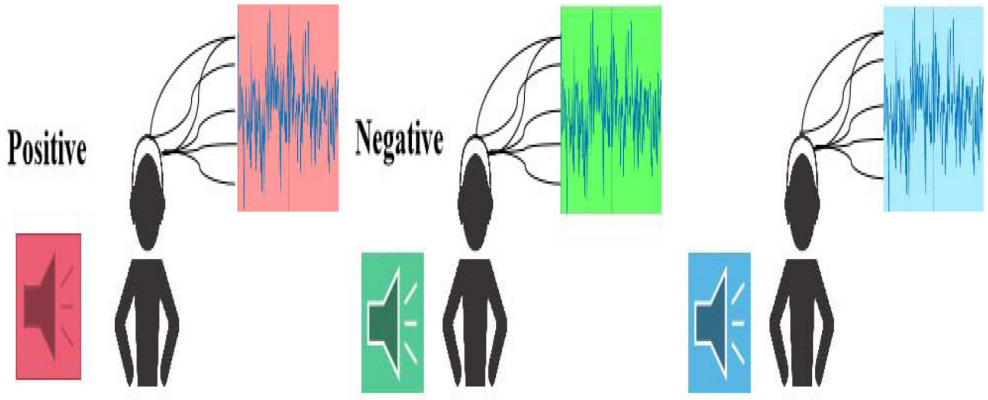

FIGURE 1
The workflow of the music emotion EEG signals collection.

- Participants must read the experimental precautions and execution process, including the errors that could be caused by physical shaking and emotional tension;
- Participants need to fill in the personal information form and check whether the electrode is in good contact and whether the electrode cap is placed correctly;
- The participants put on the electrode cap, adjusted the best physical and psychological state, and pressed the button to start the test when ready.

During the EEG collection process, 32 participants (16 females and 16 males) aged 18–31 (mean age is 24.7) were enrolled, all working or studying at the same University. These 32 subjects are in good physical and mental condition, without mental illness or brain damage. In addition, two specialists (two females) at the elicitation of emotion majoring in psychology contributed to determine the adopted music clips.

During the data collection process, each subject needs to listen to 12 pieces of music excerpts (four pieces of positive music excerpts, four pieces of negative music excerpts, and four pieces of neutral music excerpts, as shown in Table 1), which are different kinds of music clips with a uniform duration of 1 min. To note that the sequence of playing the music clips is randomized for the participants to neutralize the retention effect of the emotion evoked by the previous music clip on the next one. The EEG acquisition equipment used is the Biosemi ActiveTwo system, with 32 EEG signal channels, using 512 Hz sampling, 128 Hz complex sampling (pre-processing), and using the 10-20 system EEG signals.

The positions of 32 EEG channels are selected according to the international 10-20 system, namely Fp1, AF3, F3, F7, FC5, FC1, C3, T7, CP5, CP1, P3, P7, PO3, O1, Oz, Pz, Fp2, AF4, Fz, F4, F8, FC6, FC, Cz, C4, T8, Cp6, Cp2, P4, P8, PO4, O2. The position distribution of electrode placement covers the four primary areas of the brain, with moderate spacing, which can effectively collect the required EEG raw data.

The specific process of EEG acquisition is shown in the following figure, mainly including the following steps:

TABLE 1 The types and descriptions of the music excerpts used in this study.

ID	Type	Title	Singer
1	Positive	A man should strengthen himself	Zixiang Lin
2	Negative	A dream of misery	Wei Dou
3	Neutral	Reiki meditation	Not applicable
4	Positive	I really love you	Beyond
5	Negative	In case	SHIN
6	Neutral	Calm tech	Not applicable
7	Positive	Flying higher	Feng Wang
8	Negative	Collapse	Catcher in the Rye
9	Neutral	Let the sun shine	Not applicable
10	Positive	Friend	Huajian Zhou
11	Negative	Negative	Bill Do
12	Neutral	Make it rain	Not applicable

- The Serial number of the progress of the experiment. Give the serial number through the prompt sound to let participants know what music is currently in progress;
- Collection of benchmark records. This process will last for 5 s. At this time, participants try to keep calm and record the mark of the beginning of the EEG signal;
- Music playback. This process will last for 75 s. The 15 s is the time for each music switch, and the 60 s is the time for the music to play. During this process, participants need to keep their body balance and reduce movement as much as possible;
- Self-assessment scoring. After listening to each music excerpt, the participants must evaluate themselves in time (−1 denotes negative, 0 denotes neutral, and +1 denotes positive). This process will last <15 s, equal to the period for the music switch. In addition, the participants need to take a short rest after self-assessment according to their real emotional experience after listening to music;

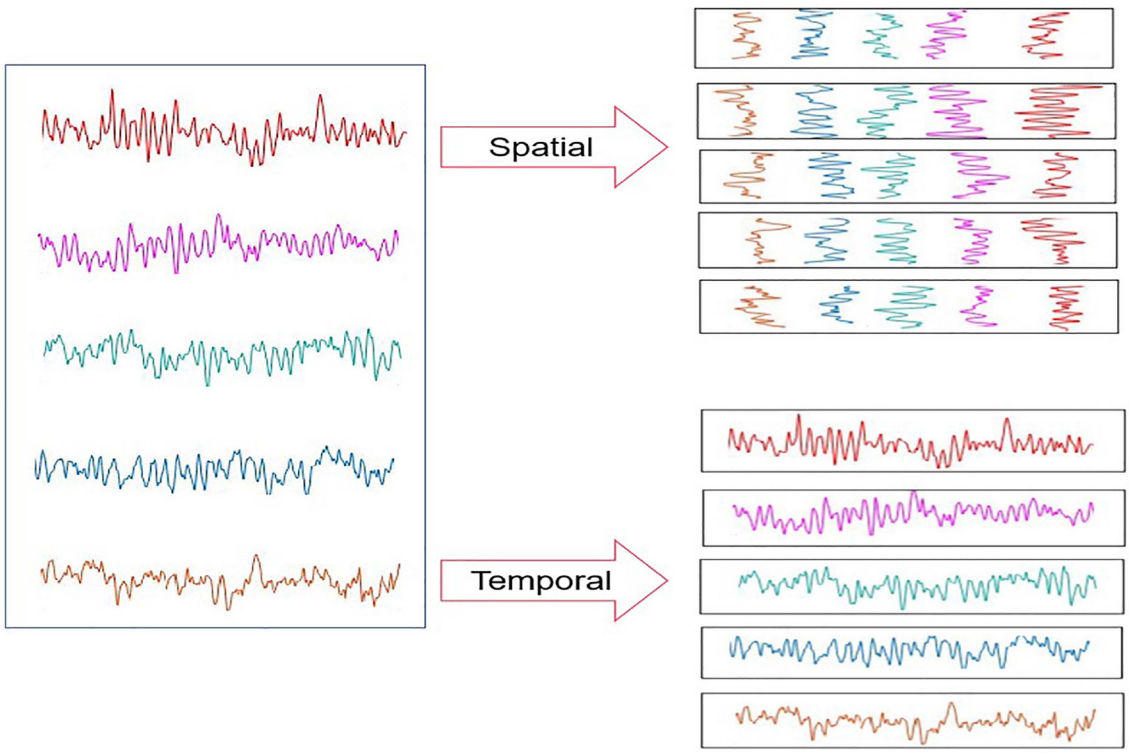

FIGURE 2
The input of the proposed spatial-temporal transformer.

- Start playing the next piece of music. Repeat the above two steps until all 12 music materials have been played.

The experiment uses the software provided by the company to play music. The CPU of the computer playing video is Core i3; the memory is 4 GB, and the Windows 10 operating system. Notably, the EEG signals were recorded while listening to music. All EEG signals were recorded between 9 and 11 a.m., and 2 and 5 p.m. to ensure the participants were not tired. Also, to avoid the noise from different sources, including Electro-Oculogram (EOG) and Electrocardiograph (ECG), all the subjects were instructed to keep their eyes closed during the EEG recording process.

For the binary classification task (positive and negative), 8 min (positive = 240 s and negative = 240 s) of sampled EEG signals were used for two types of emotions. Furthermore, the EEG samples from each electrode were divided into overlapped 240 epochs, each lasting 4 s. On the other hand, for the ternary classification task (positive, negative, and neutral), 12 min (positive = 240 seconds, negative = 240 s, and neutral = 240 s) of sampled EEG signals were used for three types of emotions. The EEG signals from each electrode were divided into overlapped 360 epochs, each lasting 4 s. Considering the balance of classification, the number of samples for each type of emotion used in both binary and ternary classification equals each other. In addition, the overlapped epochs were leveraged to decrease the possibility of over-fitting during the classification processes. Note that the complex sampling is 128 Hz for the EEG signals in this study. Thus each epoch of 4 s contains 512 sampling points in total.

2.2. Spatial-temporal transformer

In this section, details about the proposed spatial-temporal transformer model are provided. Note that this transformer is constructed using a self-attention mechanism, which has been extensively employed in many previous works (Fan et al., 2021; Liu et al., 2021; Wang et al., 2021). In addition, we conceived that the transformer-based deep learning architecture could produce competitive outcomes over the CNN-based algorithms, which are constrained by the issues caused by the local receptive field. The pipeline of the presented spatial-temporal transformer is shown in Figure 2, inspired by the model of ViT (Dosovitskiy et al., 2020) that leverages both the image patches and the corresponding sequence of these image patches as the input for the model.

In general, the input for the presented spatial-temporal transformer is composed of both the spatial and temporal epochs of the EEG signals captured from the same subject. Furthermore, to integrate the spatial and temporal associations between a batch of EEG signals (each group has eight epochs), the corresponding position embedding was fed into the proposed transformer (as shown in Figure 2). It is notable that each input batch of EEG epochs belongs to the same type of emotion.

Suppose that H and W denote the height and width of the input of each channel in the proposed model for the EEG epochs divided from the original EEG signals, $H = 5$ and $W = 512$; Let C represent the number of channels, and the epochs were flattened and mapped into the vector in a length of D. Inspired by the embedding used in ViT (Dosovitskiy et al., 2020), we also added

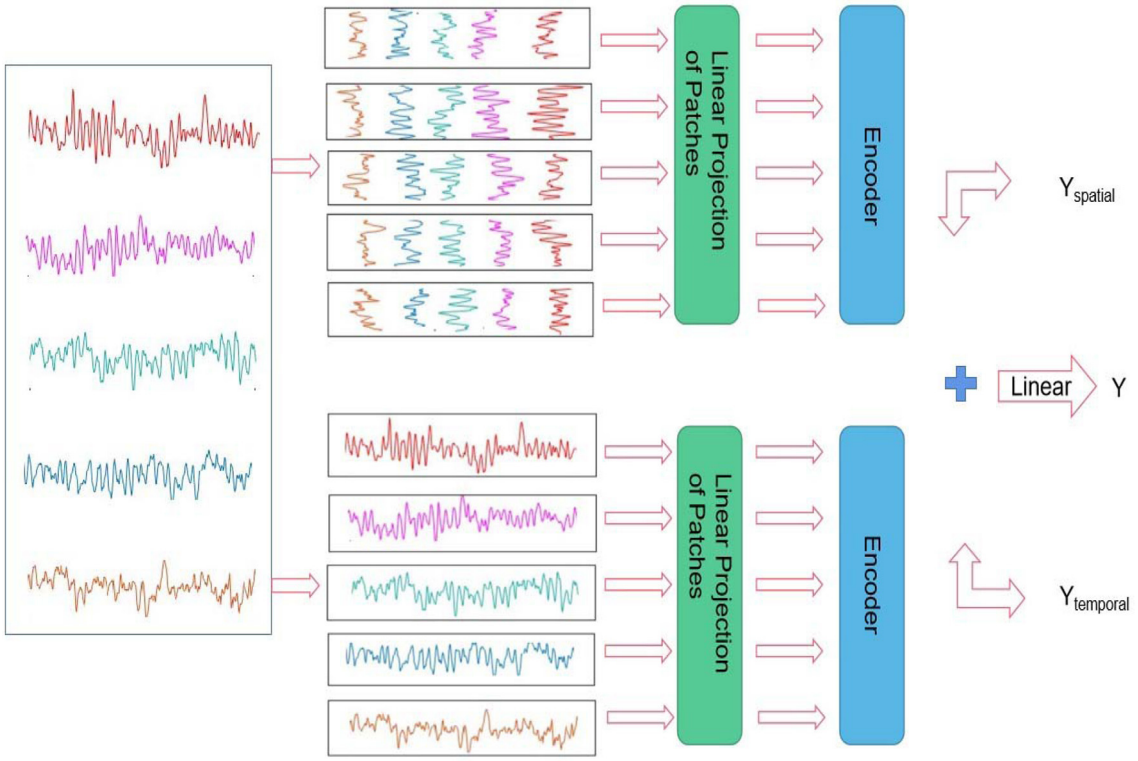

FIGURE 3
Architecture of the presented spatial-temporal transformer.

a learnable embedding along with the epoch embedding sequence. The output corresponding to the input epochs from the spatial channel is Yspatial, the output corresponding to the input epochs from the temporal channel is Ytemporal, and the composite output is denoted as Y (as shown in Figure 3).

As shown in Figure 3, the proposed spatial-temporal transformer model consists of two channels designed for the spatial and temporal EEG epochs without sharing weighting parameters between the channels. According to the upper and bottom components on the right side of Figure 3, the upper and bottom channels are supposed to address the spatial and temporal information within the input EEG samples, respectively. Furthermore, inspired by the structure of ViT (Dosovitskiy et al., 2020) and the original transformer model (Vaswani et al., 2017), the proposed model adopts a sequence of token embedding and the component of the transformer Encoder (as shown in Figure 4).

The Encoder in each channel of the proposed transformer consists of L layers. Furthermore, each layer contains multiple self-attention units and a multi-layer perception unit. Along with the attention unit, the layer Normalization mechanism is exploited before both units. The multiple attention unit with H heads in the proposed model is constructed using the self-attention mechanism, which can be used to measure the similarity between each query and the keys by allocating a weight for each value (Vaswani et al., 2017). Thus, the output of the proposed approach can be obtained from a weighted sum of the outputs from the two channels. Moreover, a linear operator is leveraged to integrate the outputs from two channels, which can be mathematically formulated as (Equation 1):

$$Y = Linear(layernNormalization(Y_{spatial}, Y_{temporal}), \quad (1)$$

To be specific, besides the composite output Y, the outcomes of $Y_{spatial}$ and $Y_{temporal}$ can also be used to implement the classification of input EEG samples, respectively.

3. Results

In this section, the experimental results of the proposed method was provided for automated music emotion classification.

3.1. Implementation details

The proposed algorithm was realized using the PyTorch (Paszke et al., 2019) framework and two NVidia Telsa V100 Graphical Processing Units (GPUs) with 32GB RAM. The hyper-parameters for the proposed network were determined by using a trial-and-error fashion. To be specific, the model training was performed by adopting GELU (Hendrycks and Gimpel, 2016) activation function and AMSGrad optimizer (Reddi et al., 2019) with the learning rate of 0.001.

To evaluate the performance of the proposed approach, a set of experiments were carried out in sequence. To determine the

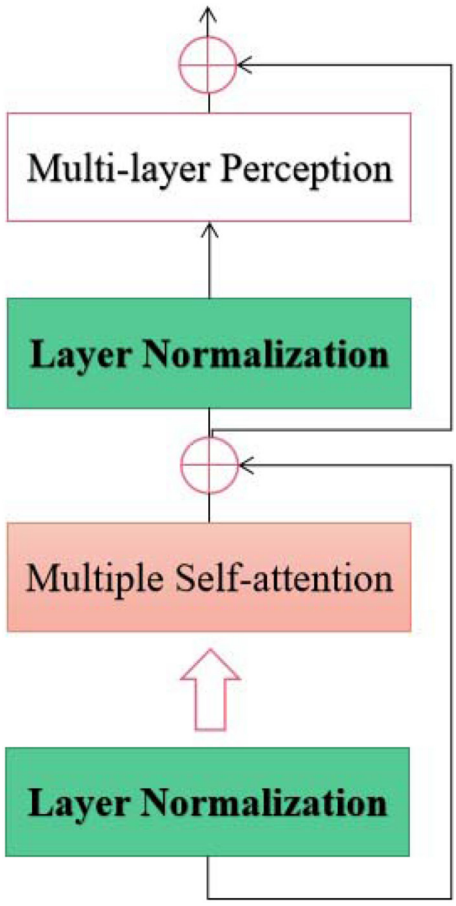

FIGURE 4
The Encoder Unit used in the proposed spatial-temporal transformer.

optimal combination of the spatial and temporal channels in the proposed approach, we adopted the loss hyperparameter α and found the optimal value of α by training on half of the training samples before starting the following experiments (as shown in Figure 5). As shown in Figure 5, the optimal value of α is set to **0.3** in the experiments. Firstly, we carried out the ablation study by comparing the outcomes of the separate spatial channel, separate temporal channel, and the composite channels with each other. In addition, we evaluated the impact of number of heads and layers (L) of the proposed transformer model in the ablation study. We then conducted the experiments on all subjects in the manually collected dataset to build the baseline for the remaining experiments. In addition, the comparison experiments were conducted between the state-of-the-art deep learning methods and ours on the manually collected dataset. Finally, we provided the comparison between the state-of-the-art algorithms and ours on the publicly available DEAP dataset (Koelstra et al., 2011). To be specific, sensitivity, specificity, and accuracy were leveraged as the evaluation metrics in the experiments.

Meanwhile, to improve the performance of the proposed method, an integrated loss function was used by integrating both the spatial and temporal components, as shown in Equation (2).

$$Loss = \alpha Loss_{spatial} + (1-\alpha)Loss_{temporal}, \tag{2}$$

where α denotes the weighting parameter for determining the combination of spatial and t temporal channels of the proposed model, $Loss_{spatial}$ and $Loss_{temporal}$ represent the cross entropy loss of the separate channels in the proposed transformer, respectively. In addition, the following metrics were used to evaluate the performance of the techniques in the experiments, including sensitivity, specificity, and accuracy, which are provided in Equations (3)–(5).

$$Sensitivity = \frac{TP}{TP+FN}, \tag{3}$$

$$Specificity = \frac{TN}{TN+FP}, \tag{4}$$

$$Accuracy = \frac{TP+TN}{TP+FN+TN+FP}, \tag{5}$$

where TP, FN, TN, FP denote true positive, false negative, true negative, and false positive, respectively.

Furthermore, a 10-fold cross-validation mechanism was employed in the experimental process. First, all the input EEG samples were divided into ten subsets with the same number of samples. In each round of ten rounds, one subgroup was taken as the testing set, and the remaining groups were used as the training set. At last, the average of 10-folds was leveraged as the outcome for the methods used in the experiments.

3.2. Ablation study

3.2.1. Hybrid model and individual transformers

The proposed approach could be treated as a hybrid model due to the spatial and temporal channels employed. Therefore, we first compared the proposed model with individual channels and the model with both channels. Due to the difference in sensitivity, specificity, and accuracy values of the three different models, including spatial, temporal, and combined models in the binary classification task, it can be observed that the integration of the two channels is superior in various performances over the individual channels (as shown in Figure 6).

3.2.2. Impact of number of heads and layers of the proposed transformer

In addition, with the proposed hybrid model with both spatial and temporal channels, we further evaluated the impact of number of heads and layers on the binary-classification performance of the proposed the transformer model. As shown in Table 2, the optimal combination of the number of heads and layers is 4 and 8.

To ensure that the number of weighting parameters of the proposed model remains at a low level, we did not take more combinations of the number of heads and layers (e.g., 12 or 24) in the ablation study.

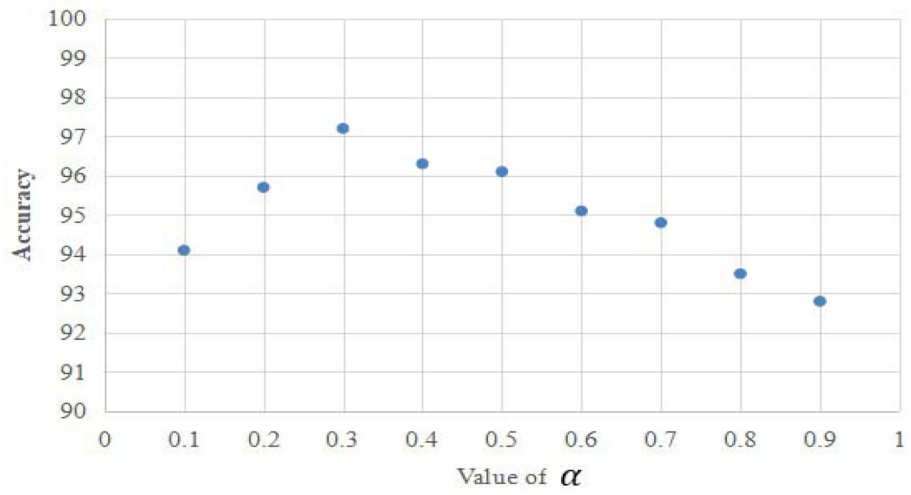

FIGURE 5
The optimal value of the hyperparameter α used in the experiments.

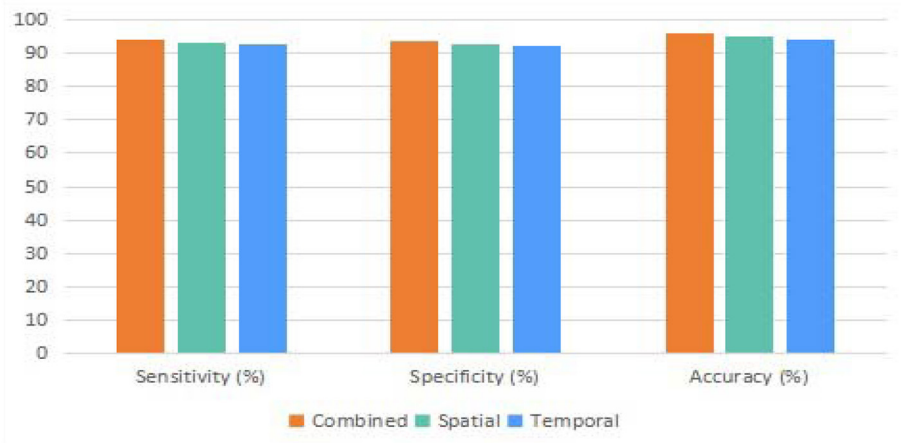

FIGURE 6
The binary classification comparison between the combined, spatial, and temporal versions of the proposed method.

3.3. Binary and ternary classification outcomes of the proposed approach

As shown in Figure 7, the sensitivity, specificity, and accuracy of the proposed approach in the binary classification task are 95.3, 94.8, and 97.3%. Meanwhile, the sensitivity, specificity, and accuracy of the proposed approach in the ternary classification task are 94.1, 93.2, and 97.1%. Note that both the evaluation metrics generated from the binary classification are more significant than the ternary classification by using the proposed algorithm since there are more samples and more types of samples in the ternary classification process.

3.4. Comparison experiments between the state-of-the-arts and the proposed approach

Furthermore, the following algorithms were incorporated into the comparison experiments between the state-of-the-art techniques and the presented approach, including U-Net (Ronneberger et al., 2015), Mask R-CNN (He et al., 2017), ExtremeNet (Zhou et al., 2019), TensorMask (Chen et al., 2019), Visual Transformer (Wu et al., 2020), ViT (Dosovitskiy et al., 2020), MViT (Fan et al., 2021), PVT (Wang et al., 2021), PiT (Heo et al., 2021), and Swin Transformer (Liu et al., 2021). As

shown in Tables 3, 4, the proposed approach has achieved superior performance over the state-of-the-art algorithms.

To note that the specificity of the algorithm (Rudakov, 2021) is superior over the proposed approach.

To evaluate the performance of the proposed approach in a fair fashion, we further evaluated the performance of the proposed approach and the state-of-the-art algorithms (Shawky et al., 2018; Yang et al., 2018, 2019; Xing et al., 2019; Shen et al., 2020; Pan and Zheng, 2021; Rudakov, 2021; Kan et al., 2022; Zhang et al., 2022) on the DEAP dataset (as shown in Table 5). Accordingly, we used the DEAP dataset during the training phase of the proposed approach in this stage.

4. Discussion

In this study, a dual-channel transformer for music emotion classification was presented. The proposed model consists of two channels designed to extract spatial and temporal information from the EEG signals. To our best knowledge, this is an early work of transformer architecture in this type of machine-learning task. From the experimental results, it can be observed that our method has achieved superior performance (Binary classification: sensitivity 94.3%, specificity 93.8%, and accuracy 96.8%; Ternary classification: sensitivity 93.6%, specificity 92.2%, and accuracy

TABLE 2 The influence of number of heads and number of layers on the binary-classification outcome of the proposed model.

Model	Number of heads (H)	Number of layers (L)	Accuracy (%)
STT_1_4	1	4	96.1
STT_1_8	1	8	96.7
STT_1_12	1	12	96.6
STT_2_4	2	4	95.8
STT_2_8	2	8	96.7
STT_2_12	2	12	97.1
STT_4_4	4	4	96.8
STT_4_8	4	8	**97.3**
STT_4_12	4	12	97.1

The bold values denote the best performance.

TABLE 3 Binary classification comparison between the state-of-the-arts and the proposed approach.

Method	Sensitivity (%)	Specificity (%)	Accuracy (%)
U-Net (Ronneberger et al., 2015)	82.3	83.1	84.8
Mask R-CNN (He et al., 2017)	82.5	82.9	83.6
ExtremeNet (Zhou et al., 2019)	81.9	84.7	85.1
TensorMask (Chen et al., 2019)	82.9	83.8	84.7
Visual Transformer (Wu et al., 2020)	85.5	84.1	86.2
ViT (Dosovitskiy et al., 2020)	86.7	87.1	88.5
MViT (Fan et al., 2021)	89.4	87.9	90.3
PVT (Wang et al., 2021)	89.7	86.5	90.1
PiT (Heo et al., 2021)	92.4	90.5	93.8
Swin Transformer (Liu et al., 2021)	91.7	92.5	94.4
Pan and Zheng (2021)	91.2	90.8	90.4
Shen et al. (2020)	93.6	93.2	93.9
Kan et al. (2022)	94.2	93.9	94.3
Rudakov (2021)	95.2	94.5	95.8
The proposed approach	**95.3**	**94.8**	**97.3**

The bold values denote the best performance.

FIGURE 7
The binary and ternary classification outcome of the proposed transformer.

TABLE 4 Ternary classification comparison between the state-of-the-arts and the proposed approach.

Method	Sensitivity (%)	Specificity (%)	Accuracy (%)
U-Net (Ronneberger et al., 2015)	80.5	81.6	82.2
Mask R-CNN (He et al., 2017)	81.7	82.1	82.8
ExtremeNet (Zhou et al., 2019)	81.2	83.1	84.3
TensorMask (Chen et al., 2019)	82.3	83.1	85.2
Visual transformer (Wu et al., 2020)	83.4	84.2	85.9
ViT (Dosovitskiy et al., 2020)	85.6	86.9	88.7
MViT (Fan et al., 2021)	88.6	86.7	88.1
PVT (Wang et al., 2021)	88.2	87.2	89.5
PiT (Heo et al., 2021)	90.2	89.4	91.1
Swin transformer (Liu et al., 2021)	91.2	91.9	92.5
Pan and Zheng (2021)	90.7	91.0	90.2
Shen et al. (2020)	92.8	92.3	92.9
Kan et al. (2022)	93.5	93.8	93.9
Rudakov (2021)	93.8	**93.5**	95.3
The proposed approach	**94.1**	93.2	**97.1**

The bold values denote the best performance.

TABLE 5 Comparison between the state-of-the-arts and the proposed approach on DEAP dataset.

Method	Detail	Accuracy	
		Valence	Arousal
Xing et al. (2019)	LSTM*	81.10	74.38
Shawky et al. (2018)	CNN	87.44	88.49
Yang et al. (2019)	CNN	90.01	90.65
Pan and Zheng (2021)	CNN	90.26	88.90
Yang et al. (2018)	LSTM	90.82	86.13
Shen et al. (2020)	CRNN**	94.22	94.58
Zhang et al. (2022)	GAN***	93.52	94.21
Kan et al. (2022)	Contrastive Learning	94.72	92.65
Rudakov (2021)	CNN	96.28	96.62
Ours	Transformer	**96.36**	**96.91**

*LSTM denotes long short term memory.
**CRNN denotes convolutional recurrent neural network.
***GAN denotes generative adversarial network.
The bold values denote the best performance.

95.3%) over the state-of-the-art algorithms in both the binary and ternary classification tasks. To note that the neutral emotions evoked by the music clips get more misclassified than both the positive and negative emotions in the experiments. The competing methods used in the comparison experiments include the CNN-based and transformer-based deep learning models, which focus on local receptive and global receptive fields during the classification tasks, respectively. Since the spatial and temporal information can be extracted from the input EEG signals, the proposed approach has achieved superior performance over both CNNs and transformers. Notably, the input of a batch of EEG signals could be treated the same as an image for the CNNs and vision transformers mentioned in the comparison experiments.

Furthermore, according to the results of the ablation study, the combination of the spatial and temporal channels in the proposed transformer network is superior to the individual spatial channel or temporal channel. Meanwhile, the outcome of the proposed model in the ablation study has also proved that a hybrid model could be a valuable mechanism for enhancing the performance of individual deep-learning models. In addition, the impact of number of heads and layers of the proposed transformer was evaluated in another ablation study. The corresponding results demonstrate that the optimal number of heads and layers of are 12 and 6 for the proposed model.

Notably, the proposed music emotion classification framework is inspired by the ViT (Dosovitskiy et al., 2020) architecture. Different from the original transformer (Vaswani et al., 2017) used for natural language processing and the vision transformers (He et al., 2017; Chen et al., 2019; Zhou et al., 2019; Dosovitskiy et al., 2020; Wu et al., 2020; Fan et al., 2021; Heo et al., 2021; Liu et al., 2021; Wang et al., 2021) designed for image analysis, our method is primarily exploited to classify the EEG recordings. Moreover, the proposed approach needs to adapt to the requirements of EEG classification by using both the spatial and temporal modules. Moreover, the self-attention mechanism first presented in the work of Vaswani et al. (2017), the association between the global associations between the EEG fragments could be fully exploited and unveiled. Moreover, this is a primary strategy to enhance the discriminating ability of the proposed transformer.

In addition, this study has some limitations despite its contributions. First, we leveraged a private dataset in the experiments, and the publicly available dataset should be considered in the following steps. Secondly, the weighting parameters used in the proposed approach were selected using a trial-and-error strategy, which might not be the optimal choice. Moreover, a strategy with more interpretability should be exploited instead. Furthermore, the training process of the proposed approach was time-consuming, which could be resolved using more powerful GPU platforms. However, to implement the proposed approach in practical applications, the computation resources and the execution time should be decreased through lightweight parameter configuration. In the future, more types of emotions in music will be incorporated into our study to satisfy the requirements of different categories of the practical applications.

5. Conclusion

In this study, the transformer-based deep learning model generally enhances the performance of music emotion classification in EEG recordings compared with traditional deep learning techniques. Specifically, the ability of the global receptive field

provided by the transformer is conceived as beneficial for unveiling the long-range associations between different components in EEG signals. Both the binary classification and ternary classification of emotions evoked by music of the proposed approach are promising. In addition, the proposed approach has achieved superior performance over the state-of-the-art deep learning techniques.

In the future, we will incorporate more participants and more music clips from various countries and cultures to improve the robustness of the proposed approach. In addition, more types of emotions evoked by music will also be taken into consideration in our future work.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving human participants were reviewed and approved by the Shandong Management University's Human Research Ethics Committee (2023031501). The patients/participants provided their written informed consent to participate in this study.

Author contributions

JL: conceptualization, formal analysis, funding acquisition, methodology, project administration, and supervision. YZ and JL: data curation. YZ: investigation, validation, visualization, and writing—original draft. All authors contributed to the article and approved the submitted version.

Funding

This study was supported by the Natural Science Foundation of Shandong Province (ZR2020MF133), Key Laboratory of Public Safety Management Technology of Scientific Research and Innovation Platform in Shandong Universities during the 13th 5-Year Plan Period, Collaborative Innovation Center of Internet plus intelligent manufacturing of Shandong Universities, and Intelligent Manufacturing and Data Application Engineering Laboratory of Shandong Province.

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

Alarcao, S. M., and Fonseca, M. J. (2017). Emotions recognition using EEG signals: a survey. *IEEE Trans. Affect. Comput.* 10, 374–393. doi: 10.1109/TAFFC.2017.2714671

Balasubramanian, G., Kanagasabai, A., Mohan, J., and Seshadri, N. G. (2018). Music induced emotion using wavelet packet decomposition–an EEG study. *Biomed. Signal Process. Control* 42, 115–128. doi: 10.1016/j.bspc.2018.01.015

Chen, X., Girshick, R., He, K., and Dollár, P. (2019). "Tensormask: a foundation for dense object segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (Seoul), 2061–2069. doi: 10.1109/ICCV.2019.00215

Cui, X., Wu, Y., Wu, J., You, Z., Xiahou, J., and Ouyang, M. (2022). A review: music-emotion recognition and analysis based on EEG signals. *Front. Neuroinform.* 16:997282. doi: 10.3389/fninf.2022.997282

Daly, I. (2023). Neural decoding of music from the EEG. *Sci. Rep.* 13:624. doi: 10.1038/s41598-022-27361-x

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al. (2020). An image is worth 16 × 16 words: transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Eerola, T., and Vuoskoski, J. K. (2012). A review of music and emotion studies: approaches, emotion models, and stimuli. *Mus. Percept.* 30, 307–340. doi: 10.1525/mp.2012.30.3.307

Eskine, K. (2022). Evaluating the three-network theory of creativity: effects of music listening on resting state EEG. *Psychol. Mus.* 51, 730–749. doi: 10.1177/03057356221116141

Fan, H., Xiong, B., Mangalam, K., Li, Y., Yan, Z., Malik, J., et al. (2021). "Multiscale vision transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (Montreal Canada), 6824–6835. doi: 10.1109/ICCV48922.2021.00675

Han, D., Kong, Y., Han, J., and Wang, G. (2022). A survey of music emotion recognition. *Front. Comput. Sci.* 16:166335. doi: 10.1007/s11704-021-0569-4

He, K., Gkioxari, G., Dollár, P., and Girshick, R. (2017). "Mask R-CNN," in *Proceedings of the IEEE International Conference on Computer Vision* (Venice), 2961–2969. doi: 10.1109/ICCV.2017.322

Hendrycks, D., and Gimpel, K. (2016). Bridging nonlinearities and stochastic regularizers with Gaussian error linear units. *CoRR*. abs/1606.08415. doi: 10.48550/arXiv.1606.08415

Heo, B., Yun, S., Han, D., Chun, S., Choe, J., and Oh, S. J. (2021). "Rethinking spatial dimensions of vision transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (Montreal Canada), 11936–11945. doi: 10.1109/ICCV48922.2021.01172

Hou, Y., and Chen, S. (2019). Distinguishing different emotions evoked by music via electroencephalographic signals. *Comput. Intell. Neurosci.* 2019:3191903. doi: 10.1155/2019/3191903

Kan, H., Yu, J., Huang, J., Liu, Z., and Zhou, H. (2022). Self-supervised group meiosis contrastive learning for eeg-based emotion recognition. *arXiv:2208.00877*. doi: 10.48550/arXiv.2208.00877

Koelstra, S., Muhl, C., Soleymani, M., Lee, J.-S., Yazdani, A., Ebrahimi, T., et al. (2011). Deap: a database for emotion analysis; using physiological signals. *IEEE Trans. Affect. Comput.* 3, 18–31. doi: 10.1109/T-AFFC.2011.15

Lin, W.-C., Chiu, H.-W., and Hsu, C.-Y. (2006). "Discovering EEG signals response to musical signal stimuli by time-frequency analysis and independent component analysis," in *2005 IEEE Engineering in Medicine and Biology 27th Annual Conference* (Shanghai), 2765–2768.

Lin, Y.-P., Wang, C.-H., Jung, T.-P., Wu, T.-L., Jeng, S.-K., Duann, J.-R., et al. (2010). EEG-based emotion recognition in music listening. *IEEE Trans. Biomed. Eng.* 57, 1798–1806. doi: 10.1109/TBME.2010.2048568

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., et al. (2021). "Swin transformer: hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (Montreal Canada), 10012–10022. doi: 10.1109/ICCV48922.2021.00986

McIntosh, T. (2021). Exploring the relationship between music and emotions with machine learning. *EVA*. doi: 10.14236/ewic/EVA2021.49

Nag, S., Basu, M., Sanyal, S., Banerjee, A., and Ghosh, D. (2022). On the application of deep learning and multifractal techniques to classify emotions and instruments using Indian classical music. *Phys. A* 597:127261. doi: 10.1016/j.physa.2022.127261

Ozel, P., Akan, A., and Yilmaz, B. (2019). Synchrosqueezing transform based feature extraction from EEG signals for emotional state prediction. *Biomed. Signal Process. Control* 52, 152–161. doi: 10.1016/j.bspc.2019.04.023

Pan, B., and Zheng, W. (2021). Emotion recognition based on EEG using generative adversarial nets and convolutional neural network. *Comput. Math. Methods Med.* 2021:2520394. doi: 10.1155/2021/2520394

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., et al. (2019). "Pytorch: an imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems, Vol. 32*. Vancouver.

Reddi, S. J., Kale, S., and Kumar, S. (2019). On the convergence of Adam and beyond. *CoRR*. abs/1904.09237. doi: 10.48550/arXiv.1904.09237

Ronneberger, O., Fischer, P., and Brox, T. (2015). "U-Net: convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention-MICCAI 2015: 18th International Conference* (Munich: Springer), 234–241. doi: 10.1007/978-3-319-24574-4_28

Rudakov, E. (2021). "Multi-task CNN model for emotion recognition from EEG brain maps," in *2021 4th International Conference on Bio-Engineering for Smart Technologies (BioSMART)* (New York, NY). doi: 10.1109/BioSMART54244.2021.9677807

Sammler, D., Grigutsch, M., Fritz, T., and Koelsch, S. (2007). Music and emotion: electrophysiological correlates of the processing of pleasant and unpleasant music. *Psychophysiology* 44, 293–304. doi: 10.1111/j.1469-8986.2007.00497.x

Shawky, E., El-Khoribi, R., Shoman, M., and Wahby Shalaby, M. (2018). Eeg-based emotion recognition using 3d convolutional neural networks. International Journal of Advanced Computer Science and Applications, 9. doi: 10.14569/IJACSA.2018.090843

Shen, F., Dai, G., Lin, G., Zhang, J., Kong, W., and Zeng, H. (2020). Eeg-based emotion recognition using 4d convolutional recurrent neural network. *Cogn. Neurodyn.* 14, 1–14. doi: 10.1007/s11571-020-09634-1

Thammasan, N., Moriyama, K., Fukui, K.-I., and Numao, M. (2017). Familiarity effects in EEG-based emotion recognition. *Brain Inform.* 4, 39–50. doi: 10.1007/s40708-016-0051-5

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). "Attention is all you need," in *Advances in Neural Information Processing Systems, Vol. 30*. Long Beach, CA.

Vuilleumier, P., and Trost, W. (2015). Music and emotions: from enchantment to entrainment. *Ann. N. Y. Acad. Sci.* 1337, 212–222. doi: 10.1111/nyas.12676

Wang, W., Xie, E., Li, X., Fan, D.-P., Song, K., Liang, D., et al. (2021). "Pyramid vision transformer: a versatile backbone for dense prediction without convolutions," in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (Montreal Canada), 568–578. doi: 10.1109/ICCV48922.2021.00061

Wu, B., Xu, C., Dai, X., Wan, A., Zhang, P., Yan, Z., et al. (2020). Visual transformers: token-based image representation and processing for computer vision. *arXiv preprint arXiv:2006.03677*. doi: 10.48550/arXiv.2006.03677

Xing, X., Li, Z., Xu, T., Shu, L., and Xu, X. (2019). SAE+LSTM: a new framework for emotion recognition from multi-channel EEG. *Front. Neurorobot.* 13:37. doi: 10.3389/fnbot.2019.00037

Yang, H., Han, J., and Min, K. (2019). A multi-column CNN model for emotion recognition from EEG signals. *Sensors* 19:4736. doi: 10.3390/s19214736

Yang, Y., Wu, Q., Qiu, M., Wang, Y., and Chen, X. (2018). "Emotion recognition from multi-channel EEG through parallel convolutional recurrent neural network," in *2018 International Joint Conference on Neural Networks (IJCNN)* (Rio de Janeiro), 1–7. doi: 10.1109/IJCNN.2018.8489331

Zhang, Z., Zhong, S.-h., and Liu, Y. (2022). Ganser: a self-supervised data augmentation framework for EEG-based emotion recognition. *IEEE Trans. Affect. Comput.* abs/2109.03124. doi: 10.1109/TAFFC.2022.3170369

Zheng, W.-L., and Lu, B.-L. (2015). Investigating critical frequency bands and channels for eeg-based emotion recognition with deep neural networks. *IEEE Trans. Auton. Mental Dev.* 7, 162–175. doi: 10.1109/TAMD.2015.2431497

Zheng, W.-L., Zhu, J.-Y., and Lu, B.-L. (2017). Identifying stable patterns over time for emotion recognition from EEG. *IEEE Trans. Affect. Comput.* 10, 417–429. doi: 10.1109/TAFFC.2017.2712143

Zhou, X., Zhuo, J., and Krahenbuhl, P. (2019). "Bottom-up object detection by grouping extreme and center points," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Long Beach, CA), 850–859. doi: 10.1109/CVPR.2019.00094