frontiers | Frontiers in Neuroscience

TYPE Original Research
PUBLISHED 24 March 2023
DOI 10.3389/fnins.2023.1148855

OPEN ACCESS

EDITED BY
Xin Huang,
Renmin Hospital of Wuhan University, China

REVIEWED BY
Hongzhi Kuai,
Maebashi Institute of Technology, Japan
Yaofei Xie,
Xuzhou Medical University, China

*CORRESPONDENCE
Wenfeng Duan
✉ ndyfy02345@ncu.edu.cn

SPECIALTY SECTION
This article was submitted to Visual Neuroscience, a section of the journal Frontiers in Neuroscience

RECEIVED 20 January 2023
ACCEPTED 06 March 2023
PUBLISHED 24 March 2023

CITATION
Wan Z, Li M, Liu S, Huang J, Tan H and Duan W (2023) EEGformer: A transformer–based brain activity classification method using EEG signal.
Front. Neurosci. 17:1148855.
doi: 10.3389/fnins.2023.1148855

EEGformer: A transformer–based brain activity classification method using EEG signal

Zhijiang Wan[1,2,3], Manyu Li[2], Shichang Liu[4], Jiajin Huang[5], Hai Tan[6] and Wenfeng Duan[1]*

[1]The First Affiliated Hospital of Nanchang University, Nanchang University, Nanchang, Jiangxi, China, [2]School of Information Engineering, Nanchang University, Nanchang, Jiangxi, China, [3]Industrial Institute of Artificial Intelligence, Nanchang University, Nanchang, Jiangxi, China, [4]School of Computer Science, Shaanxi Normal University, Xi'an, Shaanxi, China, [5]Faculty of Information Technology, Beijing University of Technology, Beijing, China, [6]School of Computer Science, Nanjing Audit University, Nanjing, Jiangsu, China

Background: The effective analysis methods for steady-state visual evoked potential (SSVEP) signals are critical in supporting an early diagnosis of glaucoma. Most efforts focused on adopting existing techniques to the SSVEPs-based brain–computer interface (BCI) task rather than proposing new ones specifically suited to the domain.

Method: Given that electroencephalogram (EEG) signals possess temporal, regional, and synchronous characteristics of brain activity, we proposed a transformer–based EEG analysis model known as EEGformer to capture the EEG characteristics in a unified manner. We adopted a one-dimensional convolution neural network (1DCNN) to automatically extract EEG-channel-wise features. The output was fed into the EEGformer, which is sequentially constructed using three components: regional, synchronous, and temporal transformers. In addition to using a large benchmark database (BETA) toward SSVEP-BCI application to validate model performance, we compared the EEGformer to current state-of-the-art deep learning models using two EEG datasets, which are obtained from our previous study: SJTU emotion EEG dataset (SEED) and a depressive EEG database (DepEEG).

Results: The experimental results show that the EEGformer achieves the best classification performance across the three EEG datasets, indicating that the rationality of our model architecture and learning EEG characteristics in a unified manner can improve model classification performance.

Conclusion: EEGformer generalizes well to different EEG datasets, demonstrating our approach can be potentially suitable for providing accurate brain activity classification and being used in different application scenarios, such as SSVEP-based early glaucoma diagnosis, emotion recognition and depression discrimination.

KEYWORDS

brain activity classification, SSVEPs, EEGformer, EEG characteristics, deep learning

1. Introduction

Glaucoma is known as a "silent thief of sight," meaning that patients do not notice the health condition of their visual function until vision loss and even blindness occur (Abdull et al., 2016). According to the world health organization, the number of people with glaucoma worldwide in 2020 is 76 million, and the patient number would be increased to 95.4 million in 2030. As the population ages, the number with this condition will also increase substantially (Guedes, 2021). Glaucoma causes irreversible optic nerve vision damage. It is crucial to provide accurate early screening to diagnose patients in their early stages so that they can receive appropriate early treatment. Steady-state visual evoked potentials (SSVEPs), which refer to a stimulus-locked oscillatory response to periodic visual stimulation commonly exerted in the visual pathway of humans, can be used to evaluate the functional abnormality of the visual pathway that is essential for the complete transmission of visual information (Zhou et al., 2020). SSVEPs are always measured using electroencephalogram (EEG) measurement and have been widely used in the study of brain–computer interface (BCI). Because peripheral vision loss is a key diagnostic sign of glaucoma, patients cannot be evoked by certain repetitive stimuli with a constant frequency from vision loss regions (Khok et al., 2020). Therefore, stimuli with the corresponding frequency are not detected by the primary visual cortex. Thus, the SSVEPs-based BCI applications can be used in the early diagnosis of visual function detection for patients with glaucoma.

The effective analysis method for SSVEPs is critical in the accurate early diagnosis of glaucoma. SSVEPs are EEG activity with a spatial-spectral-temporal (SST) pattern. It is easy to understand that SSVEP signals, such as the EEG signal measured over time, could be analyzed using time series analysis methods. Brain functional connectivity (BFC) can be used to capture spatial patterns from multiple brain regions by analyzing the correlations between brain activities detected from different regions. The spectral pattern extraction method is the most popular method for analyzing the frequency characteristics of EEG signals. For instance, power spectra density–based analysis (PSDA) is a commonly used frequency detection method that can classify various harmonic frequencies from EEG signals (Zhang et al., 2020). In addition, canonical correlation analysis (CCA) (Zhuang et al., 2020) and other similar algorithms, such as multivariate synchronization index (MSI) (Qin et al., 2021) and correlated component analysis (COCA) (Zhang et al., 2019), are effective frequency detection algorithms based on the multivariate statistical analysis method. Although SST pattern extraction algorithms have demonstrated satisfactory results, most patterns or features extracted from raw EEG data require a manually predefined algorithm based on expert knowledge. The procedure of learning handcrafted features for SSVEP signals is not flexible and might limit the performance of these systems in brain activity analysis tasks.

In recent years, deep learning (DL) methods have achieved excellent performance in processing EEG-based brain activity analysis tasks (Li Z. et al., 2022; Schielke and Krekelberg, 2022). Currently, the mainstream technologies of using DL to process SSVEP signal could be divided into two aspects: convolutional neural network (CNN) based methods and transformer-based methods. For the CNN-based methods, Li et al. (2020) propose a CNN-based nonlinear model, i.e. convolutional correlation analysis (Conv-CA), to transform multiple channel EEGs into a single EEG signal and use a correlation layer to calculate correlation coefficients between the transformed single EEG signal and reference signals. Guney et al. (2021) propose a deep neural network architecture for identifying the target frequency of harmonics. Waytowich et al. (2018) design a compact convolutional neural network (Compact-CNN) for high-accuracy decoding of SSVEPs signal. For the transformer-based methods, Du et al. (2022) propose a transformer-based approach for the EEG person identification task that extracts features in the temporal and spatial domains using a self-attention mechanism. Chen et al. (2022) propose SSVEPformer, which is the first application of the transformer to the classification of SSVEP. Li X. et al. (2022) propose a temporal-frequency fusion transformer (TFF-Former) for zero-training decoding across two BCI tasks. The aforementioned studies demonstrate the competitive model performance of DL methods in performing SSVEPs-based BCI tasks. However, most existing DL efforts focused on applying existing techniques to the SSVEPs-based BCI task rather than proposing new ones specifically suited to the domain. Standard well-known network architectures are designed for data collected in natural scenes and do not consider the peculiarities of the SSVEP signals. Therefore, further research is required to understand how these architectures can be optimized for EEG-based brain activity data.

The main question is what is the specificity of the SSVEP signal analysis domain and how to use machine learning methods (particularly DL methods) to deal with the signal characteristics. Because the SSVEP signal is EEG-based brain activity, we can answer the question by analyzing the EEG characteristics in the brain activity analysis domain. Specifically, EEG characteristics are reflected in three aspects: temporal, regional, and synchronous characteristics. The temporal characteristics (e.g., mean duration, coverage, and frequency of occurrence) are easily traceable in standard EEG data and provide numerous sampling points in a short time (Zhang et al., 2021), thereby providing an efficient way to investigate trial-by-trial fluctuations of functional spontaneous activity. The regional characteristics refer to different brain regions that are linked to distinct EEG bands (Nentwich et al., 2020). The synchronous characteristics refer to the synchronous brain activity pattern over a functional network including several brain regions with similar spatial orientations (Raut et al., 2021). Traditionally, brain response to a flickering visual stimulation has been considered steady-state, in which the elicited effect is believed to be unchanging in time. In fact, the SSVEPs belongs to a signal with non-stationary nature, which indicates dynamical patterns and complex synchronization between EEG channels can be used to further understand brain mechanisms in cognitive and clinical neuroscience. For instance, Ibáñez-Soria et al. explored the dynamical character of the SSVEP response by proposing a novel non-stationary methodology for SSVEP detection, and found dynamical detection methodologies significantly improves classification in some stimulation frequencies (Ibáñez-Soria et al., 2019). Tsoneva et al. (2021) studied the mechanisms behind SSVEPs generation and propagation in time and space. They concluded that the SSVEP spatial properties appear sensitive to input frequency with higher stimulation frequencies showing a faster propagation speed. Thus, we hypothesize that a machine learning method that can capture the EEG characteristics in a unified manner can suit the SSVEPs-based BCI domain and improve the model performance in EEG-based brain activity analysis tasks.

In this study, we propose a transformer–based EEG analysis model known as the EEGformer (Vaswani et al., 2017) to capture the EEG characteristics in the SSVEPs-based BCI task. The EEGformer is an end-to-end DL model, processing SSVEP signals from the EEG to the prediction of the target frequency. The component modules of the EEG former are depicted as follows:

(1) Depth-wise convolution-based one-dimensional convolution neural network (1DCNN). The depth-wise convolution-based 1DCNN is first used to process the raw EEG input. Assuming the raw data is collected from C EEG channels, there are M depth-wise convolutional filters for generating M feature maps. Each convolutional filter is responsible for shifting across the raw data in an EEG-channel-wise manner and extracting convolutional features from the raw data of each EEG channel to form a feature map. Unlike other techniques that manually extract temporal or spectrum features based on the time course of the EEG signal, we use the depth-wise convolutional filter to extract the EEG features in a completely data–driven manner. Because the feature map is generated by the same depth-wise convolutional filter, each row of the feature map shares the same convolutional property. Follow-up convolutional layers are allocated with several depth-wise convolutional filters to enrich the convolutional features and deepen the 1DCNN network. A three-dimensional (3D) feature matrix is used to represent the output of the 1DCNN network. The x, y, and z dimensions of the 3D feature matrix represent temporal, spatial, and convolutional features, respectively.

(2) EEGformer encoder. This component module consists of three sub-modules: temporal, synchronous, and regional transformers, which are used in learning the temporal, synchronous, and regional characteristics, respectively. The core strategy of learning EEG characteristics by our model mainly include two steps: input tokens that serve as the basic elements of learning the temporal, synchronous, and regional characteristics are sliced from the 3D feature matrix along the temporal, convolutional, and spatial dimension, respectively. And then, self-attention mechanism is employed to measure the relationships between pairs of input tokens and give tokens more contextual information, yielding more powerful features for representing the EEG characteristics. The three components could be performed in a sequential computing order, allowing the encoder to learn the EEG characteristics in a unified manner.

(3) EEGformer decoder. This module contains three convolutional layers and one fully connected (FC) layer. The output of the last FC layer is fed to a softmax function which produces a distribution over several category labels. The categorical cross entropy combined with regularization was used as the loss function for training the entire EEGformer pipeline. The EEGformer decoder is used to deal with specific tasks, such as target frequency identification, emotion recognition, and depression discrimination. In addition to using a large benchmark database (BETA) (Liu et al., 2020) to validate the performance of the SSVEP-BCI application, we validate the model performance on two additional EEG datasets, one for emotion analysis using EEG signals [SJTU emotion EEG dataset (SEED)] (Duan et al., 2013; Zheng and Lu, 2015) and the other for a depressive EEG database (DepEEG) (Wan et al., 2020) obtained from our previous study, to support our hypothesis that highlights the significance of learning EEG characteristics in a unified manner for EEG-related data analysis tasks.

The main contributions of this study are as follows: (1) current mainstream DL models have superior ability in processing data collected in natural scenes and might not adept at dealing with SSVEP signals. To achieve a DL model that can be applied to the specificity of the SSVEP signal analysis domain and obtain better model performance in SSVEPs-based frequency recognition task, we propose a transformer–based EEG analysis model known as the EEGformer to capture the EEG characteristics in a unified manner. (2) To obtain a flexible method for addressing the SSVEPs-based frequency recognition and avoid the model performance limited by manual feature extraction, we adopt 1DCNN to automatically extract EEG-channel-wise features and fed them into the EEGformer. This operation transforms our method into a complete data–driven manner for mapping raw EEG signals into task decisions. (3) To ensure the effectiveness and generalization ability of the proposed model, we validate the performance of the EEGformer on three datasets for three different EEG-based data analysis tasks: target frequency identification, emotion recognition, and depression discrimination.

2. Materials and methods

2.1. Dataset preparation

Table 1 shows some detailed information about the three datasets (BETA, SEED, and DepEEG) that we used as benchmarks to validate the effectiveness of this study. The participants' column in the table describes how many subjects joined in the corresponding data collection. The experiment per participant (EPP) column shows how many experiments were performed by each participant. The trails per experiment (TPE) column shows how many trails are executed in one experiment. The channel number (CHN) column shows the CHN of the EEG dataset. The sampling rate (SR) column shows the down-sampling rate of the EEG signal. The time length per trail (TLPT) column shows the time length of a single trail in seconds. The labels column shows the categorical emotion labels for the classification task and emotional intensity for the regression task. Specifically, for the target frequency identification task, we classified 40 categories of harmonic frequencies and the frequency range is 8–15.8 HZ with 0.2 HZ intervals. For the emotion recognition task, we used arousal, valence, and dominance rating scores as the dataset labels. For the depression discrimination task, we classified EEG samples from depressive or normal control.

2.2. Pipeline of EEGformer–based brain activity analysis

Figure 1 shows the pipeline of EEGformer–based brain activity analysis. The core modules of the pipeline include 1DCNN, EEGformer encoder, and decoder. The input of the 1DCNN is an

TABLE 1 Detail information on the three datasets.

Dataset	Participants	EPP	TPE	CHN	SR (HZ)	Labels	TLPT
BETA	70 healthy subjects	4	40	64	250	40 harmonics, e.g., $f_j \in \{8,8.2,\ldots,15.8\}$	2/3 s
SEED	15 healthy subjects	3	15	62	200	Positive, neutral, negative	305 s
DepEEG	12 healthy subjects and 23 depressives	1	1	6	500	Depressive, normal control	≥480 s

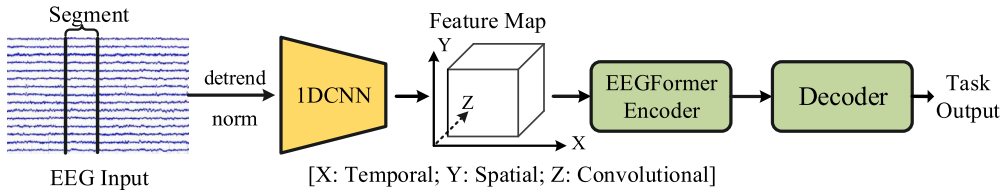

FIGURE 1
Pipeline of EEGformer for different tasks of brain activity analysis.

EEG segment represented using a two-dimensional (2D) matrix of size $S \times L$, where S represents the number of EEG channels, and L represents the segment length. The EEG segment is de-trend and normalized before being fed into the 1DCNN module, and the normalized EEG segment is represented by $x \in R^{S \times L}$. The 1DCNN adopts multiple depth-wise convolutions to extract EEG-channel-wise features and generate 3D feature maps. It shifts across the data along the EEG channel dimension for each depth-wise convolution and generates a 2D feature matrix of size $S \times L_f$, where L_f is the length of the extracted feature vector. The output of the 1DCNN module is a 3D feature matrix of size $S \times C \times L_e$, where C is the number of depth-wise convolutional kernels used in the last layer of the 1DCNN module, L_e is the features length outputted by the last layer of the 1DCNN module. More specifically, the 1DCNN is comprised of three depth-wise convolutional layers. Hence, we have the processing $x \rightarrow z_1 \rightarrow z_2 \rightarrow z_3$, where z_1, z_2, and z_3 denote the outputs of the three layers. The size of the depth-wise convolutional filters used in the three layers is 1×10, valid padding mode is applied in the three layers and the stride of the filters is set to 1. The number of the depth-wise convolutional filter used in the three layers is set to 120, ensuring sufficient frequency features for learning the regional and synchronous characteristics. We used a 3D coordinate system to depict the axis meaning of the 3D feature matrix. The X, Y, and Z axes represent the temporal, spatial, and convolutional feature information contained in the 3D feature matrix, respectively. The output of the 1DCNN module is fed into the EEGformer encoder for encoding the EEG characteristics (regional, temporal, and synchronous characteristics) in a unified manner. The decoder is responsible for decoding the EEG characteristics and inferencing the results according to the specific task.

2.3. EEGformer encoder

The EEGformer encoder is used to provide a uniform feature refinement for the regional, temporal, and synchronous characteristics contained in the output of the 1DCNN module. Figure 2 illustrates the EEGformer architecture and shows that the EEGformer encoder uses a serial structure to sequentially refine the EEG characteristics. The temporal, regional, and synchronous characteristics are refined using temporal, regional, and synchronous transformers, respectively. The outputs of the 1DCNN are defined as $z_3 \in R^{S \times C \times L_e}$ and are represented using black circles in the green rectangle box.

The specific computing procedures of each transformer module are depicted as follows:

2.3.1. Regional transformer module

The input of the regional transformer module is represented by $z_3 \in R^{C \times L_e \times S}$. The 3D matrix z_3 is first segmented into S 2D submatrices along the spatial dimension. Each submatrix is represented by $X_i^{reg} \in R^{C \times L_e}$ (i = 1,2,3,...,S). The input of the regional transformer module is represented by S black circles in the green rectangle box and each circle represents a submatrix. The vector $X_{(i,c)}^{reg} \in R^{L_e}$ is sequentially taken out from the X_i^{reg} along the convolutional feature dimension and fed into the linear mapping module. According to the terminology used in the vision of transformer (ViT) studies, we defined the vector $X_{(i,c)}^{reg}$ as a patch of the regional transformer module. Each $X_{(i,c)}^{reg}$ is represented by a tiny yellow block in the Figure 2. The $X_{(i,c)}^{reg}$ is linearly mapped into a latent vector $z_{(i,c)}^{(reg,0)} \in R^D$ using a learnable matrix $M \in R^{D \times L_e}$:

$$z_{(i,c)}^{(reg,0)} = MX_{(i,c)}^{reg} + e_{(i,c)}^{pos}, \quad (1)$$

where $e_{(i,c)}^{pos} \in R^D$ denotes a positional embedding added to encode the position for each convolutional feature changing over time. The regional transformer module also consists of $K \geq 1$ encoding blocks, each block contains two layers: a multi-head self-attention layer and a position-wise fully connected feed-forward network. The resulting $z_{(i,c)}^{(reg,0)}$ is defined as a token representing the inputs of each block, and the $z_{(0,0)}^{(reg,0)}$ indicates the classification token. The l-th block produces an encoded representation $z_{(i,c)}^{(reg,l)}$ for each token in the input sequence by incorporating the attention scores. Specifically, at each block l, three core vectors, including $q_{(i,c)}^{(l,a)}$, $k_{(i,c)}^{(l,a)}$, and $v_{(i,c)}^{(l,a)}$ are computed from the representation $z_{(i,c)}^{(reg,l-1)}$ encoded by the preceding layer:

$$q_{(i,c)}^{(l,a)} = W_Q^{(l,a)} LN(z_{(i,c)}^{(reg,l-1)}) \in R^{D_h}, \quad (2)$$

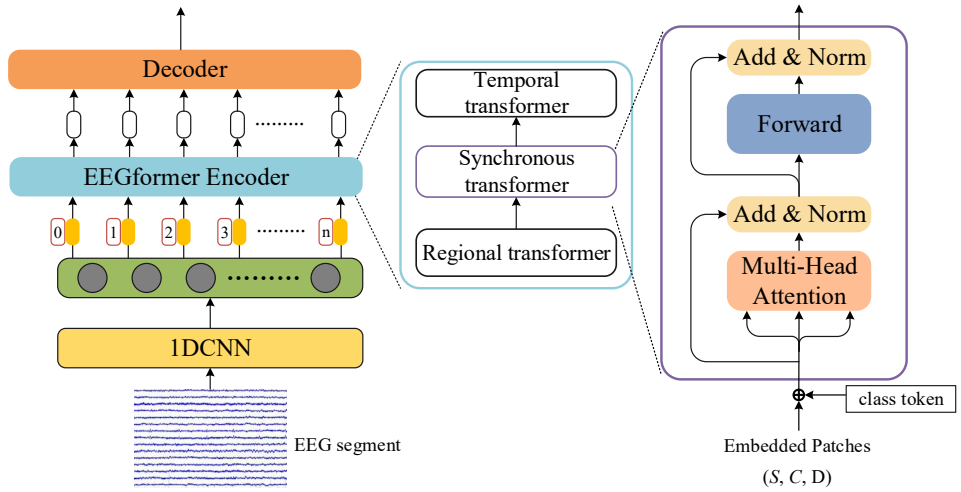

FIGURE 2
Illustration of the EEGformer architecture.

$$k_{(i,c)}^{(l,a)} = W_K^{(l,a)} LN(z_{(i,c)}^{(reg,l-1)}) \in R^{D_h}, \quad (3)$$

$$v_{(i,c)}^{(l,a)} = W_V^{(l,a)} LN(z_{(i,c)}^{(reg,l-1)}) \in R^{D_h}, \quad (4)$$

where $W_Q^{(l,a)}$, $W_K^{(l,a)}$, and $W_V^{(l,a)}$ are the matrixes of query, key, and value in the regional transformer module, respectively. LN() denotes the LayerNorm operation, and a $\in$ {1, 2, 3, . . ., A} is an index over the multi-head self-attention units. A is the number of units in a block. D_h is the quotient computed by D/A and denotes the dimension number of three vectors. The regional self-attention (RSA) scores for $z_{(i,c)}^{(reg,l-1)}$ in the a-th multi-head self-attention unit is given as follows:

$$\alpha_{(i,c)}^{(l,a)reg} = \sigma\left(\frac{q_{(i,c)}^{(l,a)}}{\sqrt{D_h}} \cdot \left[k_{(0,0)}^{(l,a)} \left\{k_{(i,c)}^{(l,a)}\right\}_{c=1,\ldots,C}\right]\right) \in R^C, \quad (5)$$

where σ denotes the softmax activation function, and the symbol · denotes the dot product for computing the similarity between the query and key vectors. $k_{(i,c)}^{(l,a)}$ and $q_{(i,c)}^{(l,a)}$ represent the corresponding key and query vectors, respectively. The equation shows that the RSA scores are merely computed over convolutional features of single brain region. That is, the RSA can calculate the contribution of a changing mono-electrode convolutional feature to the final model decision at a specific EEG channel. An intermediate vector $s_{(i,c)}^{(l,a)}$ for encoding $z_{(i,c)}^{(reg,l-1)}$ is given as follows:

$$s_{(i,c)}^{(l,a)} = \alpha_{(i,0)}^{(l,a)} v_{(i,0)}^{(l,a)} + \sum_{j=1}^{C} \alpha_{(i,j)}^{(l,a)} v_{(i,j)}^{(l,a)} \in R^{D_h}. \quad (6)$$

The encoded feature $z_{(i)}^{(reg,l)} \in R^{C \times D}$ by the l-th block is computed by first concatenating the intermediate vectors from all heads, and the vector concatenation is projected by matrix $W_O \in R^{D \times L}$, where L is equal to A $= D_h$. $z_{(i)}^{'(reg,l)}$ is the residual connection result of the projection of the intermediate vectors and the $z_{(i)}^{(reg,l-1)}$ encoded by the preceding block. Finally, the $z_{(i)}^{'(reg,l)}$ normalized by LN() is passed through a multilayer perceptron (MLP) using the residual connection. The output of the regional transformer is represented by $z_4 \in R^{S \times C \times D}$.

2.3.2. Synchronous transformer module

The input of the synchronous transformer module is represented by $z_4 \in R^{S \times L_e \times C}$. The 3D matrix z_4 is first segmented into C 2D submatrices along the convolutional feature dimension. Each submatrix is represented by $X_i^{syn} \in R^{S \times D}$ (i = 1,2,3,. . .,C). The vector $X_{(i,s)}^{syn} \in R^D$ is sequentially taken out from the X_i^{syn} along the spatial dimension and fed into the linear mapping module. The $X_{(i,s)}^{syn}$ is defined as a patch and is linearly mapped into a latent vector $z_{(i,s)}^{(syn,0)} \in R^D$ using a learnable matrix M $\in R^{D \times D}$:

$$z_{(i,s)}^{(syn,0)} = MX_{(i,s)}^{syn} + e_{(i,s)}^{pos}, \quad (7)$$

where $e_{(i,s)}^{pos} \in R^D$ denotes a positional embedding added to encode the spatial position for each EEG channel changing over time. The synchronous transformer also consists of $K \geq 1$ encoding blocks, and each block contains two layers: a multi-head self-attention layer and a position-wise fully connected feed-forward network. The resulting $z_{(i,s)}^{(syn,0)}$ is defined as a token representing the inputs of each block, and the $z_{(0,0)}^{(syn,0)}$ indicates the classification token. The l-th block produces an encoded representation $z_{(i,s)}^{(syn,l)}$ for each token in the input sequence by incorporating the attention scores. Specifically, at each block l, three core vectors, including $q_{(i,s)}^{(l,a)}$, $k_{(i,s)}^{(l,a)}$, and $v_{(i,s)}^{(l,a)}$ are computed from the representation $z_{(i,s)}^{(syn,l-1)}$ encoded by the preceding layer:

$$q_{(i,s)}^{(l,a)} = W_Q^{'(l,a)} LN(z_{(i,s)}^{(syn,l-1)}) \in R^{D_h}, \quad (8)$$

$$k_{(i,s)}^{(l,a)} = W_K^{'(l,a)} LN(z_{(i,s)}^{(syn,l-1)}) \in R^{D_h}, \quad (9)$$

$$v_{(i,s)}^{(l,a)} = W_V^{'(l,a)} LN(z_{(i,s)}^{(syn,l-1)}) \in R^{D_h}, \quad (10)$$

where $W_Q^{'(l,a)}$, $W_K^{'(l,a)}$, and $W_V^{'(l,a)}$ are the matrixes of query, key, and value in the synchronous transformer module, respectively.

Synchronous e self-attention (SSA) scores for $z_{(i,s)}^{(syn,l-1)}$ in the a-th multi-head self-attention unit are given as follows:

$$\alpha_{(i,s)}^{(l,a)syn} = \sigma\left(\frac{q_{(i,s)}^{(l,a)}}{\sqrt{D_h}} \cdot \left[k_{(0,0)}^{(l,a)}\left\{k_{(i,s)}^{(l,a)}\right\}_{s=1,\ldots,S}\right]\right) \in R^S, \quad (11)$$

where $k_{(i,s)}^{(l,a)}$ and $q_{(i,s)}^{(l,a)}$ denote the corresponding key and query vectors, respectively. The equation shows that the SSA scores are merely computed over the feature map extracted by the same depth-wise convolution. The SSA can calculate the contribution of convolution features changing over time to the final model decision at a specific EEG channel. An intermediate vector $s_{(i,s)}^{(l,a)}$ for encoding $z_{(i,s)}^{(syn,l-1)}$ is given as follows:

$$s_{(i,s)}^{(l,a)} = \alpha_{(i,0)}^{(l,a)} v_{(i,0)}^{(l,a)} + \sum_{j=1}^{C} \alpha_{(i,j)}^{(l,a)} v_{(i,j)}^{(l,a)} \in R^{D_h}. \quad (12)$$

The encoded feature $z_{(i)}^{(syn,l)} \in R^{S \times D}$ by the l-th block is computed by first concatenating the intermediate vectors from all heads, and the vector concatenation is projected by matrix $W_O \in R^{D \times L}$. $z_{(i)}^{'(syn,l)}$ is the residual connection result of the projection of the intermediate vectors and the $z_{(i)}^{(syn,l-1)}$ encoded by the preceding block. Finally, the $z_{(i)}^{'(syn,l)}$ normalized by LN() is passed through a multilayer perceptron (MLP) using the residual connection. The output of the synchronous transformer is represented by $z_5 \in R^{C \times S \times D}$.

2.3.3. Temporal transformer module

The input of the temporal transformer module is $z_5 \in R^{C \times S \times D}$. To avoid huge computational complexity, we compress the original temporal dimensionality D of z_5 into dimensionality M. That is, the 3D matrix z_5 is first segmented and then averaged into M 2D submatrices along the temporal dimension. Each submatrix is represented by $X_i^{temp} \in R^{S \times C}$ (i = 1,2,3,...,M) and the M submatrices are concatenated to form $X^{temp} \in R^{M \times S \times C}$. Each submatrix X_i^{temp} is flattened into a vector $X_i^{'temp} \in R^{L1}$, where $L1$ is equal to $S \times C$. The $X_i^{'temp}$ is defined as a patch and is linearly mapped into a latent vector $z_{(i)}^{(temp,0)} \in R^D$ using a learnable matrix $M \in R^{D \times L}$:

$$z_{(i)}^{(temp,0)} = MX_{(i)}^{'temp} + e_{(i)}^{pos}, \quad (13)$$

where $e_{(i)}^{pos} \in R^D$ denotes a positional embedding added to encode the temporal position for each EEG channel changing over the features extracted by different depth-wise convolutional kernels. The module consists of $K \geq 1$ encoding blocks, each block contains two layers: a multi-head self-attention layer and a position-wise fully connected feed-forward network. The resulting $z_{(i)}^{(temp,0)}$ is defined as a token representing the inputs of each block, and the $z_{(0)}^{(temp,0)}$ indicates the classification token. The l-th block produces an encoded representation $z_{(i)}^{(temp,l)}$ for each token in the input sequence by incorporating the attention scores. Specifically, at each block l, three core vectors, including $q_{(i)}^{(l,a)}$, $k_{(i)}^{(l,a)}$, and $v_{(i)}^{(l,a)}$ are computed from the representation $z_{(i)}^{(temp,l-1)}$ encoded by the preceding layer:

$$q_{(i)}^{(l,a)} = W_Q^{''(l,a)} LN(z_{(i)}^{(temp,l-1)}) \in R^{D_h}, \quad (14)$$

$$k_{(i)}^{(l,a)} = W_K^{''(l,a)} LN(z_{(i)}^{(temp,l-1)}) \in R^{D_h}, \quad (15)$$

$$v_{(i)}^{(l,a)} = W_V^{''(l,a)} LN(z_{(i)}^{(temp,l-1)}) \in R^{D_h}, \quad (16)$$

where $W_Q^{''(l,a)}$, $W_K^{''(l,a)}$, and $W_V^{''(l,a)}$ are the matrixes of query, key, and value in the temporal transformer, respectively. The temporal self-attention (TSA) score for $z_{(i,s)}^{(T,l-1)}$ in the a-th multi-head self-attention unit is given as follows:

$$\alpha_{(i)}^{(l,a)temp} = \sigma\left(\frac{q_{(i)}^{(l,a)}}{\sqrt{D_h}} \cdot \left[k_{(0)}^{(l,a)}\left\{k_{(i)}^{(l,a)}\right\}_{i=1,\ldots,M}\right]\right) \in R^M. \quad (17)$$

The equation shows that the TSA scores are merely computed over the temporal dimension. The TSA can calculate the contribution of multiple electrode features changing over different convolutional features to the final model decision at a specific time. An intermediate vector $s_{(i)}^{(l,a)}$ for encoding $z^{(temp,l-1)}$ is given as follows:

$$s_{(i)}^{(l,a)} = \alpha_{(i,0)}^{(l,a)} v_{(i,0)}^{(l,a)} + \sum_{j=1}^{M} \alpha_{(i,j)}^{(l,a)} v_{(i,j)}^{(l,a)} \in R^{D_h}. \quad (18)$$

The encoded feature $z^{(temp,l)} \in R^{M \times L1}$ by the l-th block is computed by first concatenating the intermediate vectors from all heads, and the vector concatenation is projected by matrix $W_O \in R^{L1 \times L}$. $z^{'(temp,l)}$ is the residual connection result of the projection of the intermediate vectors and the $z^{(temp,l-1)}$ encoded by the preceding block. Finally, the $z^{'(temp,l)}$ normalized by LN() is passed through a multilayer perceptron (MLP) using the residual connection. The output of the temporal transformer is represented by $O \in R^{M \times L1}$.

2.4. EEGformer decoder

The EEGformer is used to extract the temporal, regional, and synchronous characteristics in a unified manner, as well as to deal with various EEG-based brain activity analysis tasks. Unlike the original transformer decoder, which uses a multi-head self-attention mechanism to decode the feature output of the corresponding encoder, we designed a convolution neural network (CNN) to perform the corresponding task. The CNN contains three convolutional layers and one fully connected layer. Specifically, the output $O \in R^{M \times L1}$ of the EEGformer encoder is reshaped to $X \in R^{S \times C \times M}$, where M is the dimensional length of the encoded temporal feature. The first layer of our EEGformer decoder (with the weights $w_1 \in R^{C \times 1}$) linearly combined different convolutional features for normalization across the convolutional dimension. Thus, the output data shape of the first layer is $X_1 \in R^{S \times M}$. The motivation to convolve C feature maps along the convolutional dimension of X into one is to allow the network to make data–driven decisions about the contribution of different convolutional features to the final model decision. The second layer of our CNN was responsible for combining information across spatial dimensions of X and extracting the entire information while discarding redundancy or noninformative variations. To this end, our CNN convolved X along the spatial dimension using the weights $w_2 \in R^{S \times N}$ and returns the plane $X_2 \in R^{M \times N}$,

where N denotes the number of convolutional filters used in the second layer. The third layer halved the dimension and reduced the parameter complexity using the weights $w_3 \in R^{(M/2) \times N}$ to produce the output plane $X_3 \in R^{(M/2) \times N}$. The fourth layer of our CNN is a fully connected layer that produced classification results for the brain activity analysis task. The corresponding equation of the loss function is given as follows:

$$Loss = \frac{1}{D_n} \sum_{i=1}^{D_n} -\log\left(p_i\left(y_i\right)\right) + \lambda |w| \tag{19}$$

where D_n is the number of data samples in the training dataset, p_i and y_i are the prediction results produced by the model and the corresponding ground truth label for the i-th data sample, respectively, and λ is the constant of the L1 regularization.

3. Experiment results

3.1. Experimental setup

For generating the input of the EEGformer and other comparison models, we first extract the raw EEG data of each trial of the three datasets to form data samples and assign the corresponding label to each data sample. Further, we apply a sliding window with the step of $ratio \times SR$ (i.e., SR) on each data sample and generate the final input samples in a non-overlapping manner. The data shape of each input sample is $ratio \times SR \times N_c$, and the N_c denotes the number of EEG channels (i.e., 64). The equation for representing the relationship between segment length T and the total number of input samples N is given as follows:

$$\mathrm{N} = \frac{N_{sub} \times EPP \times TPE \times TLPT}{ratio}, \tag{20}$$

where N_{sub} denotes how many subjects joined in the corresponding data collection experiment. Taking the data splitting method for BETA dataset as an example, we remove the EEG data collected during the gaze shifting of 0.5 s guided by a visual cue and an offset of 0.5 s followed by the visual stimulation. The final BETA dataset consists of 11,200 trials and 40 categories. For the first 15 participants and the remaining 55 participants in the BETA dataset, the time length of the flickering visual stimulation in each trial is 2 and 3 s, respectively. When the number of data points of each input sample is 100, meaning the *ratio* is set to 0.4 and the *SR* is equal to 250 Hz, and the time length of each input sample is 0.4 s, the total number of input samples of the BETA dataset for training and testing models is 78,000. Under the same setting, the total number of input samples of the SEED and DepEEG dataset for training and testing models is 514,687 and 42,000.

The state-of-the-art DL models, which have performed well in previous studies, were tested on the three datasets to compare their model performance with ours. In our comparison, we followed the same test procedures for all these methods. The EEGformer and other comparison baselines were trained with a batch size of 64 and Adam optimizer with a learning rate of 0.001. In each transformer module, the number of encoding blocks is equal to three. The models were trained using an early-stop training strategy. Note that all training hypermeters were optimized using the testing data. Pytorch was used to implement these models, which were trained on an NVIDIA Tesla A100 GPU. As mentioned above, we tested our model on three datasets (BETA, SEED, and DepEEG). Fivefold cross-validation was applied to separate the dataset, and the average classification accuracy (ACC) rate, sensitivity (SEN), and specificity (SPE) and the corresponding standard deviation (SD) of them were used as model performance metrics. For multi-category classification, the accuracy rate, which means how many data samples are corrected and labeled out of all the data samples, is calculated as the sum of true positive and true negative divided by the sum of true positive, false positive, false negative, and true negative. The above metrics are calculated using the following formula:

$$\mathrm{ACC} = (\mathrm{TP}+\mathrm{TN})/(\mathrm{TP}+\mathrm{FP}+\mathrm{FN}+\mathrm{TN}) \tag{21}$$

$$\mathrm{SEN} = \mathrm{TP}/(\mathrm{TP}+\mathrm{FN}), \tag{22}$$

$$\mathrm{SPE} = \mathrm{TN}/(\mathrm{TN}+\mathrm{FP}), \tag{23}$$

where TP denotes true positives, TN denotes true negatives, FP denotes false positives, and FN denotes false negatives.

3.2. Comparison baselines

To show the effective model performance of EEGformer, we compared several commonly used DL methods in other studies of EEG-based data analysis tasks, which were target frequency identification, emotion recognition, and depression discrimination. The comparison models are described as follows:

(1) EEGNet (Lawhern et al., 2018). It is a Compact-CNN for EEG-based BCIs. The network starts with a temporal convolution operation to learn frequency filters. The operation is made up of F_1 convolutional filters, and each size equals $1 \times N$, where N represents the length of the convolutional filter. It used $D \times F_1$ depth-wise convolutional filters to learn frequency-specific spatial filters and the size of each filter is $C \times 1$. The separable convolution followed by point-wise convolution was used to learn the summary for each feature map and optimally combine them. The network architecture shows that EEGNet considers temporal and spatial information of EEG signals.

(2) Conv-CCA (Waytowich et al., 2018). It is designed for SSVEPs-based target frequency identification and can be used in other EEG-based classification tasks. Unlike pure DL models, the Conv-CCA uses a signal-CNN with three-layers to transform multiple channel EEGs ($N_s \times N_c \times 1$) into a single $\bar{x}$ with a shape of $N_s \times 1 \times 1$, where N_s and N_c are the numbers of sampling points and channels, respectively. Another reference CNN with two-layers was used to transform the reference signal ($N_s \times N_f \times N_c$) into a 2D signal $\bar{Y}$ with a shape of $N_s \times N_f$, where N_f is the number of target frequencies. Correlation analysis was used to calculate the coefficients of $\bar{x}$ and each $\bar{Y}^n$ for all $n \in [1, N_f]$. A dense layer with N_f units and a softmax activation function was used as the final layer for classification.

(3) 4DCRNN (Shen et al., 2020). It is a DL model known as a four-dimensional (4D) convolutional recurrent neural network that extracts and fuses frequency, spatial and temporal information from raw EEG data to improve model performance of emotion recognition. It is not an end-to-end model for BCI tasks because it requires the Butterworth filter to decompose frequency bands and manually extract differential entropy features from each frequency band. The model input is represented as a 4D structure $X \in R^{h \times w \times d \times 2T}$, where h and w are the height and width of the 2D brain topographical map, respectively, d denotes the number of frequency bands and T denotes the length of the signal segment. CNN was used to extract the frequency and spatial information from each temporal segment of an EEG sample, and long short-term memory (LSTM) was adopted to extract temporal information from CNN outputs.

(4) EmotionNet (Wang et al., 2018). Instead of using 2D convolution filters to extract features from input data, EmotionNet used a 3D convolution filter to learn spatial and temporal features from raw EEG data. The first two layers and the third layer of the model used a 3D convolution filter to learn spatiotemporal and fuse spatial features, respectively. The fourth and fifth layers of the model performed temporal feature extraction using a 2D convolutional filter. The sixth layer of the model is a fully connected layer for dense predictions.

(5) PCRNN (Yang et al., 2018). The model is an end-to-end DL model known as a parallel convolutional recurrent neural network for EEG-based emotion classification tasks. It also takes 3D shape ($X \in R^{h \times w \times T}$) of raw EEG data as model input. CNN model was first used to learn spatial feature maps from each 2D map, and the LSTM was used to extract temporal features from the CNN outputs. Note that the CNN and LSTM were organized by a parallel structure to extract the spatial and temporal features from the model input. The outputs of the parallel structure were integrated to classify emotions.

3.3. Ablation studies

3.3.1. Effect of the EEGformer decoder constructed by different transformer combination

We conducted an ablation study to show the effectiveness of the EEGformer by constructing the encoder with different combinations of temporal, synchronous, and regional transformers. The classification results (ACC, SPE, SEN, and their corresponding SDs) on the three EEG datasets using different transformer module combinations to construct the EEGformer encoder are shown in Table 2. The table shows that the EEGformer encoder constructed by the combinations of the three transformers achieves the best classification results. For BETA dataset, the average sensitivity, specificity, and accuracy are 69.86, 75.86, and 70.15%, respectively. For SEED dataset, the average sensitivity, specificity, and accuracy are 89.14, 92.75, and 91.58%, respectively. For DepEEG dataset, the average sensitivity, specificity, and accuracy are 77.83, 70.95, and 72.19%, respectively. The result supports our hypothesis that a machine learning method can capture EEG characteristics in a unified manner that can suit the EEG-based brain activity analysis tasks.

The table also demonstrates that the EEGformer that contains a synchronous transformer achieves better model performance than the EEGformer without a synchronous transformer. For instance, the EEGformer constructed using a single synchronous transformer outperforms the EEGformer constructed using the other two types of single transformers, with better accuracy of 57.29, 80.12, and 60.12% on BETA, SEED, and DepEEG, respectively. The EEGformer constructed using a transformer pair consisting of a synchronous transformer outperforms the EEGformer constructed using the transformer pair without a synchronous transformer, with better accuracy on the BETA, SEED, and DepEEG datasets. The results indicate the significance of learning spatial distribution characteristics of EEG activity generated by multiple brain regions for the task of SSVEPs-based frequency discrimination. In addition, the EEGformer constructed using synchronous transformer and regional transformer outperforms the EEGformer constructed using other transformer pairs, with better classification results on SEED and DepEEG dataset. On the one hand, the result demonstrates that the convolutional features could represent regional and spatial characteristics of EEH signal well. On the other hand, the result indicates that the integration of the synchronous and regional EEG characteristics improves discrimination ability of our model.

3.3.2. Effect of using 1DCNN or not to construct the EEGformer pipeline

The model performance affected by using 1DCNN or not is validated to show the rationality of using a 1D depth-wise convolutional filter to learn regional characteristics in a completely data–driven manner. Figure 3 compares the results of using 1DCNN or not constructing the EEGformer pipeline. The figure shows that using a 1D depth-wise convolutional filter to learn regional characteristics is beneficial for improving model performance in EEG-based classification tasks.

3.3.3. Effect of EEG channel number on the model performance

Table 3 reports the classification results (ACC, SPE, SEN, and their corresponding SDs) of our model with varying number of EEG channel. The EEG CHN and the corresponding name of brain regions are illustrated as follows: 3 (O1, Oz, and O2), 6 (O1, Oz, O2, POz, PO3, and PO4), 9 (O1, Oz, O2, Pz, PO3, PO5, PO4, PO6, and POz), 32 channels (all channels from occipital, parietal, central-parietal regions and C3, C1, Cz, C2, C4, and FCz) as well as all 64 channels. From the table, we can know that as the EEG CHN increases, the classification results of the EEGformer show an upward trend. This result indicates that although the EEG channels that are placed over the occipital and parietal regions provide perhaps the most informative SSVEP signals, other channels are informative as well. The result also illustrates the data mining ability of our model, which can learn representational features from complex data structure.

3.4. Comparison studies

Leave-one-subject-out (LOSO) cross-validation method is utilized to compare the model performance between EEGformer

TABLE 2 Classification results (ACC, SPE, SEN, and their corresponding SDs) on the three EEG datasets by using different transformer module combinations to construct EEGformer encoders.

Combinations	BETA			SEED			DepEEG		
	ACC (%)	SPE (%)	SEN (%)	ACC (%)	SPE (%)	SEN (%)	ACC (%)	SPE (%)	SEN (%)
Reg	41.63 ± 5.91	46.59 ± 3.58	35.67 ± 3.26	76.53 ± 1.68	77.26 ± 2.41	73.58 ± 1.94	58.78 ± 5.21	60.51 ± 2.58	57.25 ± 3.42
Syn	57.29 ± 6.50	62.86 ± 5.89	55.28 ± 4.69	80.12 ± 5.12	82.83 ± 4.65	78.86 ± 2.71	60.12 ± 4.86	65.94 ± 3.59	55.26 ± 4.27
Temp	45.36 ± 7.18	53.38 ± 6.38	43.86 ± 5.68	77.28 ± 4.12	78.29 ± 3.83	76.69 ± 3.82	61.73 ± 4.12	65.82 ± 4.78	60.83 ± 2.65
Temp + Syn	66.52 ± 3.82	70.25 ± 2.97	62.23 ± 4.32	85.36 ± 3.61	88.36 ± 4.75	83.45 ± 2.86	70.15 ± 3.18	68.97 ± 3.56	75.65 ± 4.81
Temp + Reg	59.29 ± 3.27	65.93 ± 2.65	58.79 ± 3.54	80.12 ± 3.19	82.33 ± 2.08	79.16 ± 3.19	65.21 ± 2.89	62.14 ± 4.72	72.31 ± 3.75
Syn + Reg	65.72 ± 2.91	70.85 ± 2.58	61.23 ± 5.12	86.73 ± 2.95	88.04 ± 2.36	83.77 ± 3.76	71.46 ± 2.85	61.96 ± 2.36	75.64 ± 3.19
Temp + Syn + Reg	70.15 ± 2.18	75.86 ± 2.04	69.86 ± 3.29	91.58 ± 2.77	92.75 ± 3.72	89.14 ± 2.98	72.19 ± 2.67	70.95 ± 2.38	77.83 ± 2.15

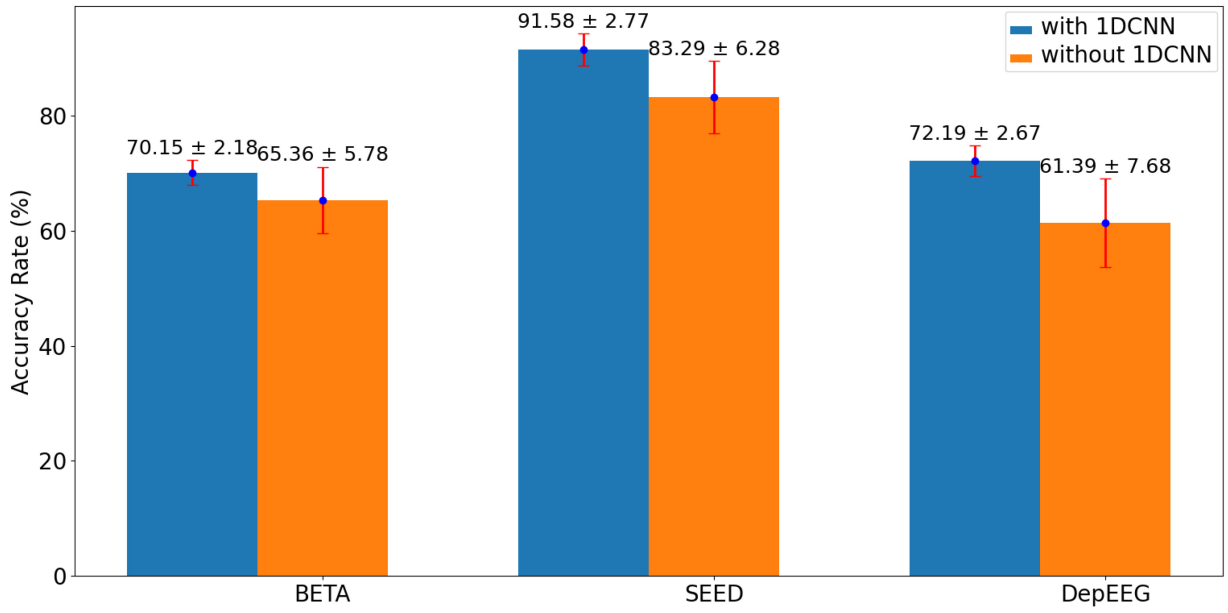

FIGURE 3
Comparison results of using 1DCNN or not to construct the EEGformer pipeline.

TABLE 3 Classification results (ACC, SPE, SEN, and their corresponding SDs) of our model is reported versus varying number of channels and 1.0 s of stimulation.

Channel number	BETA			SEED			DepEEG		
	ACC (%)	SPE (%)	SEN (%)	ACC (%)	SPE (%)	SEN (%)	ACC (%)	SPE (%)	SEN (%)
3	42.73 ± 3.60	50.73 ± 5.17	36.83 ± 4.39	69.54 ± 3.86	70.49 ± 2.96	66.76 ± 4.85	51.29 ± 2.99	50.86 ± 3.75	55.71 ± 4.51
6	50.86 ± 4.49	63.69 ± 2.38	55.17 ± 6.73	73.21 ± 2.83	74.62 ± 3.79	73.61 ± 2.73	56.74 ± 3.85	54.14 ± 2.64	60.26 ± 3.29
9	56.52 ± 2.17	70.46 ± 3.96	65.89 ± 5.26	76.37 ± 3.72	77.24 ± 4.21	78.18 ± 3.82	61.21 ± 4.74	59.75 ± 3.82	65.78 ± 2.79
32	65.21 ± 3.05	72.17 ± 2.57	65.36 ± 4.74	85.98 ± 3.16	86.91 ± 2.64	86.27 ± 4.54	68.56 ± 2.38	65.37 ± 3.57	70.39 ± 4.26
64	70.15 ± 2.18	75.86 ± 2.04	69.86 ± 3.29	91.58 ± 2.77	92.75 ± 3.72	89.14 ± 2.98	72.19 ± 2.67	70.95 ± 2.38	77.83 ± 2.15

and other five comparison methods. As shown in Figure 4, the upper figure shows accuracy comparison results between EEGformer and Conv-CCA across using BETA dataset, and the lower figure shows standard deviation comparison between EEGformer and other five comparison methods across subjects using BETA dataset. The reason of only choosing Conv-CCA to compare with EEGformer is both of them achieve high accuracy on the BETA dataset. From the Figure 5, we can find that EEGformer achieves the lowest standard deviation among other comparison methods, indicating the proposed method generalizes well on unseen data and potentially requires little to model training and calibration for new users, suitable for SSVEP classification tasks.

1. Accuracy comparison between EEGformer and Conv-CCA across subjects using BETA dataset.

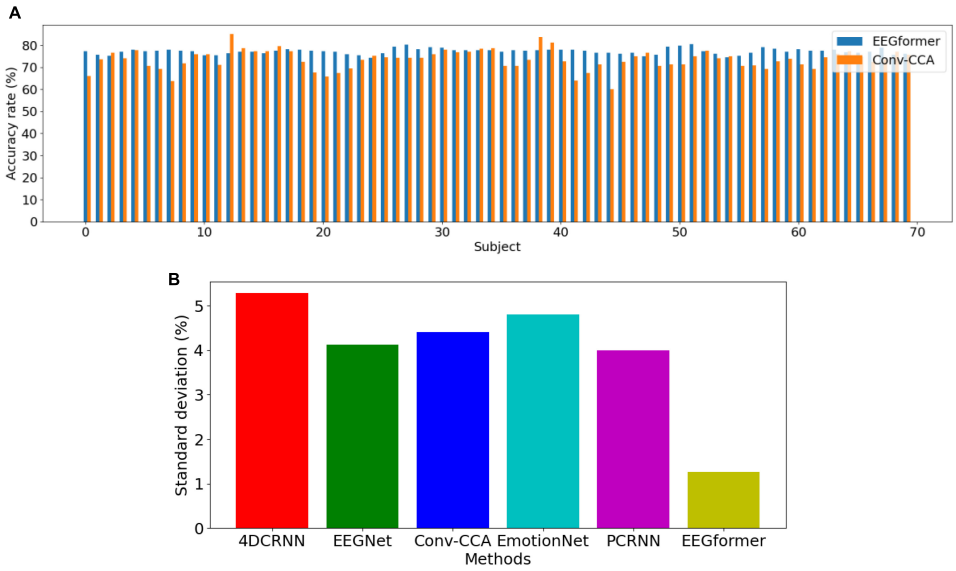

FIGURE 4
Performance comparison between EEGformer and other five comparison methods using leave-one-subject-out cross-validation method based on BETA dataset. **(A)** Accuracy comparison between EEGformer and Conv-CCA across subjects using BETA dataset. **(B)** Standard deviation comparison between EEGformer and other five comparison methods across subjects using BETA dataset.

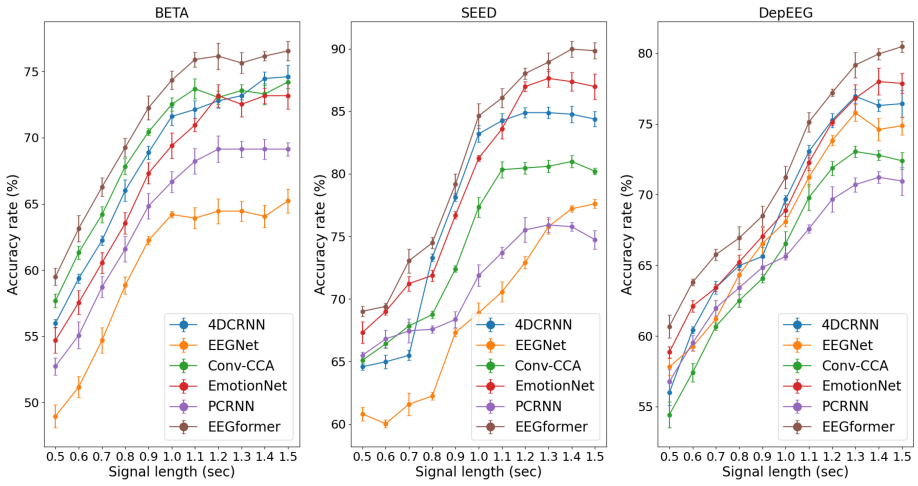

FIGURE 5
Performance (average ACC ± SD %) of segment length *T* using the EEGformer and other comparable models on the three EEG datasets.

2. Standard deviation comparison between EEGformer and other five comparison methods across subjects using BETA dataset.

Furthermore, according to the SSVEP studies, they pursue a higher information transfer rate by not using long EEG segments to execute the target frequency identification task. The model performance can be improved by increasing the segment length *T* because longer EEG segments contain more information about brain activity. Therefore, we investigated the impact of segment length *T* ranges [0.5, 0.6, 0.7, 0.8, 0.9, 1.0, 1.1, 1.2, 1.3, 1.4, 1.5] on model performance. The performance (average ACC and SD) of segment length *T* using the EEGformer and other comparable models on the three EEG datasets are shown in Figure 5. The figure shows that our model achieves the best accuracy rate across the three datasets. For other comparison baseline models, the model performance reduces in some cases if the segment length *T* exceeds 1.2 s. The model performance of the EEGformer on the three datasets showed an increasing trend as the segment length *T* increases, indicating that our method can extract inherent temporal information from EEG and is unaffected by segment length. In addition, the model performance of 4DRCNN and EmotionNet outperforms the performance of other comparison baselines. Because 4DRCNN and the EmotionNet are models that learn spatiotemporal features simultaneously, this operation may facilitate the DL model to learn better feature representation of EEG regional and synchronous characteristics.

4. Discussion

The abovementioned ablation and comparison studies show the rationality of our EEGformer architecture and demonstrate that our model performs outperforms other comparison baselines. This section covers several noteworthy points and future works:

(1) The unified manner, sequentially maps an input sequence into an abstract continuous representation that holds temporal, convolutional, and spatial information of that input outperforms the 2D and 3D structures that integrate frequency, spatial and temporal information of EEG. The EEGformer achieved the highest accuracy rate compared with other comparison baselines, which could be due to the unified EEG characteristics learning manner. Compared with 4DRCNN, which requires the user to manually extract frequency information from raw EEG data and use it as model input, our model is an end-to-end deep method because it uses depth-wise 1DCNN to learn the feature in an EEG-channel-wise manner. In the EEGformer encoder, we sequentially encode the convolutional results generated by the 1DCNN from temporal, convolutional, and spatial dimensions. The temporal, regional, and synchronous transformers were responsible for learning the temporal, regional, and synchronous characteristics of EEG signals. This type of feature learning strategy contains more cues of EEG characteristics than other model structures and performs better than them.

(2) EEG signals are well-known to exhibit data statistics that can drastically change from one subject to another in various aspects (e.g., regional characteristics), but also share similarities in certain other aspects (e.g., synchronous characteristics). To exploit the commonalities while tackling variations, we require a large data sample to train the model and improve its generalization ability. However, the performance of a DL model is always affected by the dataset size. Compare with the dataset size in the computer vision studies, researchers find it difficult to collect a dataset with a similar size in EEG-based clinical studies. Therefore, increasing the number of EEG datasets used for training DL models is crucial to reduce the influence of small dataset size on model performance. To this end, many studies separate the EEG signal collected in a trial into several segments and label them with the same label. Those segments were then used in cross-subject and within-subject classifications, which are two commonly used experimental designs, to execute model training and validate model performance. Meanwhile, those studies also designed model training strategies to improve the model generalization ability. For instance, Guney et al. (2021) trained their model in two stages: the first stage trains globally with all the available data from all the subjects, and then the second stage fine-tunes the model individually using the data of each subject separately. In the future, we can also design a training strategy to reduce the influence of small dataset size on model performance.

(3) Although the experimental results demonstrated that learning temporal, regional, and spatial characteristics in a unified manner facilitates the EEGformer to achieve promising classification performance across three EEG datasets, this result might be unable to provide strong support for clinical treatment that is associated with EEG biomarkers. Because DL methods are essentially considered black boxes, we require novel methods to open the box and visualize the feature learned by the DL model. To this end, an emerging technique known as explainable artificial intelligence (AI) enables the understanding of how DL methods work and what drives their decision-making. The competitive model performance of DL methods and the explainable AI provided a promising way to support effective EEG-based brain activity analysis. By using the explainable AI method, we could visualize the form of the temporal, regional, and spatial characteristics learned by the EEGformer and use it to connect with BFC, as well as perform brain activity analysis.

5. Conclusion

In this study, we proposed a transformer–based EEG analysis model known as EEGformer to capture EEG characteristics in a unified manner. The EEGformer consists of 1DCNN, an EEGformer encoder (sequentially constructed by three components: regional, synchronous, and temporal transformers), and an EEGformer decoder. We conducted ablation studies to demonstrate the rationality of the EEG former. The results not only supported our hypothesis that a machine learning method capable of capturing the EEG characteristics in a unified manner can be applied to EEG-based brain activity analysis tasks but also demonstrated that convolutional features could accurately represent regional and spatial characteristics of EEG signals. The LOSO cross-validation method is utilized to compare the model performance between EEGformer and other five comparison methods, the result shows the proposed method generalizes well on unseen data and potentially requires little to model training and calibration for new users, suitable for SSVEP classification tasks. We also investigate the impact of segment length T on model performance, and the results show that our method can extract inherent temporal information from EEG and is unaffected by the segment length. The proposed EEGformer outperforms the comparison models, which perform well in other studies on the three EEG datasets.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: http://bci.med.tsinghua.edu.cn/download.html and https://bcmi.sjtu.edu.cn/home/seed/seed.html.

Ethics statement

The studies involving human participants were reviewed and approved by the Institutional Review Board of Beijing Anding Hospital of Capital Medical University. The patients/participants provided their written informed consent to participate in the

data collection. Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable data included in this article.

Author contributions

ZW, ML, and WD contributed to the conception and design of the study. SL and ML performed the data analysis. ZW and WD drafted the manuscript. JH and HT participated in editing the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant (62161024), China Postdoctoral Science Foundation under Grant (2021TQ0136 and 2022M711463), and the State Key Laboratory of Computer Architecture (ICT, CAS) Open Project under Grant (CARCHB202019).

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

Abdull, M. M., Chandler, C., and Gilbert, C. J. (2016). Glaucoma,"the silent thief of sight": Patients' perspectives and health seeking behaviour in Bauchi, northern Nigeria. *BMC Ophthalmol.* 16:44. doi: 10.1186/s12886-016-0220-6

Chen, J., Zhang, Y., Pan, Y., Xu, P., and Guan, C. (2022). A Transformer-based deep neural network model for SSVEP classification. *arXiv* [Preprint]. arXiv:2210.04172.

Du, Y., Xu, Y., Wang, X., Liu, L., and Ma, P. (2022). EEG temporal–spatial transformer for person identification. *Sci. Rep.* 12:14378.

Duan, R. N., Zhu, J. Y., and Lu, B. L. (2013). "Differential entropy feature for EEG-based emotion classification," in *Proceedings of the international IEEE/EMBS conference on neural engineering* (San Diego, CA), 81–84. doi: 10.1109/NER.2013.6695876

Guedes, R. A. P. (2021). Glaucoma, collective health and social impact. *Rev. Bras.Oftalmol.* 05–07. doi: 10.5935/0034-7280.20210001

Guney, O. B., Oblokulov, M., and Ozkan, H. J. (2021). A deep neural network for ssvep-based brain-computer interfaces. *IEEE Trans. Biomed. Eng.* 69, 932–944. doi: 10.1109/TBME.2021.3110440

Ibáñez-Soria, D., Soria-Frisch, A., Garcia-Ojalvo, J., and Ruffini, G. (2019). Characterization of the non-stationary nature of steady-state visual evoked potentials using echo state networks. *PLoS One* 14:e0218771. doi: 10.1371/journal.pone.0218771

Khok, H. J., Koh, V. T., and Guan, C. (2020). "Deep multi-task learning for SSVEP detection and visual response mapping," in *Proceedings of the IEEE international conference on systems, man, and cybernetics (SMC)*, (Toronto, ON), 1280–1285. doi: 10.1109/SMC42975.2020.9283310

Lawhern, V. J., Solon, A. J., Waytowich, N. R., Gordon, S. M., Hung, C. P., and Lance, B. J. (2018). EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces. *J. Neural Eng.* 15:056013. doi: 10.1088/1741-2552/aace8c

Li, X., Wei, W., Qiu, S., and He, H. (2022). "TFF-Former: Temporal-frequency fusion transformer for zero-training decoding of two BCI tasks," in *Proceedings of the 30th ACM international conference on multimedia*, (Lisboa), 51–59. doi: 10.1145/3503161.3548269

Li, Y., Xiang, J., and Kesavadas, T. J. (2020). Convolutional correlation analysis for enhancing the performance of SSVEP-based brain-computer interface. *IEEE Trans. Neural Syst. Rehabil. Eng.* 28, 2681–2690. doi: 10.1109/TNSRE.2020.3038718

Li, Z., Wang, Q., Zhang, S. F., Huang, Y. F., and Wang, L. Q. (2022). Timing of glaucoma treatment in patients with MICOF: A retrospective clinical study. *Front. Med.* 9:986176. doi: 10.3389/fmed.2022.986176

Liu, B., Huang, X., Wang, Y., Chen, X., and Gao, X. J. (2020). BETA: A large benchmark database toward SSVEP-BCI application. *Front. Neurosci.* 14:627. doi: 10.3389/fnins.2020.00627

Nentwich, M., Ai, L., Madsen, J., Telesford, Q. K., Haufe, S., Milham, M. P., et al. (2020). Functional connectivity of EEG is subject-specific, associated with phenotype, and different from fMRI. *Neuroimage* 218:117001. doi: 10.1016/j.neuroimage.2020.117001

Qin, K., Wang, R., and Zhang, Y. (2021). Filter bank-driven multivariate synchronization index for training-free SSVEP BCI. *IEEE Trans. Neural Syst. Rehabil. Eng.* 29, 934–943. doi: 10.1109/TNSRE.2021.3073165

Raut, R. V., Snyder, A. Z., Mitra, A., Yellin, D., Fujii, N., Malach, R., et al. (2021). Global waves synchronize the brain's functional systems with fluctuating arousal. *Sci. Adv.* 7:eabf2709. doi: 10.1126/sciadv.abf2709

Schielke, A., and Krekelberg, B. (2022). Steady state visual evoked potentials in schizophrenia: A review. *Front. Neurosci.* 16:988077. doi: 10.3389/fnins.2022.988077

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, 815–828. doi: 10.1007/s11571-020-09634-1

Tsoneva, T., Garcia-Molina, G., and Desain, P. (2021). SSVEP phase synchronies and propagation during repetitive visual stimulation at high frequencies. *Sci. Rep.* 11:4975. doi: 10.1038/s41598-021-83795-9

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., et al. (2017). Attention is all you need. *Adv. Neural Inf. Process. Syst.* 30:15. doi: 10.48550/arXiv.1706.03762

Wan, Z., Huang, J., Zhang, H., Zhou, H., Yang, J., and Zhong, N. (2020). HybridEEGNet: A convolutional neural network for EEG feature learning and depression discrimination. *IEEE Access* 8, 30332–30342. doi: 10.1109/ACCESS.2020.2971656

Wang, Y., Huang, Z., McCane, B., and Neo, P. (2018). "EmotioNet: A 3-D Convolutional Neural Network for EEG-based Emotion Recognition," in *Proceedings of the 2018 International Joint Conference on Neural Networks (IJCNN)* (Rio de Janeiro, Brazil), 1–7. doi: 10.1109/IJCNN.2018.8489715

Waytowich, N., Lawhern, V. J., Garcia, J. O., Cummings, J., Faller, J., Sajda, P., et al. (2018). Compact convolutional neural networks for classification of asynchronous steady-state visual evoked potentials. *J. Neural Eng.* 15, 066031. doi: 10.1088/1741-2552/aae5d8

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 *Proceedings of the international joint conference on neural networks*, (Rio de Janeiro: IEEE), 1–7. doi: 10.1109/IJCNN.2018.8489331

Zhang, X., Yao, L., Wang, X., Monaghan, J., McAlpine, D., and Zhang, Y. (2021). A survey on deep learning-based non-invasive brain signals: Recent advances and new frontiers. *J. Neural Eng.* 18:031002. doi: 10.1088/1741-2552/abc902

Zhang, Y., Xie, S. Q., Wang, H., and Zhang, Z. J. (2020). Data analytics in steady-state visual evoked potential-based brain–computer interface: A review. *IEEE Sens. J.* 21, 1124–1138. doi: 10.1109/JSEN.2020.3017491

Zhang, Y., Yin, E., Li, F., Zhang, Y., Guo, D., Yao, D., et al. (2019). Hierarchical feature fusion framework for frequency recognition in SSVEP-based BCIs. *Neural Netw.* 119, 1–9. doi: 10.1016/j.neunet.2019.07.007

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. Ment. Dev.* 7, 162–175. doi: 10.1109/TAMD.2015.2431497

Zhou, Y., He, S., Huang, Q., and Li, Y. J. (2020). A hybrid asynchronous brain-computer interface combining SSVEP and EOG signals. *IEEE Trans. Biomed Eng.* 67, 2881–2892. doi: 10.1109/TBME.2020.2972747

Zhuang, X., Yang, Z., and Cordes, D. J. (2020). A technical review of canonical correlation analysis for neuroscience applications. *Hum. Brain Mapp.* 41, 3807–3833. doi: 10.1002/hbm.25090