frontiers | Frontiers in Psychology

TYPE Original Research
PUBLISHED 04 April 2024
DOI 10.3389/fpsyg.2024.1275142

OPEN ACCESS

EDITED BY
Federica Marcolin,
Polytechnic University of Turin, Italy

REVIEWED BY
Hiroki Tanaka,
Nara Institute of Science and Technology (NAIST), Japan
Jiahui Pan,
South China Normal University, China

*CORRESPONDENCE
Yanan Zhou
✉ zhouyanan@bfsu.edu.cn
Hebin Cheng
✉ chenghebin@sdmu.edu.cn

†These authors have contributed equally to this work

RECEIVED 09 August 2023
ACCEPTED 20 March 2024
PUBLISHED 04 April 2024

CITATION
Wang D, Lian J, Cheng H and Zhou Y (2024) Music-evoked emotions classification using vision transformer in EEG signals. *Front. Psychol.* 15:1275142. doi: 10.3389/fpsyg.2024.1275142

Music-evoked emotions classification using vision transformer in EEG signals

Dong Wang[1,2†], Jian Lian[2†], Hebin Cheng[2*] and Yanan Zhou[3*]

[1]School of Information Science and Electrical Engineering, Shandong Jiaotong University, Jinan, China, [2]School of Intelligence Engineering, Shandong Management University, Jinan, China, [3]School of Arts, Beijing Foreign Studies University, Beijing, China

Introduction: The field of electroencephalogram (EEG)-based emotion identification has received significant attention and has been widely utilized in both human-computer interaction and therapeutic settings. The process of manually analyzing electroencephalogram signals is characterized by a significant investment of time and work. While machine learning methods have shown promising results in classifying emotions based on EEG data, the task of extracting distinct characteristics from these signals still poses a considerable difficulty.

Methods: In this study, we provide a unique deep learning model that incorporates an attention mechanism to effectively extract spatial and temporal information from emotion EEG recordings. The purpose of this model is to address the existing gap in the field. The implementation of emotion EEG classification involves the utilization of a global average pooling layer and a fully linked layer, which are employed to leverage the discernible characteristics. In order to assess the effectiveness of the suggested methodology, we initially gathered a dataset of EEG recordings related to music-induced emotions.

Experiments: Subsequently, we ran comparative tests between the state-of-the-art algorithms and the method given in this study, utilizing this proprietary dataset. Furthermore, a publicly accessible dataset was included in the subsequent comparative trials.

Discussion: The experimental findings provide evidence that the suggested methodology outperforms existing approaches in the categorization of emotion EEG signals, both in binary (positive and negative) and ternary (positive, negative, and neutral) scenarios.

KEYWORDS

music-evoked emotion, emotion classification, electroencephalographic, deep learning, transformer

1 Introduction

Emotion is intricately intertwined with all facets of the human experience and action. According to Jerritta et al. (2011), it has an impact on human attitudes and perceptions in both human-human contact and human-computer interaction. In the realm of artistic expression, music holds a paramount position as a means to convey and articulate human emotions. Music has been widely recognized as a means of evoking distinct emotive states, leading to its characterization as the language of emotions (Vuilleumier and Trost, 2015). In their investigations, Ekman (1999) and Gilda et al. (2017) introduced six distinct and quantifiable emotional states, namely happiness, sadness, anger, fear, surprise, and disgust, as the basis for implementing emotion identification.

Over time, other emotional states have been included in this collection, such as neutrality, arousal, and relaxation (Bong et al., 2012; Selvaraj et al., 2013; Goshvarpour et al., 2017; Minhad et al., 2017; Wei et al., 2018; Sheykhivand et al., 2020; Liu et al., 2022). In the context of machine learning, the establishment of distinct states for emotions serves as a significant framework for effectively addressing the challenge of emotion recognition. Numerous algorithms for music emotion identification based on machine learning have been proposed in the literature, with applications spanning composition and psychotherapy (Eerola and Vuoskoski, 2012; Cui et al., 2022).

Typically, a conventional music emotion identification system based on machine learning encompasses the subsequent stages:

- The collection of changes in emotions elicited by music is facilitated via the utilization of physiological information obtained by specialized sensors.
- The physiological samples that have been gathered are subjected to a processing procedure in order to remove any potential artifacts.
- The generation of representation pertaining to emotional states is thereafter accomplished by extracting features from the pre-processed data.
- By utilizing a classifier, it is possible to generate the corresponding category of music emotion for a given sample.

Numerous instruments utilized in the acquisition of physiological signals have been employed for the purpose of emotion recognition. Various physiological signals have been investigated for the purpose of emotion recognition. These include, body movement (Zhang et al., 2021), facial expression (Song, 2021), respiration (Siddiqui et al., 2021), galvanic skin response (Kipli et al., 2022), blood volume pulse (Semerci et al., 2022), skin temperature (Semerci et al., 2022), electromyography (Xu et al., 2023), photoplethysmographic (Cosoli et al., 2021), electrocardiogram (Hasnul et al., 2021), and EEG (Li et al., 2021). The non-invasive nature, affordability, and ability to capture data in real-time have contributed to the extensive utilization of EEG in the field of emotion identification (Alarcao and Fonseca, 2017), with a particular emphasis on music emotion categorization (Lin et al., 2006).

Several studies have introduced different approaches for emotion categorization utilizing EEG in the context of machine learning. For example, the study conducted by Sammler et al. (2007) examined the impact of valence on human emotions by analyzing EEG data and heart rate concurrently. The present study aimed to gather data on positive and negative emotions elicited by EEG signals during the auditory experience of consonant and discordant musical stimuli. Subsequently, the authors of the study (Koelstra et al., 2011) made available a publicly accessible dataset. The study conducted by Balasubramanian et al. (2018) examined the emotional reaction to various types of music using EEG data. The experimental findings have indicated that there is an increase in theta band activity in the frontal midline region when individuals are exposed to their preferred music. Conversely, the beta band would have an increase in activity when exposed to music that is perceived as undesirable. In their study, Ozel et al. (2019) introduced a methodology for emotion identification that involves the analysis of temporal-spectral EEG signals. Hou and Chen (2019) derived a set of 27-dimensional EEG characteristics to represent music-induced emotions, including calmness, pleasure, sadness, and rage. *Recently, Qiu et al. (2022) proposed an integrated framework of multi-modal EEG and functional near infrared spectroscopy to explore the influence of music on brain activity.*

In addition, the utilization of deep learning-based architectures in music emotion categorization has been widely adopted due to the shown effectiveness of deep learning in different domains such as machine vision and natural language processing. In their study, Han et al. (2022) conducted a comprehensive review of the existing literature pertaining to the assessment metrics, algorithms, datasets, and extracted features utilized in the analysis of EEG signals in the context of music emotion detection. In their publication, Nag et al. (2022) introduced the JUMusEmoDB dataset. The music emotion categorization challenge was addressed by the authors through the utilization of Convolutional Neural Network (CNN) based models, namely resnet50, mobilenet, squeezenet, and their own suggested ODE-Net. Eskine (2022) conducted a study examining the impact of music listening on creative cognition, a phenomenon that has been empirically demonstrated to enhance creative cognitive processes. The experimental findings provided evidence that cognitive function exhibited an increase inside the default mode. This was supported by the observed augmentation of spectral frequency power in the beta range throughout the entire brain, as well as in the theta range within the parietal region, and in the gamma range across the entire brain. In their study, Daly (2023) investigated the integration of functional magnetic resonance imaging (fMRI) and EEG techniques to develop an acoustic decoder for the purpose of classifying music emotions. The study employed an EEG-fMRI combined paradigm to capture neural responses during music listening among individuals. In this study, a deep learning model known as the long short-term memory (LSTM) was utilized to extract neural information from EEG signals during music listening. The objective was to rebuild the matching music clips based on this extracted information.

Both machine learning and deep learning techniques have demonstrated promising results in the categorization of music-evoked emotions. Nevertheless, there are a number of constraints associated with these approaches that must be addressed prior to their practical implementation in contexts such as medical diagnosis, namely in the realm of emotion identification. One aspect to consider is that the efficacy of machine learning techniques is heavily dependent on the selection of appropriate features. The task at hand continues to provide an unsolved problem as the extraction and selection of these characteristics from EEG data must be done in a manual manner. In addition, it should be noted that manually-designed features possess subjectivity and susceptibility to errors, perhaps rendering them unsuitable for the specific requirements of music emotion identification. In contrast, deep learning models like as CNNs have the ability to automatically

Abbreviations: EEG, Electroencephalographic; fMRI, functional magnetic resonance imaging; LSTM, long short term; CNN, convolutional neural network; ECG, electrocardiogram; EOG, electro-oculugram; GAP, global average pooling; FC, fully connected; GPU, graphical processing unit; TP, true positive; FN, false negative; TN, true negative; FP, false positive.

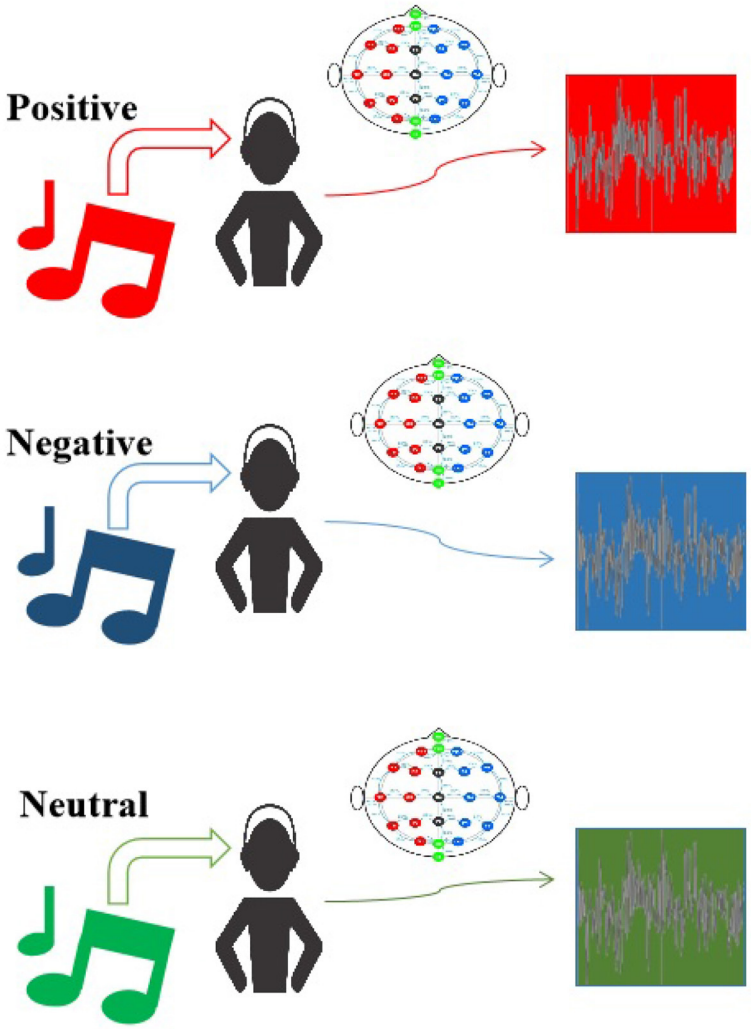

FIGURE 1
The data collecting process for classifying music-evoked emotions using an EEG equipment based on the 10–20 system (Homan et al., 1987).

extract internal representations from EEG inputs. Nevertheless, it is expected that the features derived from CNN models prioritize the consideration of the overall connection between distant EEG signals. This is due to the fact that CNN utilizes a local receptive field approach in the process of extracting features.

The present work introduces a transformer architecture for music-evoked emotion categorization, using a self-attention mechanism. This model incorporates the self-attention mechanism and positional embedding to describe the sequence of channels in EEG data, drawing inspiration from the vision transformer's work (Dosovitskiy et al., 2020). The suggested transformer model has the ability to extract both spatial representations, which correspond to self-attention modules, and temporal representations, which correspond to positional embedding. These representations are derived from multi-channel EEG data acquired from subjects who were listening to music. Furthermore, the transformer model that has been introduced has the capability to extract the relationships that exist among EEG signals across extended distances. In order to assess the efficacy of the suggested methodology, the experiments were conducted using both a publicly accessible dataset (Koelstra et al., 2011) and a privately held dataset. Furthermore, comparative tests were conducted to evaluate the performance of the proposed model in comparison to state-of-the-art algorithms. The experimental findings provide evidence that the suggested methodology exhibits superior performance compared to existing binary and ternary music emotion categorization algorithms. The suggested model has a positive conclusion, indicating its potential value as a tool for classifying music-evoked emotions.

The main contributions of this work can be summarized as follows:

- *This is an early application of the spatial-temporal transformer into the classification of music-evoked emotions.*
- *A novel dataset of music-evoked EEG signals was established.*
- *The proposed approach considers both the spatial connections among a set of EEG channels and the temporal sequence of each individual EEG signal.*

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

ID	Type	Title	Singer	Durating (mm:ss)
1	Positive	Honey	Xinling Wang	03:33
2	Negative	Advanced animals	Wei Dou	04:38
3	Neutral	Reiki meditation	Reiki	06:03
4	Positive	Wu Ha	Weibo Pan	03:46
5	Negative	In case	SHIN	04:24
6	Neutral	Calm dreams	Sleep Tech	04:27
7	Positive	In Spring	Feng Wang	05:10
8	Negative	Cloudy day	Wenwei Mo	04:02
9	Neutral	Let the sun shine	Milk & Sugar	07:02
10	Positive	As broad as the sea and sky	Beyond	03:59
11	Negative	Negative	Black sun empire	05:44
12	Neutral	Illusionary daytime	Shirfine	04:10
13	Positive	Invisible wings	Shaohan Zhang	03:44
14	Negative	Unfortunately, its not you	Jingru Liang	04:45
15	Neutral	Song from a secret garden	Secret garden	03:33

- *The performance of our approach surpassed the state-of-the-art deep learning algorithms on both public and private datasets.*

The subsequent sections of this article are structured as follows: The methodology Section 2 contains information on the acquisition of EEG signals during music listening as well as the details of the presented deep learning model. Section 3 presents a detailed account of the experimental procedures conducted in this investigation, as well as a comprehensive comparison between the existing state-of-the-art methods and the technique proposed in the current study. This research concluded at Section 4.

2 Methodology

This section provides a comprehensive overview of the data gathering process employed in the present investigation. Furthermore, the subsequent sections of the article will present a comprehensive analysis of the suggested transformer model.

2.1 Dataset and pre-processing

The initial step in this study included the creation of a private dataset using multi-channel EEG, which involved the collection of three distinct music-evoked emotions: positive, negative, and neutral. The complete workflow is depicted in Figure 1.

During the course of data gathering, a total of 48 individuals were registered, including 24 females and 24 men. The age range of the participants was between 18 and 25 years, with an average age of 20.6. All individuals involved in the study were enrolled as students at the same institution's campus. Furthermore, it should be noted that the individuals exhibit robust physical and mental well-being. During the course of the project, the research team received advice and supervision from two psychology specialists, one female and one male, who possessed significant expertise in the field.

To ensure the consistency of the data gathering process, the following challenges were proactively addressed. Additionally, all participants were provided with instructions to thoroughly review the handbook and become acquainted with the workflow of EEG signal collecting. It should be noted that the manual has identified and emphasized the entries that are prone to errors, with the intention of facilitating the reader's attention toward the vital operations. Subsequently, the participants were requested to complete a questionnaire pertaining to their personal details. Subsequently, the participants were provided with instructions and guidance from the specialists in order to properly don the EEG electrode caps. Subsequently, the specialists would assess the adequacy of the EEG electrodes' contact and ensure that no detachment has occurred. Furthermore, the participants were instructed by the experts to initiate the signal gathering procedure by hitting the designated buttons. In addition, the EEG collection device utilized in the study was the Biosemi ActiveTwo system. The system employs the international 10–20 system, consisting of 32 channels, notably 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, and O2. Additionally, the sampling rate is set at 512Hz.

During the process of data collection, each participant was provided with instructions to listen to a total of 15 music clips. These clips were categorized into three distinct emotional categories, namely positive, negative, and neutral, with each category consisting of five clips. To note that the categories of these clips were determined by three psychological experts using a majority voting mechanism. The specifics about the music may be found in Table 1. The initial duration of the music clips varies among them. Nevertheless, the participant received a standardized 1-min audio clip for each piece of music. Each participant

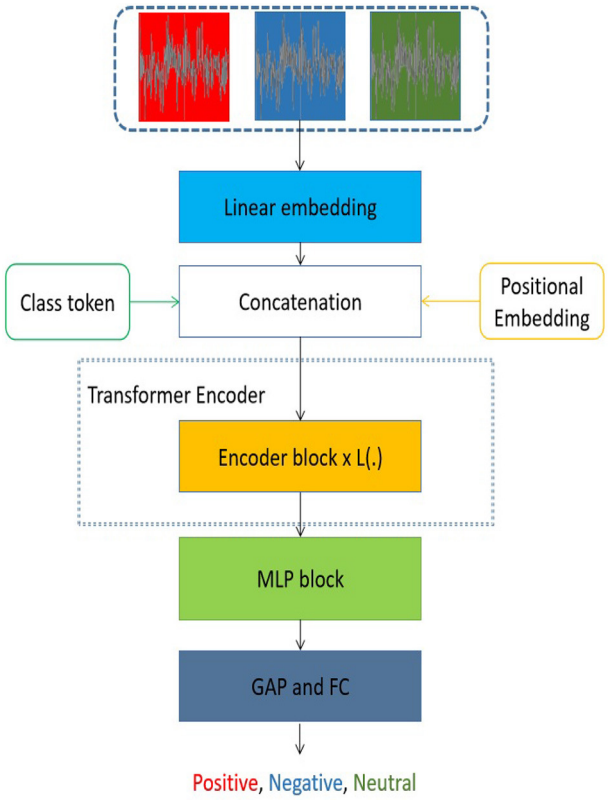

FIGURE 2
The architectural design of the proposed transformer model. The symbol L is used to describe the number of encoder blocks in the proposed model. The proposed model has several variants, with L(.) being either 12, 18, or 24. The abbreviation GAP is used to denote global average pooling, while FC represents fully-connected.

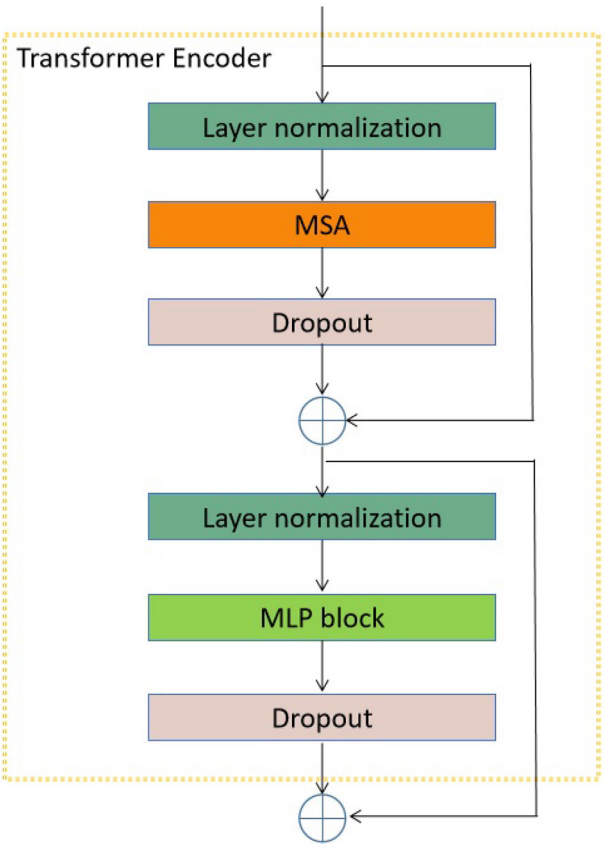

FIGURE 3
The encoder block utilized in the transformer model under consideration. The acronym MSA refers to multi-head self attention.

was instructed to listen to the music clips in a randomized sequence.

The subsequent section presents a comprehensive overview of the data collecting procedure involved in capturing EEG signals related to music-induced emotions.

(1) The participants were provided with instructions to achieve a state of calmness, following which the experts started the marking process to denote the commencement of each EEG recording. The duration of this process is expected to be 5 seconds.
(2) In each 75-second interval, the participants would undergo a 15-second pause to transition between music clips, followed by a 60-second period of actively listening to the music clip. Simultaneously, the experts would provide guidance to the participants on how to minimize superfluous bodily motions.
(3) Following the auditory experience, the individuals were directed by the experimental personnel to assign a value to the musical composition, with positive being denoted as +1, negative as −1, and neutral as 0. The duration of this procedure should not exceed 15 seconds, during which it is utilized for the purpose of transitioning the music.
(4) The participants proceeded with the auditory experience by sequentially engaging with the subsequent musical excerpt until the entirety of the 12 excerpts had been presented.

So as to guarantee the optimal state of the participants, the collection of music-evoked emotion EEG samples was limited to the time periods of 9 a.m. to 11 a.m. and 3 p.m. to 5 p.m. In order to mitigate interference from many sources such as heart rate, breathing, electrocardiogram (ECG), and electro-scalogram (EOG), the participants were given instructions to cover their eyes while the recording procedures were being conducted.

The dataset contains a total of 43,200 ($48 \times 15 \times 60 = 43,200$) seconds of EEG signals, with each second including 32 channels. Furthermore, the initial samples were partitioned into the epochs of 1 second duration, each consisting of 60, 000 data points. To note that there were still overlapping epochs in the samples since the trivial errors are difficult to avoid due to the human reaction times. Given the absence of any imbalance issue within the dataset, it can be observed that each category of music emotion EEG signals is comprised of an equal number of samples, specifically 20,000 epochs. Hence, in the context of binary classification, namely distinguishing between positive and negative classes, the proposed model was trained using a dataset including 40,000 epochs as input samples. In contrast, in the context of the ternary classification job, the entirety of the 60,000 epochs were utilized as the input. It should

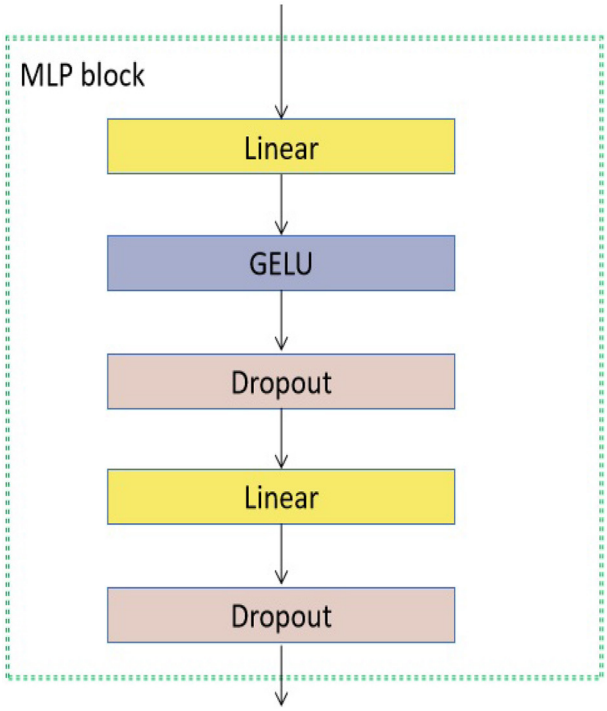

FIGURE 4
The MLP block used in the proposed transformer model. GELU denotes the activation function (Lee, 2023).

be noted that the presence of overlapping epochs has the potential to somewhat mitigate over-fitting.

In the pre-processing phase, the acquired EEG signals were subjected to a Notch filter (Serra et al., 2017) in order to remove the 50 Hz components originating from the power supply. Subsequently, a first-order low-pass filter with a frequency range of 0.5 to 45 Hz was utilized. Subsequently, the electroencephalography (EEG) data underwent a normalization process resulting in a range of values between 0 and 1.

2.2 The proposed transformer architecture

The transformer model presented in Figure 2 draws inspiration from the architecture of the vision transformer (Dosovitskiy et al., 2020). The suggested transformer model comprises three main components: (1) a linear embedding layer, (2) an encoder block, and (3) a multiple-layer perception (MLP) block. Initially, the linear embedding unit was utilized to turn a sequence of EEG data into a fixed-length input for the suggested transformer model. The flattened embedding includes the class token of the music emotion for each series of EEG data. In addition, the linear embedding is constructed by including the positional embedding, which encodes the sequential order of an individual EEG signal inside a sequence of EEG signals. It should be noted that every input sequence of EEG data pertains to the identical category of emotion elicited by music. Furthermore, the pivotal self-attention module (Fan et al., 2021; Liu et al., 2021; Wang et al., 2021), which aims to reveal the

TABLE 2 The proposed transformer model exhibits binary and ternary classification outcomes (average values and standard deviations).

Number of classes	Accuracy (%)	Sensitivity (%)	Specificity (%)
Binary	96.85 (1.73)	95.17 (1.68)	95.69 (2.01)
Ternary	95.74 (2.32)	94.32 (1.97)	95.25 (1.69)

connections among distant EEG data, is located within the encoder block. In order to create a cohesive encoder module, it is necessary for the encoder block to be iteratively repeated. In addition to the self-attention layer included in each encoder block, there are many additional sorts of layers, namely layer normalization, dropout, and MLP block. The generation of representations for music emotion EEG signals may be achieved by the utilization of stacked transformer encoder blocks. Ultimately, the use of the MLP block was implemented to get the classification result by integrating a global average pooling (GAP) layer and a fully connected (FC) layer, commonly referred to as a linear layer. The transformer model under consideration has the potential to significantly expand the scope of receptive fields in comparison to designs based on CNNs. Additionally, the recovered representation from the multi-channel EEG data encompasses both local information pertaining to a series of signals and the global association between signals that are far apart.

In the proposed transformer model, the input sequences consist of individual EEG signals, each spanning a duration of 1 second and including 30 channels. Subsequently, the EEG signal sequence was flattened and transformed into a vector. In addition, it should be noted that the encoder block is iterated a varying number of times (12, 18, or 24) across different versions of the proposed transformer model. Furthermore, the structural composition of this encoder block is illustrated in Figure 3.

As seen in Figure 3, the encoder block has many components, namely layer normalization, MSA, dropout, and MLP block. The study did not include a comprehensive examination of the MSA unit due to its extensive coverage in existing studies (Vaswani et al., 2017; Dosovitskiy et al., 2020). The unit consisting of H heads was employed to assess the similarity between a query and its associated keys based on the assigned weight for each value (Vaswani et al., 2017). Furthermore, the utilization of the Layer normalizing module is employed to calculate the mean and variance required for normalizing from the entirety of the inputs to the neurons within a layer throughout a singular training instance (Ba et al., 2016). In this study, the dropout layer (Choe and Shim, 2019) is utilized as a regularization technique to mitigate the risk of overfitting. The architecture of the MLP block is seen in Figure 4.

The proposed technique allows for the formulation of the process of music emotion categorization in Equation 1–4:

$$z_0 = [x_{class}; x_p^1 E; x_p^2 E; \ldots; x_p^N] + E_{position}, \quad (1)$$

where the variable z_0 represents the output of the linear embedding layer. In this context, $N = 30$ represents the number of channels

used as input. The variables x_{class} and $E_{position}$ refer to the class token and positional embedding, respectively.

$$z_l^{'} = MSA(LN(z_{l-1})) + z_{l-1}, \quad (2)$$

$$z_l = MLP(LN(z_l^{'})) + z_l^{'}, \quad (3)$$

$$y = LN(z_L^0), \quad (4)$$

where the layer normalization unit is denoted as $LN(.)$, where z_l represents the output of layer l, and y represents the output classification outcome.

3 Experimental results

3.1 Implementation details

The transformer model described in this study was constructed using the PyTorch framework (Paszke et al., 2019). The computational resources employed for the implementation were four NVidia RTX 3080 Graphical Processing Units (GPUs) with a total of 64 GB RAM. The best parameters of the proposed network were discovered using a trial and error technique. The learning rate is configured to be 0.004, accompanied by a weight decay of 0.05. Subsequently, a 10-fold cross-validation procedure was employed to assess the resilience of the suggested methodology. Initially, the input EEG data were partitioned into ten equitably sized groups. During each iteration, one out of the 10 groups was designated as the testing set, while the remaining nine groups were utilized as the

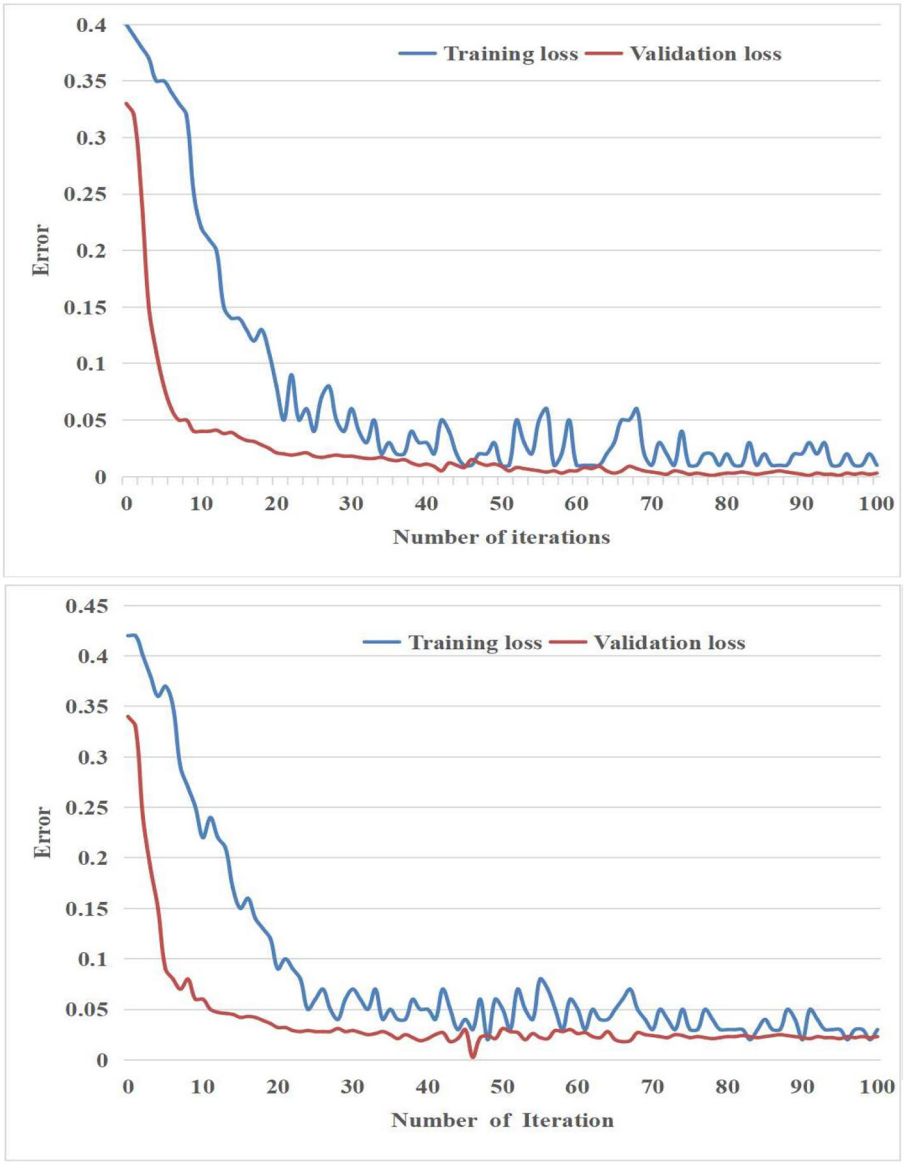

FIGURE 5
The suggested model's inaccuracy in **(Top)** binary classification and **(Bottom)** ternary classification.

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

Method	Accuracy (%)	Sensitivity (%)	Specificity (%)
U-Net (Ronneberger et al., 2015)	88.56	88.71	89.05
Mask R-CNN (He et al., 2017)	87.43	86.39	86.56
ExtremeNet (Zhou et al., 2019)	89.49	89.87	88.51
TensorMask (Chen et al., 2019)	90.56	90.18	91.27
4D-CRNN (Shen et al., 2020)	92.57	92.32	93.08
FBCCNN (Pan and Zheng, 2021)	92.53	91.68	91.24
MTCNN (Rudakov, 2021)	93.02	93.55	94.17
SSGMC (Kan et al., 2022)	94.82	94.18	94.23
MViT (Fan et al., 2021)	90.42	91.39	90.72
PVT (Wang et al., 2021)	92.27	91.15	92.01
PiT (Heo et al., 2021)	93.53	92.85	93.78
Swin Transformer (Liu et al., 2021)	95.32	94.64	94.37
GPViT (Yang et al., 2022)	96.38	94.88	95.27
The proposed approach	96.85	95.17	95.69

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

Method	Accuracy (%)	Sensitivity (%)	Specificity (%)
U-Net (Ronneberger et al., 2015)	85.52	83.86	84.20
Mask R-CNN (He et al., 2017)	85.24	84.21	85.41
ExtremeNet (Zhou et al., 2019)	86.28	83.17	84.53
TensorMask (Chen et al., 2019)	88.32	86.51	87.02
4D-CRNN (Shen et al., 2020)	91.57	92.24	91.89
FBCCNN (Pan and Zheng, 2021)	91.27	91.38	92.24
MTCNN (Rudakov, 2021)	92.21	92.19	93.43
SSGMC (Kan et al., 2022)	92.18	91.57	94.28
MViT (Fan et al., 2021)	92.15	91.93	92.78
PVT (Wang et al., 2021)	91.23	90.46	91.37
PiT (Heo et al., 2021)	92.43	92.14	91.62
Swin transformer (Liu et al., 2021)	92.57	91.38	93.27
GPViT (Yang et al., 2022)	93.14	92.25	93.18
The proposed approach	95.74	94.32	95.25

training set. Hence, the mean result of 10 iterations was utilized as the ultimate output.

Furthermore, the assessment measures utilized in the experiments involved sensitivity, specificity, and accuracy. The mathematical formulation of these metrics is elucidated in in Equations 5–7.

$$Sensitivity = \frac{TP}{TP + FN}, \tag{5}$$

$$Specificity = \frac{TN}{TN + FP}, \tag{6}$$

$$Accuracy = \frac{TP + TN}{TP + FN + TN + FP}, \tag{7}$$

where TP, FN, TN, and FP represent the terms true positive, false negative, true negative, and false positive, respectively.

3.2 Outcome of the proposed approach

Table 2 presents a summary of the average values and standard deviations (SD) obtained from the proposed method in the binary classification task, specifically in terms of average accuracy, sensitivity, and specificity. The average accuracy was found to be 96.85%, while the sensitivity and specificity were measured at 95.17% and 95.69% respectively. Furthermore, in the ternary categorization, the outcome rates were recorded as 95.74%, 94.32%, and 95.25%.

Furthermore, the loss curves of the suggested methodology throughout both the training and validation procedures were illustrated in Figure 5 It should be noted that the results presented in Figure 5 only include the initial 100 iterations of both the training and validation processes.

3.3 Comparison experiments between the state-of-the-arts and the proposed approach

To assess the efficacy of our suggested technique for music-evoked emotion categorization, we conducted comparative tests between our work and the state-of-the-art algorithms. Tables 2–4 present a comparative analysis of the current state-of-the-art deep learning models and our proposed approach. The proposed

TABLE 5 Comparison between the state-of-the-arts and ours on DEAP dataset (Koelstra et al., 2011).

Method	Detail	Accuracy	
		Valence	Arousal
3DCNN (Shawky et al., 2018)	CNN	88.52	89.36
CNN-LSTM (Yang et al., 2018)	LSTM	92.43	89.51
SAE-LSTM (Xing et al., 2019)	LSTM	86.32	81.27
Multi-column CNN (Yang et al., 2019)	CNN	93.81	94.15
4D-CRNN (Shen et al., 2020)	CRNN	95.34	93.62
FGCCNN (Pan and Zheng, 2021)	CNN	91.72	90.28
MTCNN (Rudakov, 2021)	CNN	95.34	95.49
GANSER (Zhang et al., 2022)	GAN	94.18	93.58
SSGMC (Kan et al., 2022)	Contrastive learning	96.12	94.62
The proposed approach	Transformer	97.41	97.02

TABLE 6 The impact of *H* and *L* on the performance of the proposed model in binary classification.

Model	Number of heads (H)	Number of layers (L)	Accuracy (%)
M_4_4	4	4	90.08
M_4_8	4	8	90.37
M_8_4	8	4	91.15
M_8_8	8	8	91.63
M_8_12	8	12	93.35
M_12_12	12	12	93.21
M_16_12	16	12	94.16
M_8_18	8	18	94.58
M_12_18	12	18	95.39
M_16_18	16	18	95.65
M_8_24	8	24	96.28
M_12_24	12	24	96.12
M_16_24	16	24	**96.53**

The bold value represents the best performance of accuracy with 16 heads and 24 layers.

methodology demonstrated superior performance compared to the current leading method. *To note that we did not take the traditional machine learning models (Qiu et al., 2022) into the comparison since they usually relied on manually-designed features.* The comparison experiments included the following models: U-Net (Ronneberger et al., 2015), Mask R-CNN (He et al., 2017), ExtremeNet (Zhou et al., 2019), TensorMask (Chen et al., 2019), 4D-CRNN (Shen et al., 2020), FBCCNN (Pan and Zheng, 2021), MTCNN (Rudakov, 2021), SSGMC (Kan et al., 2022) for the CNN-based models, and MViT (Fan et al., 2021), PVT (Wang et al., 2021), PiT (Heo et al., 2021), Swin Transformer (Liu et al., 2021), and GPViT (Yang et al., 2022) for the transformer-based models.

In order to conduct a comprehensive evaluation of the proposed approach, we proceeded to assess its performance alongside several state-of-the-art algorithms (Shawky et al., 2018; Yang et al., 2018; Xing et al., 2019; Yang et al., 2019; Shen et al., 2020; Pan and Zheng, 2021; Rudakov, 2021; Kan et al., 2022; Zhang et al., 2022) using the publicly accessible DEAP dataset (Koelstra et al., 2011). The results of this evaluation are presented in Table 5.

4 Discussion

Based on the empirical findings, it can be concluded that this approach exhibits greater efficacy compared to the existing state-of-the-art algorithms. It is worth mentioning that the comparative trials encompassed both CNN-based and transformer-based models. In contrast to CNN-based models, the suggested model has the capability to extract global connections between long-range multi-channels in EEG data, in addition to the local information already present in the EEG signals. In contrast to transformer-based models (He et al., 2017; Chen et al., 2019; Zhou et al., 2019; Wu et al., 2020; Fan et al., 2021; Heo et al., 2021; Wang et al., 2021), the proposed approach has been specifically optimized to accommodate the unique characteristics of multi-channel EEG signals. For instance, the linear embedding layer of the proposed approach has been tailored to effectively align with the structural properties of multi-channel EEG signals. Furthermore, the outcomes shown in the ablation research also exhibited the efficacy of self-attention modules and encoder blocks.

4.1 Ablation study

As demonstrated in Table 6, the optimal configuration of the primary hyper-parameters was determined through comparison experiments. These experiments involved testing different combinations of the number of heads (H) in the MSA module and the number of transformer encoder layers (L) on a dataset that was manually collected and constituted 50% of the total dataset. The trials solely included binary music emotion categorization in order to streamline the ablation study procedure.

Therefore, the suggested model exhibits an ideal configuration while utilizing 16 heads ($H = 16$) and 24 layers ($L = 24$).

4.2 Limitations and future research

In addition, this study possesses certain limitations in addition to its contributions. The tests solely focused on the binary and ternary classification problems. In order to enhance the evaluation of the proposed approach, it is recommended to integrate the categorization of other types of emotions and employ a multi-label classification methodology. Meanwhile, this study adopted an

offline learning strategy since the vision transformer-based models suffering from high resource occupancy. In addition, this study did not take cross-subject emotion recognition (He et al., 2021; Pan et al., 2023) into consideration, which may affect the applicability and universality of this study.

In subsequent investigations, further electroencephalography (EEG) data pertaining to the elicitation of emotions through music will be gathered. Furthermore, the suggested methodology holds potential for the identification of emotions across a wide range of applications.

5 Conclusion

The present work introduces a transformer model as a means of classifying music-evoked emotions. The model under consideration consists of three distinct phases, namely linear embedding, transformer encoder, and MLP layer. The purpose of the first phase is to generate flattened input features for the proposed model. These features are aimed to extract both local and global correlations between the multi-channel EEG data. Additionally, the MLP blocks aim to enhance the classification outcome. This study presents an initial implementation of a vision transformer-based model for the purpose of music emotion identification.

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 humans were approved by the Shandong Management University's Human Research Ethics Committee. The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

Author contributions

DW: Writing – review & editing, Formal analysis, Validation. JL: Writing – original draft, Supervision, Project administration, Methodology, Funding acquisition, Formal analysis, Data curation, Conceptualization. HC: Writing – original draft, Validation, Investigation, Formal analysis, Data curation. YZ: Writing – original draft, Validation, Investigation, Data curation.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. The present study received support from the Natural Science Foundation of Shandong Province (Grant No. ZR2020MF133), the Key Laboratory of Public Safety Management Technology of Scientific Research and Innovation Platform in Shandong Universities during the 13th Five-Year Plan Period, the Collaborative Innovation Center of "Internet plus intelligent manufacturing" of Shandong Universities, and the 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

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

Ba, J., Kiros, J. R., and Hinton, G. E. (2016). Layer normalization. *ArXiv, abs/1607.06450*.

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

Bong, S. Z., Murugappan, M., and Yaacob, S. B. (2012). "Analysis of electrocardiogram (ECG) signals for human emotional stress classification," in *IEEE International Conference on Robotics and Automation*. doi: 10.1007/978-3-642-35197-6_22

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*, 2061–2069. doi: 10.1109/ICCV.2019.00215

Choe, J., and Shim, H. (2019). "Attention-based dropout layer for weakly supervised object localization," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2214–2223. doi: 10.1109/CVPR.2019.00232

Cosoli, G., Poli, A., Scalise, L., and Spinsante, S. (2021). "Heart rate variability analysis with wearable devices: influence of artifact correction method on classification accuracy for emotion recognition," in *2021 IEEE International Instrumentation and Measurement Technology Conference (I2MTC)*, 1–6. doi: 10.1109/I2MTC50364.2021.9459828

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. Neuroinf.* 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 16x16 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. *Music Percept.* 30, 307–340. doi: 10.1525/mp.2012.30.3.307

Ekman, P. (1999). *Basic Emotions, chapter 3*. London: John Wiley Sons, Ltd. 45–60. doi: 10.1002/0470013494.ch3

Eskine, K. (2022). Evaluating the three-network theory of creativity: Effects of music listening on resting state EEG. *Psychol. Music* 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*, 6824–6835. doi: 10.1109/ICCV48922.2021.00675

Gilda, S., Zafar, H., Soni, C., and Waghurdekar, K. (2017). "Smart music player integrating facial emotion recognition and music mood recommendation," in *2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*, 154–158. doi: 10.1109/WiSPNET.2017.8299738

Goshvarpour, A., Abbasi, A., and Goshvarpour, A. (2017). An accurate emotion recognition system using ecg and gsr signals and matching pursuit method. *Biomed. J.* 40, 355–368. doi: 10.1016/j.bj.2017.11.001

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

Hasnul, M. A., Aziz, N. A. A., Alelyani, S., Mohana, M., and Aziz, A. A. (2021). Electrocardiogram-based emotion recognition systems and their applications in healthcare–a review. *Sensors* 21:5015. doi: 10.3390/s21155015

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

He, Z., Zhong, Y., and Pan, J. (2021). An adversarial discriminative temporal convolutional network for EEG-based cross-domain emotion recognition. *Comput. Biol. Med.* 141:105048. doi: 10.1016/j.compbiomed.2021.105048

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*, 11936–11945. doi: 10.1109/ICCV48922.2021.01172

Homan, R., Herman, J. H., and Purdy, P. (1987). Cerebral location of international 10–20 system electrode placement. *Electroencephalogr Clin. Neurophysiol.* 66, 376–382. doi: 10.1016/0013-4694(87)90206-9

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

Jerritta, S., Murugappan, M., Nagarajan, R., and Wan, K. (2011). "Physiological signals based human emotion recognition: a review," in *2011 IEEE 7th International Colloquium on Signal Processing and its Applications*, 410–415. doi: 10.1109/CSPA.2011.5759912

Kan, H., Yu, J., Huang, J., Liu, Z., and Zhou, H. (2022). Self-supervised group meiosis contrastive learning for EEG-based emotion recognition. *Appl. Intel.* 53, 27207–27225. doi: 10.1007/s10489-023-04971-0

Kipli, K., Latip, A. A. A., Lias, K. B., Bateni, N., Yusoff, S. M., Suud, J. B., et al. (2022). "Evaluation of galvanic skin response (GSR) signals features for emotion recognition," in *International Conference on Applied Intelligence and Informatics* (Cham: Springer Nature Switzerland), 260–274. doi: 10.1007/978-3-031-24801-6_19

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

Lee, M. (2023). Gelu activation function in deep learning: a comprehensive mathematical analysis and performance. *ArXiv, abs/2305.12073*.

Li, Y., Zheng, W., Zong, Y., Cui, Z., Zhang, T., and Zhou, X. (2021). A bi-hemisphere domain adversarial neural network model for EEG emotion recognition. *IEEE Trans. Affect. Comput.* 12, 494–504. doi: 10.1109/TAFFC.2018.2885474

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* (IEEE), 2765–2768.

Liu, J., Sun, L., Huang, M., Xu, Y., and Li, R. (2022). Enhancing emotion recognition using region-specific electroencephalogram data and dynamic functional connectivity. *Front. Neurosci.* 16:884475. doi: 10.3389/fnins.2022.884475

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*, 10012–10022. doi: 10.1109/ICCV48922.2021.00986

Minhad, K. N., Ali, S. H. M., and Reaz, M. B. I. (2017). A design framework for human emotion recognition using electrocardiogram and skin conductance response signals. *J. Eng. Sci. Technol.* 12, 3102–3119. doi: 10.1587/transinf.2017EDP7067

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. *Physica 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. Sig. Proc. 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. Mathem. Methods Med.* 2021:2520394. doi: 10.1155/2021/2520394

Pan, J., Liang, R., He, Z., Li, J., Liang, Y., Zhou, X., et al. (2023). St-scgnn: a spatio-temporal self-constructing graph neural network for cross-subject EEG-based emotion recognition and consciousness detection. *IEEE J. Biomed. Health Inf.* 28, 777–788. doi: 10.1109/JBHI.2023.3335854

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,* 32.

Qiu, L., Zhong, Y., Xie, Q., He, Z., Wang, X., Chen, Y., et al. (2022). Multi-modal integration of EEG-fnirs for characterization of brain activity evoked by preferred music. *Front. Neurorob.* 16:823435. doi: 10.3389/fnbot.2022.823435

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, Germany, October 5–9, 2015, Proceedings, Part III 18* (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 USA* (IEEE). 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

Selvaraj, J., Murugappan, M., Wan, K., and Yaacob, S. B. (2013). Classification of emotional states from electrocardiogram signals: a non-linear approach based on hurst. *BioMed. Eng. OnLine* 12:44. doi: 10.1186/1475-925X-12-44

Semerci, Y. C., Akgün, G., Toprak, E., and Barkana, D. E. (2022). "A comparative analysis of deep learning methods for emotion recognition using physiological signals for robot-based intervention studies," in *2022 Medical Technologies Congress (TIPTEKNO)*, 1–4. doi: 10.1109/TIPTEKNO56568.2022.9960200

Serra, H., Oliveira, J. P., and Paulino, N. F. (2017). "A 50 hz sc notch filter for iot applications," in *2017 IEEE International Symposium on Circuits and Systems (ISCAS)*, 1–4. doi: 10.1109/ISCAS.2017.8050904

Shawky, E., El-Khoribi, R., Shoman, M., and Wahby Shalaby, M. (2018). EEG-based emotion recognition using 3D convolutional neural networks. *Int. J. Adv. Comput. Sci. Applic.* 9:843. 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. Neurodyna.* 14, 1–14. doi: 10.1007/s11571-020-09634-1

Sheykhivand, S., Mousavi, Z., Rezaii, T. Y., and Farzamnia, A. (2020). Recognizing emotions evoked by music using CNN-LSTM networks on EEG signals. *IEEE Access* 8, 139332–139345. doi: 10.1109/ACCESS.2020.3011882

Siddiqui, H. U. R., Shahzad, H. F., Saleem, A. A., Khakwani, A. B. K., Rustam, F., Lee, E., et al. (2021). Respiration based non-invasive approach for emotion recognition using impulse radio ultra wide band radar and machine learning. *Sensors (Basel, Switzerland)* 21:8336. doi: 10.3390/s21248336

Song, Z. (2021). Facial expression emotion recognition model integrating philosophy and machine learning theory. *Front. Psychol.* 12:759485. doi: 10.3389/fpsyg.2021.759485

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,* 30.

Vuilleumier, P., and Trost, W. (2015). Music and emotions: from enchantment to entrainment. *Ann. NY 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*, 568–578. doi: 10.1109/ICCV48922.2021.00061

Wei, W., Jia, Q., Feng, Y., and Chen, G. (2018). Emotion recognition based on weighted fusion strategy of multichannel physiological signals. *Comput. Intell. Neurosci.* 2018:5296523. doi: 10.1155/2018/5296523

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*.

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

Xu, M., Cheng, J., Li, C., Liu, Y., and Chen, X. (2023). Spatio-temporal deep forest for emotion recognition based on facial electromyography signals. *Comput. Biol. Med.* 156:106689. doi: 10.1016/j.compbiomed.2023.106689

Yang, C., Xu, J., Mello, S. D., Crowley, E. J., and Wang, X. (2022). Gpvit: a high resolution non-hierarchical vision transformer with group propagation. *ArXiv, abs/2212.06795*.

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)*, 1–7. doi: 10.1109/IJCNN.2018.8489331

Zhang, H., Yi, P., Liu, R., and Zhou, D. (2021). "Emotion recognition from body movements with as-LSTM," in *2021 IEEE 7th International Conference on Virtual Reality (ICVR)*, 26–32. doi: 10.1109/ICVR51878.2021.9483833

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.* 14, 2048–2063. doi: 10.1109/TAFFC.2022.3170369

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*, 850–859. doi: 10.1109/CVPR.2019.00094