Wang et al.
*Health Information Science and Systems (2023) 11:38*
https://doi.org/10.1007/s13755-023-00239-6

Health Information Science and Systems

RESEARCH

# Combining temporal and spatial attention for seizure prediction

Yao Wang[1], Yufei Shi[2], Zhipeng He[2], Ziyi Chen[3] and Yi Zhou[2*]

**Abstract**

*Purpose:*: Approximately 1% of the world population is currently suffering from epilepsy. Successful seizure prediction is necessary for those patients. Influenced by neurons in their own and surrounding locations, the electroencephalogram (EEG) signals collected by scalp electrodes carry information of spatiotemporal interactions. Therefore, it is a great challenge to exploit the spatiotemporal information of EEG signals fully.

*Methods:*: In this paper, a new seizure prediction model called Gatformer is proposed by fusing the graph attention network (GAT) and the Transformer. The temporal and spatial attention are combined to extract EEG information from the perspective of spatiotemporal interactions. The model aims to explore the temporal dependence of single-channel EEG signals and the spatial correlations among multi-channel EEG signals. It can automatically identify the most noteworthy interaction in brain regions and achieve accurate seizure prediction.

*Results:*: Compared with the baseline models, the performance of our model is significantly improved. The false prediction rate (FPR) on the private dataset is 0.0064/h. The average accuracy, specificity and sensitivity are 98.25%, 99.36% and 97.65%.

*Conclusion:*: The proposed model is comparable to the state of the arts. Experiments on different datasets show that it has good robustness and generalization performance. The high sensitivity and low FPR prove that this model has great potential to realize clinical assistance for diagnosis and treatment.

**Keywords:** Seizure prediction, EEG, GAT, Transformer, Spatiotemporal attention

## Introduction

The local neurons in the brain discharge abnormally during epileptic seizures. There is a transient disturbance in the system function, which is manifested as limb convulsions, violent clonus or loss of consciousness [1]. The suddenness and intermittent recurrence of seizures, which do not distinguish between time and place, limit the daily work and life of patients and place a heavy burden on them. Persistent seizures may cause permanent damage to patients with mental impairment, decreased intelligence and even death [2]. Therefore, the timely seizure prediction and intervention are of great importance for those patients who cannot be cured by drugs and surgery.

Researchers typically use two different methods to achieve seizure predictions. One is based on the threshold analysis of the trend of EEG changes during seizures [3–5]. Once the characteristic value of a certain metric exceeds the threshold, the system is activated and an alert will sound immediately. The other is to train a classifier to achieve the classification of two EEG states according to the difference between preictal and interictal periods [6–9]. Therefore, the selection and the design of a classifier with high sensitivity and low FPR will determine the accuracy of the constructed model [10]. Threshold-based methods are highly interpretable, but their generalization abilities are weak. And the handcrafted low-dimensional features are difficult to work for all patients. Epilepsy is highly patient-specific, so such methods are susceptible to the threshold fluctuations, which can lead to a high

*Correspondence: zhouyi@mail.sysu.edu.cn
[2] Department of Medical Informatics, Zhongshan School of Medicine, Sun Yat-sen University, Guangzhou 510080, Guangdong, China
Full list of author information is available at the end of the article

FPR. Due to the development of the machine learning and the deep learning theories, classification-based methods are more commonly used in seizure prediction studies.

In the researches based on machine learning methods, scholars usually focus on the extraction and study of features in the time domain, frequency domain, time–frequency domain and nonlinear dynamics [11–14]. The rules learned from different features are crucial for the performance of the machine learning models. And the computational complexity will increase as the data scales up [15, 16]. The purpose of the deep learning approach is to select and construct an end-to-end classifier that enables automatic classification of preictal and interictal EEG signals. This method can automatically extract feature information for different data, thus eliminating the step of manually extracting features.

Wei et al. [17] established a deep learning model for seizure prediction based on the long-term recurrent convolutional network (LRCN), which was able to predict seizures 21 min in advance. Usman et al. [18] used the convolutional neural networks (CNN) with the support of vector machines to classify EEG signals in the CHB-MIT dataset with a sensitivity of 92.7%. Daoud and Bayoumi [19] used a fusion model of the deep CNN (DCNN) and the bi-directional long short-term memory (BiLSTM) to select the optimal EEG channels for the seizure prediction tasks. Jana and Mukherjee [20] used the CNN and Zhao et al. [21] improved the convolutional kernel of CNN, both of them obtained good performance. Although the above studies were able to substantially improve the seizure prediction performance of the models, most of these studies focused on the performance of the algorithms. They did not analyze the models from the perspective of EEG signals and were unable to integrate the models with the complex features of EEGs.

The generation of EEG is the result of the establishment of communication relationships between multiple regions of the human brain. The rich information it contains about the spatial activity of the brain is yet to be explored. At the same time, EEG has a high temporal resolution, as evidenced by the rapid and disordered waveform changes observed within a relatively short seizure duration. Traditional neural networks usually focus on the information of local neighborhoods, ignoring the long-distance dependence of EEGs. Therefore, it is not conducive to the extraction of spatiotemporal information of EEGs. How to extract the key information carried by EEG recordings more accurately and efficiently has become compelling methods for scholars.

By mimicking the network structure of human attention, the attention mechanism is able to focus on the critical information and assign weights to each input feature. Many studies have combined the attention mechanism from different perspectives with the traditional models and applied it to the prediction of epileptic seizure. Jiang et al. [22] designed a time-attention based simulation module. They used it to extract the temporal information of the intracranial EEG (iEEG) fragments that were converted into images. All information was then fed into a pre-trained residual network (ResNet) to achieve the classification of preictal and interictal periods. Sun et al. [23] constructed a channel attention module to improve the CNN, Ma et al. [24] introduced the channel attention and spatial attention into the batch normalization long short-term memory (BNLSTM) model, and Yang et al. [25] combined a spectral and channel attention module with ResNet. All of them achieved successful seizure predictions on different datasets.

The EEG signals can be considered as the graph-structured data in non-Euclidean space. The analysis of EEG signals on the graph domain enables the efficient use of structural features embedded between nodes. It is also able to provide an adequate representation of the relationships between the different channels. Therefore, some researchers used the graph convolution neural network (GCN) to extract the spatial information of EEG signals [26–28]. Veličković et al. [29] constructed a new network GAT using an attention mechanism instead of the normalized convolution operation of GCN. This improvement allows the model to get rid of the dependence of the GCN on Laplacian matrix and enables automatic capture of key information on arbitrary graphs. Different studies have demonstrated that the application of GAT to EEG analysis is an effective option [30–32]. Zhao et al. [33] learned the spatial connectivity of the brain based on the optimized GAT model. The sensitivity of the model on CHB-MIT reached 98.33%.

The epileptic EEG signal is a typical time-series data with interrelated time periods. Recurrent neural networks (RNN) can handle temporal data by fully considering the input of the previous moment and applying it to the current output information. However, this inherent sequential nature makes it impossible to compute the model in parallel during training, reducing operational efficiency. In 2017, Vaswani et al. [34] proposed the Transformer model, which uses the self-attention mechanism to allow modeling of dependencies. This model can realize parallel computing and improve the efficiency and accuracy of task processing. Therefore, once the Transformer-based framework was proposed, it obtained state-of-the-art results in several areas [35–38]. Bhattacharya et al. [39] extracted time and frequency domain features to achieve more accurate seizure prediction using the Transformer. Hu et al. [40]

also demonstrated the great potential of Transformer in seizure prediction applications by using a hybrid Transformer model.

Relying on the self-attention mechanism in the model, GAT and Transformer can capture the spatial and temporal dependence of EEG signals better. Considering the above research status, we fuse GAT and Transformer to construct a new model named Gatformer, which is able to focus on both temporal sequence and spatial correlation of EEG signals and jointly extract complex information of EEGs. Our main research work is as follows:

1. Based on the attention to the time and space of EEGs, we have developed a novel seizure prediction framework Gatformer and used the $k$ of $n$ post-processing technique to decrease false predictions.

2. The Gatformer uses the attention mechanism to capture the complex information carried by EEG recordings from different perspectives to improve the interpretation of the model. The attention mechanism in the GAT module is used to model the interaction between EEG signals, and the attention mechanism in the Transformer enables the simultaneous extraction of temporal information from multiple channels of EEG recordings.

3. Experiments are carried out on both the private dataset and the CHB-MIT dataset, which can achieve seizure prediction with a high sensitivity and a low FPR. The results on the CHB-MIT dataset are close to or better than the current optimal methods, demonstrating the effectiveness of our model.

The rest of this paper is organized as follows. Section 2 describes the architecture of the proposed Gatformer algorithm, and Sect. 3 formulates the dataset and the experimental results. The discussions are given in Sect. 4, and Sect. 5 draws conclusions.

## Methods

In this section, we discuss the theoretical knowledge involved in the proposed model in detail. The basic architecture of the model have been introduced in Sect. 2.1. In the next part, the GAT and Transformer are discussed separately. We mainly focus on the attention mechanism of the two models and make a detailed introduction.

### Overall architecture

As shown in Fig. 1, the EEG signals contain the spatial features of the interaction between multi-channel electrodes and the signal fluctuation information of single-channel electrode over time. By extracting the spatial correlation and temporal features of EEG signals, Gatformer can determine the category to which EEG segments belong. The classification results are output in the end.

The raw EEG consists of $T$ sequential EEG segments with feature dimension $F$. Each EEG fragment contains $N$ electrode channels. During a seizure, the EEG signal of the $n$th electrode at moment $t$ is constantly influenced by the surrounding electrodes, which can be expressed as a spatial interaction of EEG signals. To represent this interaction, we constructed an $N * N$ adjacency matrix using Pearson coefficients to calculate the correlation between electrodes. When the Pearson coefficient between electrodes is greater than a threshold, the value of the adjacency matrix is 1. Otherwise, it is 0. We input a segment of EEG of duration $T$ into the model and learn spatial information via GAT. After processing and pooling by GAT layer, the vector dimension of EEG becomes $(T, F')$.

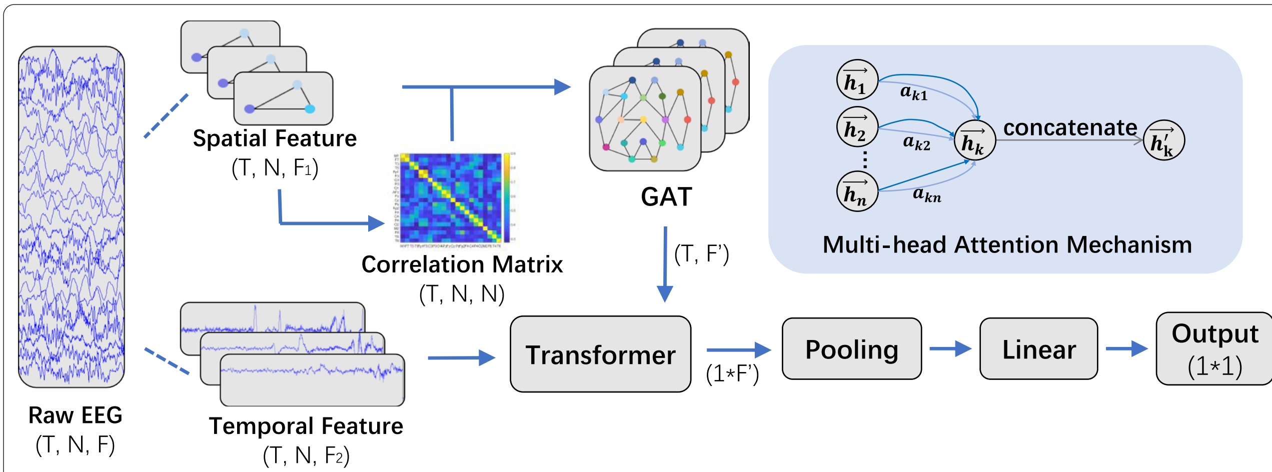


**Fig. 1** The framework of the proposed seizure prediction algorithm

At time $t$, the signal of the $n$th electrode fluctuates on the basis of the electrode signal at $t-1$ moment. It shows the time dependence of the EEG signal. We use Transformer to memorize the change of the signal in the time dimension and capture the backward and forward order relationship of the EEG sequence. The dimension of the data becomes $1 * F'$ after the transformer layer. Finally, this vector is classified to obtain the recognition result.

## Gatformer

The EEG signals are the result of information processing and transmission by neurons between various regions of the brain. During the transmission of information in the brain, different neurons interact with each other and different signal channels have different strengths. Therefore, the EEG signals collected by the scalp contain a large amount of spatial information. The graph network is a collection of functions organized in the topological space [41]. It can reason about relationships based on the diagram structures. The basic idea is that the node itself and its neighbors determine the properties of this node. The GAT uses an attention mechanism to weigh the summation of neighboring node features and is able to assign different weights among different nodes. Therefore, it can learn the mapping relationships between different brain regions and seizure prediction.

In the temporal dimension, single-channel EEG data can be represented as a collection of amplitude magnitudes [42]. Many RNNs and related variants have been applied to learn the temporal dependencies of EEG data. The Transformer-based architectures are able to focus on the features associated with seizures by exploiting the self-attentive mechanism and the strong ability to learn long-term dependencies. In the following subsections, we will describe the GAT and Transformer separately in detail.

### *Spatial attention: GAT*

In this model, we use a 2-layer GAT block to process the spatial information of EEGs. We use the electrode channels of EEG recordings as the nodes of the graph. The Pearson correlation matrix is used to calculate the spatial correlation. The index size represents the closeness of the relationship between EEG channels, which completes the construction of the input graph.

Firstly, the node features $H = \left\{ \overrightarrow{h_1}, \overrightarrow{h_2}, \ldots, \overrightarrow{h_N} \right\}$, $\overrightarrow{h_i} \in R^F$ and the correlation matrix at moment $t$ are input into the GAT network. $N$ represents the number of electrode channels and $F$ is the feature dimension of each node. The self-attention mechanism of nodes determines the weight of neighboring node features in the feature update process. Here, we transform the dimensionality of the input features according to a learnable weight matrix $W \in R^{F' \times F}$, where $F'$ denotes the dimension of the output node. The weight matrix is first Xavier initialized during the training of the model, and the optimal weight matrix is obtained by iterative training. Then, the influence degree of Node $i$ on Node $j$ is calculated by $\overrightarrow{h_i}$ and $\overrightarrow{h_j}$ .

$$e_{ij} = \mathrm{a}\left( W\overrightarrow{h_i}, W\overrightarrow{h_j} \right), \tag{1}$$

where the feed forward neural network $\mathrm{a}(\cdot)$ can concatenate the resultant vector to complete the mapping of features.

Afterward, we calculate the attention coefficients of all nodes for Node $i$ and complete the normalization operation of attention weights using the softmax to get the final attention coefficients. The related formula is as follows.

$$a_{ij} = \frac{\exp\left( \mathrm{LeakyReLU}\left( \overrightarrow{a}^T \left[ W\overrightarrow{h_i} \parallel W\overrightarrow{h_j} \right] \right) \right)}{\sum_{k \in N_i} \exp\left( \mathrm{LeakyReLU}\left( \overrightarrow{a}^T \left[ W\overrightarrow{h_i} \parallel W\overrightarrow{h_k} \right] \right) \right)}, \tag{2}$$

where $\parallel$ is the concatenation operator.

As a nonlinear activation function, $LeakyReLU(\cdot)$ can enhance the generalization ability of the model. Finally, the multi-head attention mechanism [34] is used to learn the attention weights of node features to enhance the learning ability of the model. After being processed by the GAT attention layer, the characteristic of Node $i$ can be expressed as Eq. 3.

$$\overrightarrow{h_i'} = \sigma\left( \frac{1}{K} \sum_{k=1}^{K} \sum_{j \in N_i} a_{ij}^k W^k \overrightarrow{h_j} \right). \tag{3}$$

We introduce the residual connection to sum the processed data. The final output of the GAT module is denoted by $X_G$.

### *Temporal attention: Transformer*

In 2017, Vaswani et al. [34] proposed the Transformer model in "attention is all you need", which enabled parallel processing tasks for long sequences, thus extending the speed and capacity of sequential deep learning models to unprecedented speeds. The Transformer follows the encoder–decoder architecture of the neural sequence transduction models. In this study, we use the Transformer encoder to process $X_G$ to capture the remote dependencies of EEG sequences.

As shown in Fig. 2, the encoder used in this paper consists of a positional encoding layer and a stack of four identical encoder layers. Positional encoding modifies the value

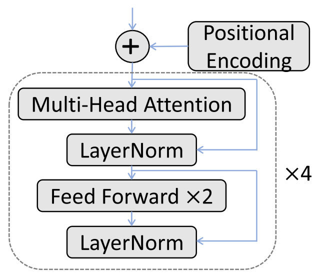


**Fig. 2** The architecture of the Transformer

of each embedding vector to represent its position in the EEG sequence. The sequence information is encoded using Eqs. 4 and 5 to produce an encoding array to maintain the temporal order.

$$PE_{(pos,2x)} = \sin\left(\frac{pos}{10}, \frac{10,000^{2x}}{\gamma}\right), \tag{4}$$

$$PE_{(pos,2x+1)} = \cos\left(\frac{pos}{10}, \frac{10,000^{2x}}{\gamma}\right). \tag{5}$$

where the dimension of the linear embedding $X_G$ exported from the GAT module is expressed as $\gamma$, $x$ is the dimension of the features from 1 to $\gamma$, and $pos$ denotes the position of the time series, i.e., from 1 to the length of the EEG series.

The input embedding layer can be directly sum up with the position coding at the element level to obtain a new embedding vector, $X'_G$. There are two sub-layers in the encoder layer: the Multi-Head Attention layer and the Feed Forward layer. Each of them has a residual connection and layer normalization to accelerate the convergence of the network and avoid overfitting. In the Multi-Head Attention layer, the attention mechanism completes the linear mapping of $X'_G$ to generate the query $Q$, key $K$ and value $V$. The calculations are as follows.

$$\mathrm{Q} = X'_G * W_Q, \tag{6}$$

$$\mathrm{K} = X'_G * W_K, \tag{7}$$

$$\mathrm{V} = X'_G * W_V, \tag{8}$$

where $W_Q$, $W_K$ and $W_V$ are the parameter matrices.

The feature weights of EEGs can be calculated using $Q$ and $K$. The division by $\sqrt{d_k}$ prevents excessive dot product. $d_k$ is the dimension of $V$. Therefore, the output of the self-attention head is:

$$\mathrm{Head} = \mathrm{Attention}(\mathrm{Q}, \mathrm{K}, \mathrm{V}) = \mathrm{softmax}\left(\frac{\mathrm{Q}K^T}{\sqrt{d_k}}\right)V. \tag{9}$$

The role of the softmax activation function is to perform normalization of the data. At the same time, it is able to increase the data gap and enhance the representational ability of attention mechanism. Finally, the multi-head attention mechanism maps Q, K and V with different linear relations and concatenates their attention with the function $Concat(\cdot)$. In this study, we have used attention heads of h = 4.

$$\mathrm{MultiHead}(\mathrm{Q}, \mathrm{K}, \mathrm{V}) = \mathrm{Concat}(\,\mathrm{head}_1, \ldots,\ \mathrm{head}_h)W^0. \tag{10}$$

In the Feed Forward layer, two linear transformations and a activation function ReLU are included.

$$\mathrm{FFN}(\mathrm{x}) = \mathrm{ReLU}(0, \mathrm{x}W_1 + b_1)W_2 + b_2. \tag{11}$$

The features generated by encoder are input to the Pooling layer and the Linear layer for data dimensionality transformation, and the final classification results can be obtained. In the model training process, we use the cross entropy as the loss function to measure the difference between the actual labels and the predicted values. Besides, the Adam optimizer is employed in experiments.

## Experiments and results

This section focuses on the dataset, details and results of the experiments. First of all, the private dataset used in the article and the rules of data selection are mentioned. In the next part, the evaluation metrics of the model and the implementation details are explained. The results on different baseline models are analyzed and compared in the end.

### Dataset

The dataset used in this paper was obtained from the First Affiliated Hospital of Sun Yat-sen University, which contained EEG recordings of 20 electrode channels from 13 patients (4 males and 9 females, aged 15–62 years). All EEG signals were collected using the international 10-20 system and sampled at a rate of 500 samples/s, i.e. 500 Hz. Table 1 shows the detailed dataset information.

To enhance the rigor and the experimental credibility of the article and to facilitate comparison with other similar studies, we also used the public dataset CHB-MIT, which is one of the most widely used datasets in the task of assessing seizure prediction. It collected 198 EEG recordings from 22 patients (one of them had a re-measurement after 1.5 years). All EEG data were sampled at a rate of 256 Hz using the International 10-20 EEG electrode location and naming system. Since the patients in

**Table 1 Characteristics of the private dataset**

| Patients | Gender | Age (years) | Consciousness | Monitoring time (h) | Seizure number |
|---|---|---|---|---|---|
| 1 | Male | 62 | Wake–Sleep | 24 | 8 |
| 2 | Male | 19 | Wake–Sleep | 24 | 4 |
| 3 | Female | 16 | Wake–Sleep | 24 | 7 |
| 4 | Female | 15 | Wake–Sleep | 24 | 3 |
| 5 | Female | 15 | Wake–Sleep | 24 | 4 |
| 6 | Female | 22 | Wake–Sleep | 48 | 7 |
| 7 | Female | 40 | Wake–Sleep | 24 | 5 |
| 8 | Female | 16 | Wake–Sleep | 24 | 5 |
| 9 | Female | 26 | Wake–Sleep | 24 | 6 |
| 10 | Female | 31 | Wake–Sleep | 24 | 8 |
| 11 | Male | 20 | Wake–Sleep | 24 | 5 |
| 12 | Male | 46 | Wake–Sleep | 24 | 4 |
| 13 | Female | 15 | Wake–Sleep | 24 | 6 |

this dataset used different electrodes at the time of EEG acquisition, we selected 18 electrodes common to all patients for follow-up experiments: FP1–F7, F7–T7, T7–P7, P7–O1, FP1–F3, F3–C3, C3–P3, P3–O1, FP2–F4, F4–C4, C4–P4, P4–O2, FP2–F8, F8–T8, T8–P8, P8–O2, FZ–CZ, CZ–PZ.

### Data selection rules

This study is dedicated to processing the raw EEG data in order to preserve its information to the maximum extent. Therefore, we only performed data cleaning and screening to exclude the interference of erroneous data in the pre-processing stage. No manual feature extraction and other processing work were performed.

For seizure prediction tasks, it is first necessary to specify the seizure prediction horizon (SPH) and the seizure occurrence period (SOP). A successful seizure prediction should be that there is one seizure after the alarm. And this seizure is within the SOP and after the SPH. The SOP indicates the time during which the patient is likely to have a seizure. The SPH means the time interval between the start of the alarm and the start of SOP. The SPH is a buffer phase to alert the physician or the patient to take interventions, which should not be too short. And it should also not be too long. Otherwise, it is prone to constant mental tension in the patient.

We set a 30-min SOP and a 5-min SPH according to the definition of Maiwald et al. [43]. All the preictal EEG signals were selected according to this rule. If the requirements of the SPH and SOP were not satisfied, the data of this seizure was discarded. The final EEG data used in this paper are shown in Table 1.

We divided the dataset into two parts. One part of the data (patients 1–9) was the general dataset, which was used to train the seizure prediction model. The other part of the data (patients 10–13) was used as the generalized dataset to verify the generalization performance of the model. We used a sliding window of 1.5 s with 20% window overlap to augment the preictal data. In addition, if the interval between two seizures was more than 4 h, the intervening 30-min EEG signals were extracted as the interictal period for research.

### Evaluation indicators

The performance of the proposed algorithm is measured by accuracy, specificity, sensitivity and FPR. Accuracy is the most common evaluation metric that assesses the global performance of the classification. By measuring the ability of the model to identify interictal data and preictal data, specificity and sensitivity are used to assess the accuracy of seizure prediction. The FPR occupies an important position in the task of seizure prediction. It indicates the proportion of interictal data incorrectly identified as preictal data, which is an indicator for calculating the probability of misdiagnosis. The calculation formulas of the above indicators are as follows.

$$\text{Accuracy} = \frac{TP + TN}{TP + FN + TN + FP}, \tag{12}$$

$$\text{Specificity} = \frac{TN}{TN + FP}, \tag{13}$$

$$\text{Sensitivity} = \frac{TP}{TP + FN}, \tag{14}$$

$$\text{FPR} = \frac{FP}{TN + FP}, \tag{15}$$

where TP is the true positive, representing the number of the preictal segments that the model correctly predicted. The FP is the false positive, which represents the wrong prediction of the positive category by the model, i.e., judging the interictal period as preictal. Similarly, the true negative, TN, is the number of interictal fragments correctly predicted by the model. The false negative, FN, means the number of interictal EEG recordings incorrectly classified by the model.

### Implementation details

Taking the private dataset as the input, the architecture of proposed model is shown in Table 2. The model inputs a time series of 750 * 20 * 1, where 750 is the length of the time series, i.e., 1.5 s. 20 represents the number of EEG channels, and 1 means the feature dimension. We

**Table 2 The parameter of Gatformer on the private dataset**

| | | Activation | Input size | Output size | Heads |
|---|---|---|---|---|---|
| GAT | GAT layer | ReLU | 750 * 20 * 1 | 750 * 20 * 64 | 2 |
| | GAT layer | ReLU | 750 * 20 * 64 | 750 * 20 * 64 | 2 |
| | Concatenate | – | 750 * 20 * 64 | 750 * 20 * 128 | – |
| | Pooling layer | – | 750 * 20 * 128 | 750 * 128 | – |
| Positional encoding | | – | 750 * 128 | 750 * 128 | – |
| Encoder (x4) | Input | – | 750 * 128 | 750 * 128 | – |
| | Multi-Head attention layer | – | 750 * 128 | 750 * 128 | 4 |
| | LayerNorm layer | – | 750 * 128 | 750 * 128 | – |
| | Feed Forward layer | ReLU | 750 * 128 | 750 * 512 | – |
| | Feed Forward layer | – | 750 * 512 | 750 * 128 | – |
| | LayerNorm layer | – | 750 * 128 | 750 * 128 | – |
| Pooling layer | | – | 750 * 128 | 1 * 128 | – |
| LayerNorm layer | | – | 1 * 128 | 1 * 128 | – |
| Linear layer | | Sigmoid | 1 * 128 | 1 * 1 | – |

Activation means that each layer is followed by an activation function

construct two 2-head GAT layers and aggregate them. After downscaling, they are used as the input to the Transformer layer. The Transformer layer captures the temporal information using positional encoding and 4 encoder layers. The model is finally dimensionally transformed to output a 1 * 1 classification result. We built our model based on the Pytorch 1.9.0. All experiments were done on the NVIDIA P6000. The hyperparameters of our model were as follows. The learning rate was set to 1e−3. The batch size was 256. And the dropout rate was 0.3. The cross entropy was used as the loss function.

### Baseline models

To demonstrate the effective performance of the model, we selected the typical models as the baseline models for comparison. We mainly focused on the graph neural networks, temporal networks and spatiotemporal fusion neural networks to conduct experiments.

GAT: we ablated the Transformer layer in the Gatformer model and built a 2-layer GAT model as a way to observe the effect of the model when only the spatial attention mechanism is available. The parameters of the model such as the number of attention heads are kept the same as the original Gatformer.

GCN: the GCN uses the Laplacian matrix to perform spectral decomposition of features. It aggregates information from neighboring nodes to obtain channel information of EEG signals. We constructed a 2-layer GCN model for comparison in this study.

Transformer: we ablated the GAT layer in the Gatformer model and used a single Transformer to build a time-series-based seizure prediction model that captures the temporal attention of EEG signals.

LSTM: we used the LSTM model for comparison with the Transformer. It is a RNN with selective memory function. By updating the cell states at each moment, it can learn key information from historical time series and infer future trends of temporal fluctuations.

Gcnformer: we replaced the 2-layer GAT in the Gatformer with a 2-layer GCN module. The purpose is to allow a good observation of the effect of removing the spatial attention mechanism on the model performance.

GatLSTM: we fused a 2-layer GAT with a LSTM to construct the spatiotemporal fusion model GatLSTM, which can also extract spatiotemporal information. Compared with the original model, it pays less attention to the temporal correlation of EEG.

The hyperparameter settings of all the above baseline models are consistent with the Gatformer, i.e., learning rate is 1e−3, batch size is 256, and hidden units is 64.

### Result analyses and comparisons

We compared our method with the baseline models using EEG recordings from the general dataset. We performed experiments using the fivefold cross-validation and used the average performance as the performance of experiments. The results are shown in Table 3.

As graph neural networks, both GCN and GAT can aggregate information of surrounding nodes to the central vertex and then learn the feature representation on the graph domain. In Table 3, GAT performs better in terms of accuracy, sensitivity and FPR. This is the fact that the GCN utilizes a fixed Laplacian matrix to treat each node equally, while the GAT uses attention coefficients to calculate the feature weights. The graph of

**Table 3 Comparison of the results on the general dataset using different algorithms**

| | Accuracy (%) | Specificity (%) | Sensitivity (%) | FPR (/h) |
|---|---|---|---|---|
| GCN | 65.87 ± 3.46 | 86.68 ± 4.29 | 76.86 ± 3.78 | 0.1332 ± 0.0429 |
| GAT | 95.75 ± 1.25 | 96.82 ± 0.88 | **98.83 ± 1.02** | 0.0318 ± 0.0088 |
| LSTM | 72.82 ± 2.36 | 81.47 ± 3.85 | 85.61 ± 3.29 | 0.1853 ± 0.0385 |
| Transformer | 85.82 ± 3.45 | 90.98 ± 1.31 | 94.05 ± 2.34 | 0.0902 ± 0.0131 |
| Gcnformer | 87.49 ± 1.28 | 84.37 ± 0.98 | 85.88 ± 1.05 | 0.1563 ± 0.9800 |
| GatLSTM | 88.72 ± 1.35 | 90.65 ± 1.19 | 91.36 ± 0.78 | 0.0935 ± 1.1900 |
| Gatformer | **98.25 ± 0.25** | **99.36 ± 0.16** | 97.65 ± 0.42 | **0.0064 ± 0.0016** |

Bold indicates the best results

each convolution in GCN is identical, which is more dependent on the structure of the prior graph with greater limitations. The GAT introduces the attention mechanism to replace the static normalized convolution operation of GCN. It can assign different importance to the edges between nodes, thus helping the model to learn structural information better.

In the classification experiments of epileptic EEG data, extensive studies have been conducted using the LSTM and its improved models. By capturing the long-term and short-term dependencies of sequences through various gate operations, the LSTM is able to reduce the gradient disappearance and gradient explosion phenomenon. The Transformer uses positional embedding to understand the sequence order and performs computation based on the self-attentive mechanism. In this experiment, compared with the LSTM, the accuracy and sensitivity of the Transformer are increased by 13% and 8.44%, which are quite good results.

We constructed the Gcnformer and GatLSTM by replacing different layers in the Gatformer. Each of these combined models can focus on the spatiotemporal correlation of EEG signals in various degrees. The difference is that the Gatformer has a multilayer attention mechanism. As a result, the correlation between the electrode nodes is significantly integrated into the model. As we can see from Table 3, the proposed model outperforms in various metrics. Meanwhile, it has better stability in the fivefold cross-validation experiments. Although the GAT had been well performed in seizure prediction tasks, the Gatformer provides a higher improvement in accuracy with an average value of 98.25%.

## Discussion

In the previous section, we have given the results of the proposed model and the baseline models to demonstrate the superior performance of the Gatformer. In the next section, we will analyze the model from different perspectives, such as stability, generalization and high-dimensional visualization to prove the comprehensive performance of the Gatformer. Meanwhile, to facilitate comparison with other similar studies, we have applied the present model to the public dataset CHB-MIT for experiments.

### Exploration of model stability

If the system raises an alarm but no seizures occur in the SOP, it is called a false alarm. False alarms can be costly and lead to a waste of medical resources. Filtering the labels of EEG segments to refine the model output can reduce misdiagnosis at a lower cost. Researchers used different post-processing methods on models to reduce the FPR of seizure prediction [44–46].

We used the $k$ of $n$ [47] method to smooth the output of the proposed model. When at least $k$ of the $n$ EEG segments are in preictal states, an alarm is sounded to alert the patient of an impending seizure. In this paper, we set $k = 25$ and $n = 50$. To be specific, out of a group of 75 s of EEG recordings, at least 37.5 s of EEG segments are classified as preictal before this group is classified as preictal. Conversely, the group is considered to be normal EEG signals of the interictal period. The results processed by this strategy are shown in Fig. 3. As we can see, the raw output results produces a lot of false positive jitter. After the results are processed by the voting mechanism, the false positives are significantly reduced and the results are closer to the ground truth.

### Analysis of generalization performance

To validate the generalization performance of the model, we performed experiments on the generalized dataset using the model trained from the general dataset. All the results are reported in Table 4. It can be seen that the sensitivity and accuracy are higher than 90% for all patients, and the zero FPR is achieved on two patients, in particular. In fact, the experimental results may not be optimal because the training dataset and the testing dataset are from different patients. However, the excellent results demonstrated that the Gatformer has good

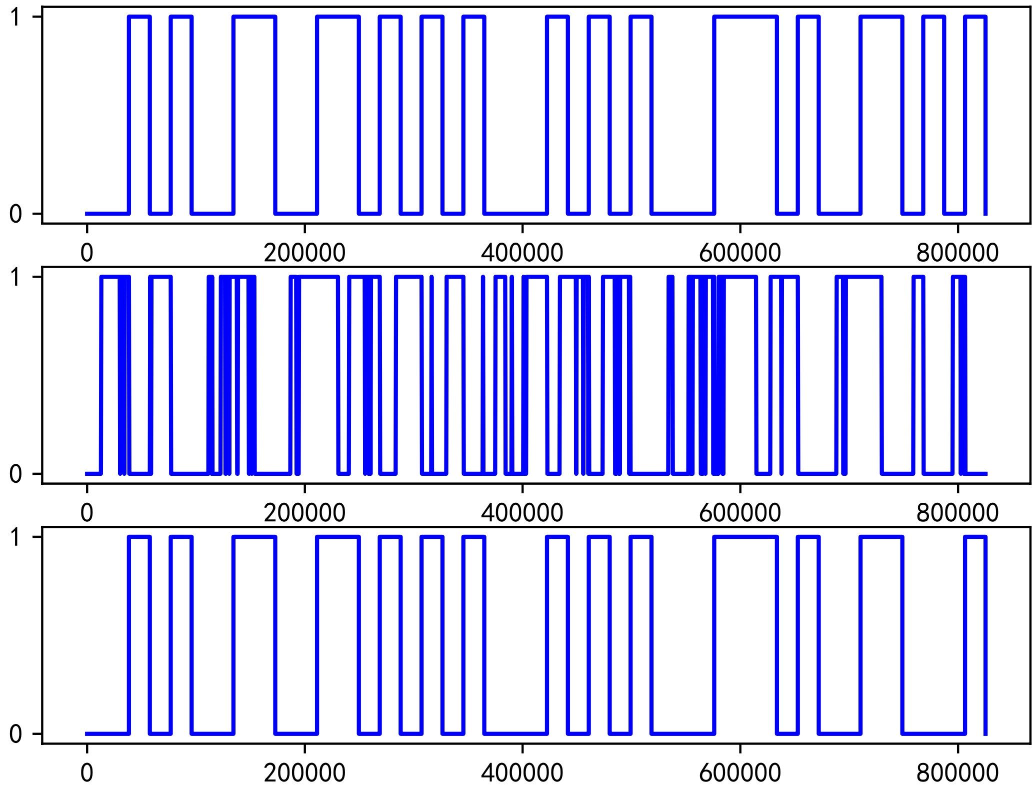


**Fig. 3** Example of vote based smoothing (Top: ground truth; Middle: model output; Bottom: smoothed result)

**Table 4 Results of the proposed algorithm on the general and the generalized datasets**

| Patient | Accuracy (%) | Specificity (%) | Sensitivity (%) | FPR (/h) |
|---|---|---|---|---|
| General dataset | 98.25 | 99.36 | 97.65 | 0.0064 |
| 10 | 94.55 | 100.00 | 95.84 | 0.0000 |
| 11 | 90.76 | 96.58 | 91.28 | 0.0342 |
| 12 | 93.78 | 97.94 | 94.66 | 0.0206 |
| 13 | 95.28 | 100.00 | 95.68 | 0.0000 |

generalization ability and can provide a reliable seizure prediction.

### Comparison with existing studies

In addition to the comparison with the baseline models, we conducted experiments on the public dataset CHB-MIT and compared our results with other similar seizure prediction algorithms reported in the recent literatures. We extracted the EEG data from the CHB-MIT dataset according to the rules mentioned in Sect. 3.2.

**Table 5 Comparison with researches in the same category**

| Authors | Methods | Channel number | Accuracy (%) | Sensitivity (%) | FPR (/h) |
|---|---|---|---|---|---|
| Zhang et al. [44] | CNN | 23 | 90.00 | 92.00 | 0.1200 |
| Tang et al. [48] | Mv-CGRN | 24 | – | 94.50 | 0.1180 |
| Abdelhameed and Bayoumi [49] | SCVAE | 12 | – | 94.45 | 0.0600 |
| Singh and Malhotra [50] | CNN | 24 | 97.40 | 98.00 | 0.0270 |
| Dissanayake et al. [51] | GDL | 23 | 95.38 | 94.47 | – |
| **This study** | **Gatformer** | **23** | **98.34** | **98.49** | **0.0108** |

Bold indicates the results obtained by our proposed strategy

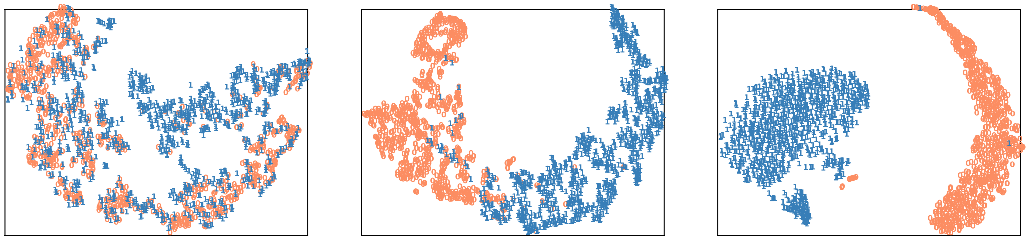

**Fig. 4** Feature distribution maps of different models (Left: Gcnformer; Middle: GatLSTM; Right: Gatformer)

The 66 seizures recorded from 23 patients were used in this experiment. The results are displayed in Table 5. It is worth noting that different data selections may affect the experimental results. However, with regard to different performance measures, our proposed method is generally comparable to advanced models.

### Analysis of feature visualization

To compare the feature extraction capabilities of different models, we visualized the classification results using the t-SNE [52]. Figure 4 provides a visual representation of the separation degree of Gcnformer, GatLSTM and Gatformer for interictal and preictal EEG data. Orange represents preictal EEGs and blue denotes interictal samples. All these three models are able to process EEG information in terms of temporal and spatial correlation. The GAT and Transformer layers of Gatformer are able to capture global EEG features using dual attention mechanisms in temporal and spatial terms, thus showing better separation effects. It can be seen that the present model enables a clear separation of the two types of EEG data, revealing its feasibility for seizure prediction.

## Conclusion

The scalp EEG signals contain rich temporal and spatial information. In this paper, we explored these information and designed a seizure prediction model using the dual attention mechanism. The Pearson adjacency matrix and EEG signals were input into the GAT module to understand the correlation between multi-channel electrodes. The attention mechanism was used to sufficiently focus on the neighbor nodes and capture the spatial relationship of EEGs. Further processing was later performed using Transformer to focus on the attention allocation of input signals at different moments. Extensive experiments on different datasets showed the superior performance of the present model in all aspects. The sensitivity of 97.65% and the FPR of 0.0064/h demonstrated the feasibility of the Gatformer in seizure prediction. The present study provides an effective choice for the clinical practice of seizure prediction.

Although the attention mechanism in our model is able to automatically learn key information in EEG and grasp the spatiotemporal features carried by EEG information, the model still has certain limitations that make it difficult to fully understand the classification mechanism of the model. In future research, the introduction of interpretable learning methods will be considered for further exploration to meet the demand of clinical requirements for interpretability.

**Funding**

This work was supported by the Key Research and Development Program of China under Grant 2022YFC3601600 and 2021YFC2009400, in part by the National Natural Science Foundation of China (NSFC) under Grant 61876194, in part by the Province Natural Science Foundation of Guangdong under Grant 2021A1515011897, in part by the Key Research and Development Program of Guangzhou under Grant 202206010028, in part by the Fundamental Research Funds for the Central Universities, Sun Yat-sen University 23ptpy119.

## Declarations

**Conflict of interest**

The authors confirm that there are no conflict of interest.

**Ethical approval**

The private dataset of this study involving human participants were reviewed and approved by Medical Ethics Committee of First Affiliated Hospital, Sun Yat-sen University. The patients provided their written informed consent to participate in this study.

**Author details**

[1]School of Biomedical Engineering, Sun Yat-sen University, Guangzhou 510006, Guangdong, China. [2]Department of Medical Informatics, Zhongshan School of Medicine, Sun Yat-sen University, Guangzhou 510080, Guangdong, China. [3]Department of Neurology, First Affiliated Hospital, Sun Yat-sen University, Guangzhou 510080, Guangdong, China.

Received: 12 November 2022 Accepted: 31 July 2023
Published: 23 August 2023

**References**

1. Assi EB, Nguyen DK, Rihana S, Sawan M. Towards accurate prediction of epileptic seizures: a review. Biomed Signal Process Control. 2017;34:144–57.

2. Calle-López Y, Ladino LD, Benjumea-Cuartas V, Castrillón-Velilla DM, Téllez-Zenteno JF, Wolf P. Forced normalization: a systematic review. Epilepsia. 2019;60(8):1610–8.
3. Zhang Z, Chen Z, Zhou Y, Du S, Zhang Y, Mei T, Tian X. Construction of rules for seizure prediction based on approximate entropy. Clin Neurophysiol. 2014;125(10):1959–66.
4. Zandi AS, Tafreshi R, Javidan M, Dumont GA. Predicting epileptic seizures in scalp EEG based on a variational Bayesian Gaussian mixture model of zero-crossing intervals. IEEE Trans Biomed Eng. 2013;60(5):1401–13.
5. Li S, Zhou W, Yuan Q, Liu Y. Seizure prediction using spike rate of intracranial EEG. IEEE Trans Neural Syst Rehabil Eng. 2013;21(6):880–6.
6. Alotaiby TN, Alshebeili SA, Alotaibi FM, Alrshoud SR. Epileptic seizure prediction using CSP and LDA for scalp EEG signals. Comput Intell Neurosci. 2017;2017:1240323.
7. Affes A, Mdhaffar A, Triki C, Jmaiel M, Freisleben B. A convolutional gated recurrent neural network for epileptic seizure prediction. In: International conference on smart homes and health telematics. Cham: Springer; 2019. pp. 85–96.
8. Xu Y, Yang J, Sawan M. Multichannel synthetic preictal EEG signals to enhance the prediction of epileptic seizures. IEEE Trans Biomed Eng. 2022;69(11):3516–25.
9. Singh K, Malhotra J. Prediction of epileptic seizures from spectral features of intracranial EEG recordings using deep learning approach. Multimed Tools Appl. 2022;81:1–24.
10. Li F, Liang Y, Zhang L, Yi C, Liao Y, Jiang Y, Si Y, Zhang Y, Yao D, Yu L, et al. Transition of brain networks from an interictal to a preictal state preceding a seizure revealed by scalp EEG network analysis. Cogn Neurodyn. 2019;13(2):175–81.
11. Wei X, Zhou, Y. A methodical approach to epileptic classification with multi-scale patterns. In: Proceedings of the 2018 5th International conference on biomedical and bioinformatics engineering, 2018, pp. 25–29.
12. Ma M, Cheng Y, Wei X, Chen Z, Zhou Y. Research on epileptic EEG recognition based on improved residual networks of 1-D CNN and indRNN. BMC Med Inform Decis Mak. 2021;21(2):1–13.
13. Parvez MZ, Paul M. Seizure prediction using undulated global and local features. IEEE Trans Biomed Eng. 2016;64(1):208–17.
14. He Z, Zhong Y, Pan J. Joint temporal convolutional networks and adversarial discriminative domain adaptation for EEG-based cross-subject emotion recognition. In: ICASSP 2022—2022 IEEE international conference on acoustics, speech and signal processing (ICASSP), 2022. IEEE; 2022. p. 3214–8.
15. Howlader KC, Satu M, Awal M, Islam M, Islam SMS, Quinn JM, Moni MA, et al. Machine learning models for classification and identification of significant attributes to detect type 2 diabetes. Health Inf Sci Syst. 2022;10(1):1–13.
16. Akbari H, Sadiq MT, Siuly S, Li Y, Wen P. Identification of normal and depression EEG signals in variational mode decomposition domain. Health Inf Sci Syst. 2022;10(1):1–14.
17. Wei X, Zhou L, Zhang Z, Chen Z, Zhou Y. Early prediction of epileptic seizures using a long-term recurrent convolutional network. J Neurosci Methods. 2019;327: 108395.
18. Usman SM, Khalid S, Aslam MH. Epileptic seizures prediction using deep learning techniques. IEEE Access. 2020;8:39998–40007.
19. Daoud H, Bayoumi MA. Efficient epileptic seizure prediction based on deep learning. IEEE Trans Biomed Circuits Syst. 2019;13(5):804–13.
20. Jana R, Mukherjee I. Deep learning based efficient epileptic seizure prediction with EEG channel optimization. Biomed Signal Process Control. 2021;68: 102767.
21. Zhao S, Yang J, Xu Y, Sawan M. Binary single-dimensional convolutional neural network for seizure prediction. In: 2020 IEEE international symposium on circuits and systems (ISCAS), 2020.
22. Jiang Y, Lu Y, Yang L. An epileptic seizure prediction model based on a time-wise attention simulation module and a pretrained ResNet. Methods. 2022;202:117–26.
23. Sun B, Lv J-J, Rui L-G, Yang Y-X, Chen Y-G, Ma C, Gao Z-K. Seizure prediction in scalp EEG based channel attention dual-input convolutional neural network. Physica A. 2021;584: 126376.
24. Ma M, Cheng Y, Wang Y, Li X, Mao Q, Zhang Z, Chen Z, Zhou Y. Early prediction of epileptic seizure based on the BNLSTM-CASA model. IEEE Access. 2021;9:79600–10.
25. Yang X, Zhao J, Sun Q, Lu J, Ma X. An effective dual self-attention residual network for seizure prediction. IEEE Trans Neural Syst Rehabil Eng. 2021;29:1604–13.
26. Zhao Y, Dong C, Zhang G, Wang Y, Chen X, Jia W, Yuan Q, Xu F, Zheng Y. EEG-Based Seizure detection using linear graph convolution network with focal loss. Comput Methods Programs Biomed. 2021;208: 106277.
27. He J, Cui J, Zhang G, Xue M, Chu D, Zhao Y. Spatial–temporal seizure detection with graph attention network and bi-directional LSTM architecture. Biomed Signal Process Control. 2022;78: 103908.
28. Chen X, Zheng Y, Dong C, Song S. Multi-dimensional enhanced seizure prediction framework based on graph convolutional network. Front Neuroinform. 2021;15: 605729.
29. Veličković P, Cucurull G, Casanova A, Romero, A, Lio P, Bengio Y. Graph attention networks. arXiv preprint; 2017. arXiv:1710.10903.
30. Zhao Y, Zhang G, Dong C, Yuan Q, Xu F, Zheng Y. Graph attention network with focal loss for seizure detection on electroencephalography signals. Int J Neural Syst. 2021;31(07):2150027.
31. Sartipi S, Torkamani-Azar M, Cetin M. EEG emotion recognition via graph-based spatio-temporal attention neural networks. In: 43rd Annual international conference of the IEEE Engineering in Medicine & Biology Society (EMBC), 2021. IEEE; 2021. p. 571–4.
32. Wang Y, Shi Y, Cheng Y, He Z, Wei X, Chen Z, Zhou Y. A spatiotemporal graph attention network based on synchronization for epileptic seizure prediction. IEEE J Biomed Health Inform. 2023;27(2):900–11.
33. Zhao Y, Xue M, Dong C, He J, Chu D, Zhang G, Xu F, Ge X, Zheng Y. Automatic seizure identification from EEG signals based on brain connectivity learning. Int J Neural Syst. 2022;32(11):2250050.
34. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is all you need. In: Advances in neural information processing systems, 2017, vol 30.
35. Devlin J, Chang M.-W, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. arXiv preprint, 2018. arXiv:1810.04805.
36. Zou C, Wang B, Hu Y, Liu J, Wu Q, Zhao Y, Li B, Zhang C, Zhang C, Wei Y, et al. End-to-end human object interaction detection with hoi transformer. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 11825–34.
37. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, et al. An image is worth 16 × 16 words: transformers for image recognition at scale. arXiv preprint, 2020. arXiv:2010.11929.
38. Kim D, Lee J, Woo Y, Jeong J, Kim C, Kim D-K. Deep learning application to clinical decision support system in sleep stage classification. J Pers Med. 2022;12(2):136.
39. Bhattacharya A, Baweja T, Karri S. Epileptic seizure prediction using deep transformer model. Int J Neural Syst. 2022;32(02):2150058.
40. Hu S, Liu J, Yang R, Wang YN, Wang A, Li K, Liu W, Yang C. Exploring the applicability of transfer learning and feature engineering in epilepsy prediction using hybrid transformer model. IEEE Trans Neural Syst Rehabil Eng. 2023;31:1321–32.
41. Battaglia PW, Hamrick JB, Bapst V, Sanchez-Gonzalez A, Zambaldi V, Malinowski M, Tacchetti A, Raposo D, Santoro A, Faulkner R, et al. Relational inductive biases, deep learning, and graph networks. arXiv preprint, 2018. arXiv:1806.01261.
42. He Z, Li Z, Yang F, Wang L, Li J, Zhou C, Pan J. Advances in multimodal emotion recognition based on brain–computer interfaces. Brain Sci. 2020;10(10):687.
43. Maiwald T, Winterhalder M, Aschenbrenner-Scheibe R, Voss HU, Schulze-Bonhage A, Timmer J. Comparison of three nonlinear seizure prediction methods by means of the seizure prediction characteristic. Physica D. 2004;194(3–4):357–68.
44. Zhang Y, Guo Y, Yang P, Chen W, Lo B. Epilepsy seizure prediction on EEG using common spatial pattern and convolutional neural network. IEEE J Biomed Health Inform. 2019;24(2):465–74.
45. Park Y, Luo L, Parhi KK, Netoff T. Seizure prediction with spectral power of EEG using cost-sensitive support vector machines. Epilepsia. 2011;52(10):1761–70.
46. Pedoeem J, Bar Yosef G, Abittan S, Keene S. TABS: transformer based seizure detection. In: Biomedical sensing and analysis. Berlin: Springer; 2022. p 133–160.

47. Truong ND, Nguyen AD, Kuhlmann L, Bonyadi MR, Yang J, Ippolito S, Kavehei O. Convolutional neural networks for seizure prediction using intracranial and scalp electroencephalogram. Neural Netw. 2018;105:104–11.
48. Tang L, Xie N, Zhao M, Wu X. Seizure prediction using multi-view features and improved convolutional gated recurrent network. IEEE Access. 2020;8:172352–61.
49. Abdelhameed AM, Bayoumi M. An efficient deep learning system for epileptic seizure prediction. In: IEEE international symposium on circuits and systems (ISCAS), 2021. IEEE; 2021. p. 1–5.
50. Singh K, Malhotra J. Predicting epileptic seizures from EEG spectral band features using convolutional neural network. Wirel Pers Commun. 2022;125:1–18.
51. Dissanayake T, Fernando T, Denman S, Sridharan S, Fookes C. Geometric deep learning for subject independent epileptic seizure prediction using scalp EEG signals. IEEE J Biomed Health Inform. 2021;26(2):527–38.
52. Van der Maaten L, Hinton G. Visualizing data using t-SNE. J Mach Learn Res. 2008;9(11):2579–605.