B2-ViT Net: Broad Vision Transformer Network With Broad Attention for Seizure Prediction

Shuiling Shi and Wenqi Liu

***Abstract*— Seizure prediction are necessary for epileptic patients. The global spatial interactions among channels, and long-range temporal dependencies play a crucial role in seizure onset prediction. In addition, it is necessary to search for seizure prediction features in a vast space to learn new generalized feature representations. Many previous deep learning algorithms have achieved some results in automatic seizure prediction. However, most of them do not consider global spatial interactions among channels and long-range temporal dependencies together, and only learn the feature representation in the deep space. To tackle these issues, in this study, an novel bi-level programming seizure prediction model, B2-ViT Net, is proposed for learning the new generalized spatio-temporal long-range correlation features, which can characterize the global interactions among channels in spatial, and long-range dependencies in temporal required for seizure prediction. In addition, the proposed model can comprehensively learn generalized seizure prediction features in a vast space due to its strong deep and broad feature search capabilities. Sufficient experiments are conducted on two public datasets, CHB-MIT and Kaggle datasets. Compared with other existing methods, our proposed model has shown promising results in automatic seizure prediction tasks, and provides a certain degree of interpretability.**

***Index Terms*— Automatic seizure prediction, electroencephalogram (EEG), vision transformer (ViT), multi-head self-attention, broad attention, broad learning system (BLS).**

I. Introduction

EPILEPSY is a chronic non-infectious disease caused by paroxysmal abnormal super-synchronous discharge activity of brain neurons. It is one of the most common neurological diseases worldwide and covers all age groups, around 50 million epileptic patients worldwide [1]. Epilepsy is associated with adverse outcomes, including serious comorbidities, injury and death [2]. The central problem of epilepsy is the unpredictability of seizures, which can have a persistent negative impact on patients' life.

If seizures can be predicted a few minutes before onset, patients will be able to take precautions against injury and open the door to new and timely treatment for the prevention or control of impending seizures [3]. In addition, doctors usually provide treatment plans for patients with epilepsy based on the type and number of seizure onset recorded by patients. But, epilepsy data recorded by patients and their caregivers are often unreliable. It takes a lot of time and energy for doctors to detect seizures from long-term electroencephalogram (EEG) records. To make effective treatment plans, it is necessary to use seizure prediction algorithms to identify seizure events automatically. Therefore, an automatic seizure prediction algorithm is vitally important for patients with epilepsy. EEG is generated by synchronous activity of a large number of neurons in the brain, which is consistent with the super-synchronous discharge mechanism of epilepsy, so EEG is an indispensable source of data for predicting seizures. These seizure prediction algorithms usually have two main functions: (1) They can be integrated into wearable technology and combined with an online alarm system to start therapeutic interventions [4], [5]. (2) It can assist medical workers in reviewing offline long-term EEG records to detect seizures automatically [6].

A complete seizure often includes interictal, preictal, ictal and postictal [7], [8]. The seizure prediction tasks can be simplified as a classification of interictal and preictal. When a certain amount of preictal data is predicted, it can provide early warning for the impending seizure onset.

In recent years, deep learning algorithms have attracted extensive attention in various fields because of their great generalization ability and more automatic feature extraction ability, encouraging their application in the field of seizure prediction. Truong et al. [9] used short-time fourier transform (STFT) to extract EEG features from the original EEG signals and used convolutional neural network (CNN) [10] to classify the interictal and preictal. Ozcan and Erturk [11] extracted spectral band power, statistical moment and hjorth parameters to reveal the frequency and time domain features of the EEG signals. The features are given as input to a 3D CNN [12]. Daoud and Bayoumi [13] used deep convolutional neural network (DCNN) and concatenated with a bidirectional long short-term memory (Bi-LSTM) network as the back-end of model to classify.

Manuscript received 15 August 2023; revised 6 December 2023; accepted 21 December 2023. Date of publication 25 December 2023; date of current version 16 January 2024. This work was supported by the National Natural Science Foundation of China under Grant 12371460. *(Corresponding author: Wenqi Liu.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Massachusetts Institute of Technology Division of Health Sciences and Technology, and Kaggle database.

The authors are with the Data Science Research Center, Kunming University of Science and Technology, Kunming 650500, China, and also with the Faculty of Science, Kunming University of Science and Technology, Kunming 650500, China (e-mail: shishuiling0409@sina.com; liuwenq2215@sina.com).

Digital Object Identifier 10.1109/TNSRE.2023.3346955

Many studies have shown that seizures involve not only the seizure onset zone and its surroundings, but also the brain areas far away from seizure onset zone [3], [14]. Abnormal interactions among different brain areas may lead to seizure onset. To characterize interactions among different brain areas within a whole-brain range, recent studies generally construct brain functional connectivity networks based on scalp EEG using channels as nodes [15], [16], [17]. According to the international standard electrode positions, in multi-channel EEG data, different channels correspond to different brain regions, so abnormal interactions among different brain regions can be reflected by the interactions among different channels. Furthermore, seizures do not occur randomly and have been shown to have long-range temporal dependencies [3], [18], [19]. In summary, the global channel interactions in spatial, and long-range temporal dependencies are crucial to seizure prediction algorithms. However, most of the previous traditional deep learning algorithms, such as CNN, they can only capture local channel interactions in spatial and short-range temporal dependencies due to the regular and local receptive field of convolution operators, without considering global channel interactions and long-range temporal dependencies together, resulting in the lack of interpretability of the model and the common results.

In fact, vision transformer (ViT) [20] algorithm based on global attention mechanism can achieve the global channel interaction in spatial, and obtain long-range temporal dependence features required for seizure prediction. But ViT only considers the deep features of the last transformer modules, transformer modules with different depths may contain complementary features related to seizure prediction tasks [21]. The complementary features can be obtained through the broad connection of shallow and deep transformer modules. But these complementary features are redundant and complicated. By applying attention mechanisms to these complementary features, we can further extract critical spatio-temporal long-range correlation complementary features that are beneficial to seizure prediction. However, the broad connection above is only used for the attention mechanism part to extract connected attention information from different transformer modules, instead of mapping all features together into a new vast space to learn new generalized features. It is necessary to search for seizure prediction features in a vast space, so as to learn new generalized spatio-temporal long-range correlation features that help predict seizures [22].

Therefore, according to the neuroscience mechanism of seizure, a novel bi-level programming seizure prediction model, broad vision transformer network with broad attention, called B2-ViT Net, is proposed for learning the new generalized spatio-temporal long-range correlation features, which can characterize the global channel interaction features in spatial and long-range dependence features in temporal, captures generalized features that are beneficial to seizure prediction, thus improving the prediction performance. Compared with other black box deep learning models, our model can quantify the interaction weights among channels, and evaluate the importance of each channel at any time, thus providing a certain degree of interpretability.

Specifically, the contributions of our proposed method can be summarized in the following aspects.

1) Based on the neuroscience mechanism of seizure onset, we proposed a novel bi-level programming seizure prediction model B2-ViT Net, which considers the global spatial interactions among channels and long-range temporal dependencies together through the global attention mechanism, called spatio-temporal long-range correlations. The global attention mechanism here can innovatively quantify the interaction weights among channels, and evaluate the importance of each channel at any time.

2) Both deep and broad features are crucial for seizure prediction tasks. Previous seizure prediction algorithms only focused on deep features while ignoring the generalized features that combine deep and broad. Generalized features are characterized through linear and nonlinear random mappings in our model. Our proposed model can comprehensively learn generalized spatio-temporal long-range correlation features that are conducive to automatic seizure prediction in a vast space, improve the prediction performance.

3) Sufficient experiments are conducted on two public datasets, CHB-MIT and Kaggle datasets. Compared with other existing methods, our proposed method has achieved promising results in automatic seizure prediction tasks, obtains the highest AUC and the lowest FPR. On CHB-MIT dataset, B2-ViT obtains 0.923, 93.3%, and 0.057/h on AUC, sensitivity and FPR, respectively. On the Kaggle dataset, the proposed model reached 0.816, 85.2%, and 0.013/h on AUC, sensitivity and FPR, respectively.

II. Preliminary Knowledge

This section introduces the preliminary knowledge of ViT and BLS, which helps to build B2-ViT Net.

A. ViT: Vision Transformer

Transformer is a deep neural network mainly based on self-attention mechanism, which is initially applied in natural language processing. Inspired by its powerful global presentation ability, researchers extend transformer to computer vision tasks, which is called ViT [20]. Compared with other networks (such as CNN), the model shows competitive performance on various benchmarks. The model follows the following steps: (1) Converting image data to sequences form as transformer input; (2) applying linear projection to the sequences; (3) adding extra learnable classification token, adding positional embedding; (4) a transformer encoder is applied to the processed data, which mainly includes multi-head self-attention mechanism (MHSA) block and multi-layer perceptron (MLP) block.

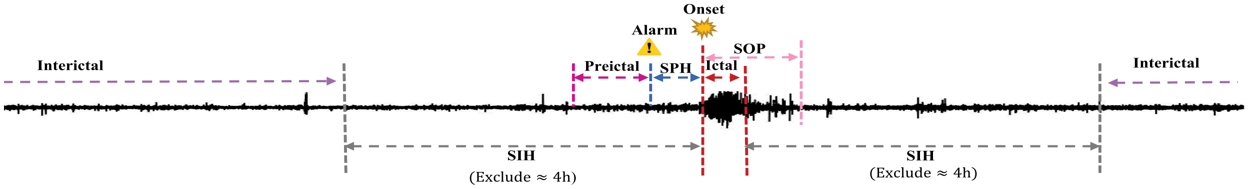

Fig. 1. Definition of interictal, preictal, SIH, SPH, SOP and seizure period (from the file *chb***01_03**.edf).

TABLE I
SUMMARY OF CHB-MIT DATASET

Patient ID	Age	Gender	No. of seizures	Interical hours	Preical hours	Patient ID	Age	Gender	No. of seizures	Interical hours	Preical hours
Pat 1	11	F	7	17	3.5	Pat 14	9	F	5	5	2.5
Pat 2	11	M	3	23	1.5	Pat 18	18	F	6	24	3.0
Pat 3	14	F	6	22	3.0	Pat 19	19	F	3	25	1.5
Pat 5	7	F	5	14	2.5	Pat 20	6	F	5	20	2.5
Pat 9	10	F	4	46.3	2.0	Pat 21	13	F	4	22	2.0
Pat 10	3	M	6	26	3.0	Pat 23	6	F	5	12.9	2.5
Pat 13	3	F	5	14	2.5	Total	-	-	64	271.2	32

TABLE II
SUMMARY OF KAGGLE DATASET

Patient ID	Age	Gender	No. of seizures	Interical hours	Preical hours	Patient ID	Age	Gender	No. of seizures	Interical hours	Preical hours
Dog 1	-	-	4	80.0	2.0	Dog 4	-	-	14	134	8.5
Dog 2	-	-	7	83.3	3.5	Dog 5	-	-	5	75	2.5
Dog 3	-	-	12	240	6.0	Total	-	-	42	612.3	22.5

B. BLS: Broad Learning System

BLS [23] has a strong ability to search broad features. It consists mainly of feature nodes and enhancement nodes. The feature nodes are obtained by a random mapping, and then the feature nodes are mapped to a possible high-dimensional vector space to obtain enhancement nodes, so that the model can automatically search features related to specific tasks in a vast vector space. Both two features yield the final output.

III. DATASETS AND METHODOLOGY

This section thoroughly introduces the datasets, data preprocessing, the modeling method of B2-ViT Net, and postprocessing. In addition, the model frame diagram and algorithm table are also provided. The structure of B2-ViT is shown in Fig. 2, the detailed implementation steps of B2-ViT are summarized in Algorithm 1.

A. Datasets

1) CHB-MIT Dataset: The CHB-MIT seizure EEG dataset [24] is obtained from Boston Children's Hospital and included in the EEG database of the Massachusetts Institute of Technology. It contains 23 records from 22 subjects (chb21 is recorded again of chb01 subjects after 1.5 years). Each subject has 9-24 recordings lasting for 1 hour (some of which are long records of 2-4 hours), and the dataset includes 884 hours of continuous scalp EEG recordings and 163 seizures. All EEG data are collected using 10-20 international standard electrode positions, EEG is recorded using 18/23 lead, and the sampling frequency is 256 Hz.

2) Kaggle Dataset: The American Epilepsy Society Seizure Prediction Challenge of Kaggle dataset [25] has iEEG data from 5 dogs and 2 patients, with 48 seizures and 627.7 hours interictal records, which is simply denoted as Kaggle dataset. Intracranial EEG (iEEG) data of 5 dogs are recorded from 16 implanted electrodes, and the sampling rate is 400 Hz. Recorded iEEG data of 2 patients from 15 deep electrodes (Patient 1) and 24 subdural electrodes (Patient 2), and the sampling rate is 5 kHz. The calculation is difficult due to the patients' high sampling rate of the Kaggle dataset, so the two patients' iEEG data are not considered, which is consistent with [9], [26], [27]. These two datasets are used in most seizure prediction tasks [8], [9], [26], [28].

B. Preprocessing

As shown in Fig. 1, a complete seizure can be divided into preictal, interictal, seizure interictal horizon (SIH), seizure prediction horizon (SPH), and seizure occurrence period (SOP). SPH is the prediction period before the seizure, during which appropriate measures can be used to prevent or control the impending seizure in advance. SOP is the interval where the seizure is expected to occur. SIH is defined as EEG signals about 4 hours before and 4 hours after the seizure [29], which can reduce the interference caused by the near seizure state. To predict correctly, seizures must be after SPH and within SOP. This paper follows the definitions of SOP and SPH proposed by [30]. In this work, SPH is set to 5 minutes and SOP is set to 30 minutes, which is consistent with most studies. CHB-MIT dataset has many seizures in a short time. For seizures less than 30 minutes from the previous seizure, we assume that there are only the leading seizure exists. In addition, this work only considers patients with seizures less than 10 times a day, because it is not very necessary to perform this task for patients who have seizures every 2 hours on average. Based on the above definition and consideration, this work evaluated 64 seizures in the CHB-MIT dataset and 42 seizures of 5 dogs in the Kaggle dataset. These two datasets' available data are summarized in Table I and Table II.

Classification tasks often face the problem of class imbalance, automatic seizure prediction tasks are no exception, interictal data is far more than preictal data. To solve this problem, the sliding overlap technique with step size s is used to obtain more preictal data. the number of extra preictal data N after oversampling is computed as:

$$N = \frac{(P - w)}{s} + 1 \tag{1}$$

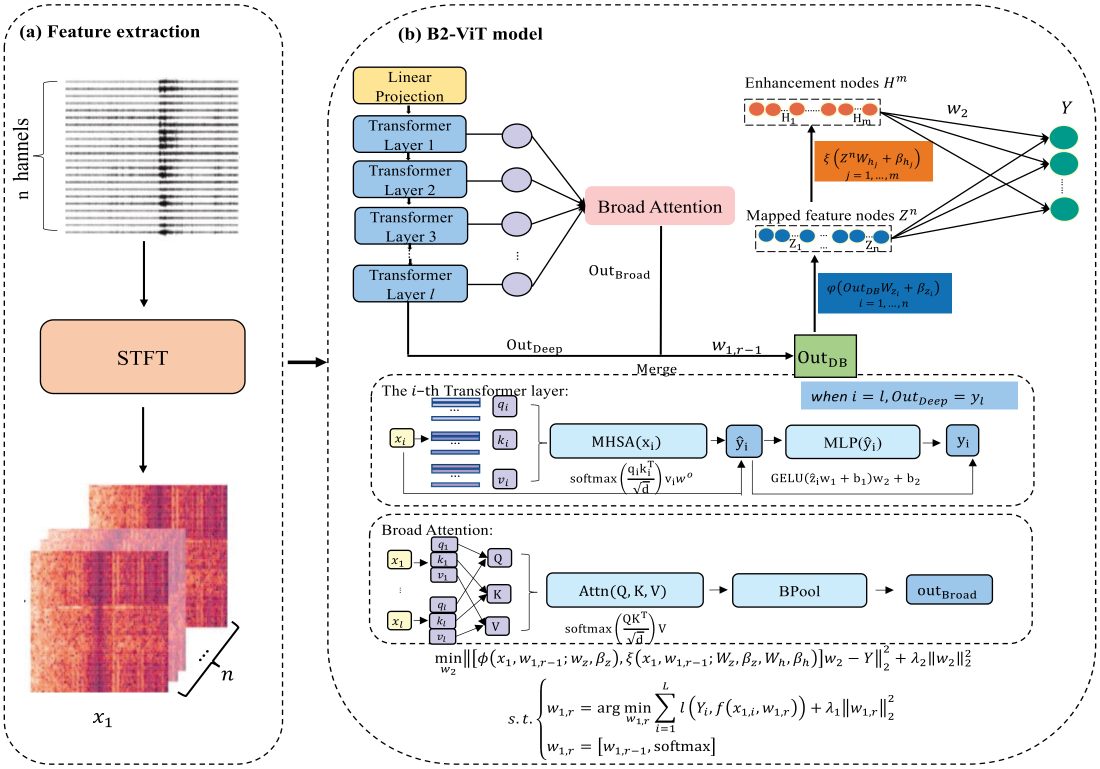
(a) Feature extraction
n hannels
STFT
x_1
n
(b) B2-ViT model
Linear Projection
Transformer Layer 1
Transformer Layer 2
Transformer Layer 3
Transformer Layer l
Broad Attention
$\mathrm{Out}_{\mathrm{Broad}}$
$\mathrm{Out}_{\mathrm{Deep}}$
Merge
$w_{1,r-1}$
$\mathrm{Out}_{\mathrm{DB}}$
Enhancement nodes H^m
H_1
H_m
$\xi\left(Z^n W_{h_j}+\beta_{h_j}\right)$ $j=1,\ldots,m$
Mapped feature nodes Z^n
Z_1
Z_n
$\varphi\left(Out_{DB} W_{z_i}+\beta_{z_i}\right)$ $i=1,\ldots,n$
w_2
Y
The i–th Transformer layer:
$when\ i=l, Out_{Deep}=y_l$
x_i
q_i
k_i
v_i
$\mathrm{MHSA}(\mathrm{x_i})$
$\mathrm{softmax}\left(\frac{q_i k_i^T}{\sqrt{d}}\right) v_i w^o$
$\hat{y}_i$
$\mathrm{MLP}(\hat{y}_i)$
$\mathrm{GELU}(\hat{z}_i w_1+b_1)w_2+b_2$
y_i
Broad Attention:
x_1
x_l
q_1
k_1
v_1
q_l
k_l
v_l
Q
K
V
Attn(Q, K, V)
$\mathrm{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V$
BPool
$\mathrm{out}_{\mathrm{Broad}}$
$\min_{w_2}\left\|\left[\phi\left(x_1, w_{1,r-1}; w_z, \beta_z\right), \xi\left(x_1, w_{1,r-1}; W_z, \beta_z, W_h, \beta_h\right)\right] w_2 - Y\right\|_2^2 + \lambda_2\|w_2\|_2^2$
$s.t.\begin{cases} w_{1,r}=\arg\min_{w_{1,r}} \sum_{i=1}^{L} l\left(Y_i, f\left(x_{1,i}, w_{1,r}\right)\right)+\lambda_1\|w_{1,r}\|_2^2 \\ w_{1,r}=[w_{1,r-1}, \mathrm{softmax}] \end{cases}$

Fig. 2. The architecture of B2-ViT Net. (a) The feature pre-extraction part using STFT. (b) The architecture of bi-level programming B2-ViT model.

where w is the sliding window length, P is the total length of preictal data, I is the total length of interictal data, R is the ratio of the total length of preictal data to the total length of interictal data and $s = w \times R$.

In this paper, STFT [31] is used to preprocess the raw EEG data to extract time-frequency features, which converts the original EEG signal into a time-frequency matrix. The window length of STFT is 30s. STFT is chosen because it can capture the dynamic changes of the frequency characteristics of EEG signals of epileptic patients, and compared with wavelet transform (WT) [32] and other signal analysis methods, it has a shorter processing time of time series, which is helpful for real-time seizure prediction. Besides, it is widely used in EEG processing, retains most of the information in the original signal, and many studies have shown its advantages in EEG [9], [33]. The datasets used are contaminated with 60 Hz power line noise, so components in the 57-63 Hz and 117-123 Hz frequency ranges are excluded to eliminate power line interference, and the DC component (0 Hz) is also removed.

C. Proposed Method

B2-ViT Net is a novel bi-level programming problem for seizure prediction. It considers the spatio-temporal long-range correlation features required for seizure prediction. In addition, it has strong global deep and broad feature search capabilities, which can comprehensively learn generalized spatio-temporal long-range correlation features that are conducive to automatic seizure prediction in a vast space, thus improving the prediction performance.

For a given preprocessed image $\mathbf{I} \in R^{L\times C_1\times W}$, L is the length of sequence, C_1 and W are the number of channels and width of image patches, which can be processed directly by the standard transformer. To get the input $x_1 \in R^{L\times C\times D}$ of the first transformer layer, linear projection is adopted for satisfying the required dimension D of transformer, C_1 is additionally added with classification token, which is recorded as C. After processing the input data, the model is first divided into two parts, one is ViT backbone to obtain deep features Out_{Deep}, and the other is broad attention to obtain local broad features Out_{Broad}. The transformer layer includes two blocks: MHSA and MLP. In addition, residual connections are used in MHSA and MLP blocks, and LayerNorm (LN) is applied before each block. Next, the calculation process of MHSA, MLP and broad attention is introduced in detail.

Multi-Head Self-Attention: Given the input $x_i \in R^{L\times C\times D}$ of i-th layer. Then query $q_i \in R^{L\times C\times(h\times d_q)}$, key $k_i \in R^{L\times C\times(h\times d_k)}$ and value $v_i \in R^{L\times C\times(h\times d_v)}$ are obtained by chunking x_i into three tensors and rearranging them, h is the number of head, where d_q, d_k and d_v are the dimension of q_i, k_i and v_i, respectively, $i \in [1, l]$, where l is the number of transform layers. Then inner product, softmax and the second linear projection are performed. The output of MHSA can be obtained by the following:

$$\begin{aligned}\mathrm{MHSA}(x_i) &= \mathrm{softmax}(q_i, k_i, v_i)w^o \\ &= \mathrm{softmax}(\frac{q_i k_i^T}{\sqrt{d_q}})v_i w^o \end{aligned} \tag{2}$$

where

$$q_i = [[q_i^1], [q_i^2], \ldots, [q_i^h]], q_i^j \in R^{L\times C\times d_q}$$
$$k_i = [[k_i^1], [k_i^2], \ldots, [k_i^h]], k_i^j \in R^{L\times C\times d_k}$$

Algorithm 1 Implementation Process of the B2-ViT Algorithm

```
Require: Input seizure EEG data X and label $\mathbf{Y_T}$;
Ensure: Prediction matrix $\mathbf{Y_P}$ for for seizure detection;
Parameters: $\mathbf{W}_{\mathrm{p1}}, \mathbf{b}_{\mathrm{p1}}, \mathbf{W}_{\mathrm{p2}}, \mathbf{b}_{\mathrm{p2}}$: Linear projection parameters; $\mathbf{W}_c$, $\mathbf{W}_p$: the class token and positional embedding matrices; $\mathbf{W}_i$: multi-head attention parameters for layer $i$; $\boldsymbol{\gamma}_{\mathrm{i}}^1, \boldsymbol{\beta}_{\mathrm{i}}^1, \boldsymbol{\gamma}_i^2, \boldsymbol{\beta}_i^2$: two sets of layer-norm parameters for layer $i$; $d$: the dim of one head; $\mathbf{W}_{11}^i, \mathbf{b}_{11}^i, \mathbf{W}_{21}^i, \mathbf{b}_{21}^i$: MLP parameters for layer $i$; $l$: the depth of transformer block; $\boldsymbol{\gamma}$: the coefficient factor; the regularization coefficient of BLS $\lambda_1$.
Fed X into the STFT to obtain the initial features $\mathbf{X}_s$;
$\mathbf{X}_p \leftarrow \mathbf{W}_{\mathrm{p2}}$ GELU($\mathbf{W}_{\mathrm{p1}}\mathbf{X}_s$+$\mathbf{b}_{\mathrm{p1}}\mathbf{1}^T$)+$\mathbf{b}_{\mathrm{p2}}\mathbf{1}^T$
$\mathbf{x}_1 \leftarrow$ cat($\mathbf{W}_c$, $\mathbf{X}_p$) + $\mathbf{W}_p$
for $i = 1, 2, \ldots, l$ do
    MHSA($\mathbf{x}_i$), $\boldsymbol{q}_i$, $\mathbf{k}_i$, $\mathbf{v}_i \leftarrow$ MHSAttention($\mathbf{x}_i|\mathbf{W}_i$)
    $\hat{\mathbf{y}}_i \leftarrow \mathbf{x}_i$ + MHSA($\mathbf{x}_i$)
    for $t \in [i]$: $\hat{\mathbf{y}}_i[:, :, t] \leftarrow$ layer_norm($\hat{\mathbf{y}}_i[:, :, t]|\boldsymbol{\gamma}_i^1, \boldsymbol{\beta}_i^1$)
    $\mathbf{y}_i \leftarrow \hat{\mathbf{y}}_i$ + $\mathbf{W}_{21}^i$GELU($\mathbf{W}_{11}^i\mathbf{x}_i$ + $\mathbf{b}_{11}^i\mathbf{1}^T$) + $\mathbf{b}_{21}^i\mathbf{1}^T$
    for $t \in [i]$: $\mathbf{y}_i[:, :, t] \leftarrow$ layer_norm($\hat{\mathbf{y}}_i[:, :, t]|\boldsymbol{\gamma}_i^2, \boldsymbol{\beta}_i^2$)
    $\mathbf{x}_{i+1} = \mathbf{y}_i$
end for
$\mathbf{Out}_{Deep} \leftarrow \mathbf{y}_l$
$\mathbf{Q} \leftarrow \boldsymbol{q}_1 \cup \boldsymbol{q}_2 \cup \ldots \cup \boldsymbol{q}_l$
$\mathbf{K} \leftarrow \boldsymbol{k}_1 \cup \boldsymbol{k}_2 \cup \ldots \cup \boldsymbol{k}_l$
$\mathbf{V} \leftarrow \boldsymbol{v}_1 \cup \boldsymbol{v}_2 \cup \ldots \cup \boldsymbol{v}_l$
Attend($\mathbf{Q}, \mathbf{K}, \mathbf{V}$) $\leftarrow$ Softmax($\mathbf{QK}^T$ $/\sqrt{d}$)$\mathbf{V}$
$\mathbf{Re} \leftarrow$ Rearrange(Attention($\mathbf{Q}, \mathbf{K}, \mathbf{V}$) $|bhnd \rightarrow bn(hd)$)
$\mathbf{Out}_{Broad} \leftarrow$ AdaptivePool($\mathbf{Re}$)
$\mathbf{Out}_{DB}$ = $\mathbf{Out}_{Deep}$ + $\boldsymbol{\gamma} \times \mathbf{Out}_{Broad}$
for $i = 1, 2, \ldots, n$ do
    Random $\mathbf{W}_{z_i}, \boldsymbol{\beta}_{z_i}$;
    Caculate $\mathbf{Z}_i$ = $\phi_i(\mathbf{Out}_{DB}\mathbf{W}_{z_i} + \boldsymbol{\beta}_{z_i})$
end for
Stack all the mapping feature nodes as
    $\mathbf{Z}^n = [\mathbf{Z_1}, \mathbf{Z_2}, \ldots, \mathbf{Z_n}]$
for $j = 1, 2, \ldots, m$ do
    Random $\mathbf{W}_{h_j}, \boldsymbol{\beta}_{h_j}$;
    Caculate $\mathbf{H}_j$ = $\xi_j(\mathbf{Z}^n\mathbf{W}_{h_j} + \boldsymbol{\beta}_{h_j})$
end for
Stack all the enhancement nodes as
    $\mathbf{H}^m = [\mathbf{H}_1, \mathbf{H}_2, \ldots, \mathbf{H}_m]$
Calculate the weight connected the hidden layer and output layer
    $\mathbf{w_2}$ by: $\mathbf{w_2} = (\mathbf{A}^T\mathbf{A} + \lambda_1\boldsymbol{I})^{-1}\mathbf{A}^T\mathbf{Y_T}$
Get the prediction matrix for seizure detection
    $\mathbf{Y_P} = [\mathbf{Z}^n|\mathbf{H}^m]\mathbf{w_2}$
```

$$v_i = [[v_i^1], [v_i^2], \ldots, [v_i^h]], v_i^j \in R^{L \times C \times d_v}$$

where q_i^j, k_i^j, v_i^j are the corresponding value of j-th head in i-th layer of q_i, k_i, v_i, w^o is the weight matrix of the second linear projection. Because of the residual connection between the layers, the hidden layer's output $\hat{y}_i$ in i-th layer is formulated by

$$\hat{y}_i = x_i + \mathrm{MHSA}(x_i) \tag{3}$$

Multi Layer Perceptron: MLP has two fully connected layers and an activation function layer, the activation function used in this paper is GELU. The output of MLP can be denoted as

$$\mathrm{MLP}(\hat{y}_i) = \mathrm{GELU}(\hat{y}_i w_{1l} + b_{1l})w_{2l} + b_{2l} \tag{4}$$

where w_{1l}, b_{1l}, w_{2l}, and b_{2l} are the weights and bias of the corresponding linear layers. The output y_i in i-th layer is formulated as

$$y_i = \hat{y}_i + \mathrm{MLP}(\hat{y}_i) \tag{5}$$

The output y_i of i-th layer is the input x_{i+1} of $(i+1)$-th layer, so the deep feature Out_{Deep} is the output of last layer:

$$Out_{Deep} = y_l \tag{6}$$

Broad Attention: Queries, keys and values of different layers are concatenated respectively as below:

$$Q = [q_1, q_2, \ldots, q_l],\ Q \in R^{L \times h \times C \times (l \times d_q)}$$
$$K = [k_1, k_2, \ldots, k_l],\ K \in R^{L \times h \times C \times (l \times d_k)}$$
$$V = [v_1, v_2, \ldots, v_l],\ V \in R^{L \times h \times C \times (l \times d_v)}$$

Self-attention is performed on the concatenated query Q, key K and value V to get Attention(Q, K, V). In this paper, 1D adaptive average pooling is introduced to solve the problem of dimension inconsistency between Out_{Deep} and Attention(Q, K, V). The output features Out_{Broad} of broad attention can be denoted as:

$$\begin{aligned} \mathrm{MHSA}(x_i) &= \mathrm{AdaptivePool}(\mathrm{Attention}(Q, K, V)) \\ &= \mathrm{AdaptivePool}(\mathrm{softmax}(\frac{QK^T}{\sqrt{d}})V) \end{aligned} \tag{7}$$

where d is the hidden dimension of transformer layer.

Combining the deep feature Out_{Deep} and local broad feature Out_{Broad}, the final output feature Out_{DB} of BViT is computed as:

$$Out_{DB} = Out_{Deep} + \gamma \times Out_{Broad} \tag{8}$$

where γ can be used to adjust the weights of two types of features. Finally, the probability of categories is calculated by a softmax function. So far, we have obtained the spatio-temporal long-range correlation features required for seizure prediction. The feature data and its labels are denoted as $\left\{(Out_{DB}, Y_T)|Out_{DB} \in R^{L \times (C \times D)}, Y_T \in R^{L \times M}\right\}$ from M classes.

It is necessary to search for seizure prediction features in a vast space to learn new generalized spatio-temporal long-range correlation features that help predict seizures. Therefore, the above algorithm is extended to a vast space through BLS to learn generalized features, so as to improve the performance and representation ability of seizure prediction tasks. Firstly, Out_{DB} are randomly extended to a vast space via a linear random mapping, that is:

$$Z_i \triangleq \phi_i(Out_{DB}W_{z_i} + \beta_{z_i}), \quad i = 1, \ldots, n \tag{10}$$

where W_{z_i} and β_{z_i} are generated by a random mapping ϕ_i. Then the set of n groups of feature nodes can be defined as $Z^n \triangleq [Z_1, Z_2, \ldots, Z_n]$.

Secondly, the j-th group of enhancement nodes can be constructed by

$$H_j \triangleq \xi_j(Z^n W_{h_j} + \beta_{h_j}),\ \ j = 1, \ldots, m \tag{11}$$

similarly, both W_{h_j} and β_{h_j} are generated by the nonlinear random mapping ξ_j. The set of m groups of enhancement

TABLE III

OVERALL PERFORMANCE COMPARISON ON CHB-MIT DATASET

Patient	CNN [9]				DCNN+Bi-LSTM [13]				Vision Transformer [20]				AdderNet [8]				Multi-scale ProtoPNet [26]				B2-ViT			
	AUC	$S_n(\%)$	FPR/h	p-value	AUC	$S_n(\%)$	FPR/h	p-value	AUC	$S_n(\%)$	FPR/h	p-value	AUC	$S_n(\%)$	FPR/h	p-value	AUC	$S_n(\%)$	FPR/h	p-value	AUC	$S_n(\%)$	FPR/h	p-value
Pat 1	0.919	85.7	0.240	<0.001	0.944	100.0	0.116	0.005	0.997	85.7	0.000	<0.001	1.000	100.0	0.000	<0.001	0.965	100.0	0.000	<0.001	0.999	100.0	0.000	<0.001
Pat 2	0.335	33.3	0.000	<0.001	0.709	66.7	0.254	0.376	0.745	33.3	0.000	<0.001	0.815	66.7	0.000	<0.001	0.771	100.0	0.000	<0.001	0.856	100.0	0.043	<0.001
Pat 3	0.968	100.0	0.180	<0.001	0.738	66.7	0.277	0.042	0.975	66.7	0.091	<0.001	0.922	83.3	0.040	<0.001	0.890	100.0	0.000	<0.001	0.979	100.0	0.000	<0.001
Pat 5	0.871	80.0	0.190	0.010	0.846	80.0	0.157	0.013	0.885	60.0	0.000	<0.001	0.981	100.0	0.000	<0.001	0.728	100.0	0.101	<0.001	0.939	80.0	0.000	<0.001
Pat 9	0.742	50.0	0.120	0.067	0.717	75.0	0.424	0.234	0.753	50.0	0.043	0.003	0.596	25.0	0.043	0.082	0.636	50.0	0.000	<0.001	0.807	50.0	0.022	<0.001
Pat 10	0.556	33.3	0.000	0.025	0.866	66.7	0.478	0.023	0.692	66.7	0.115	<0.001	0.916	85.7	0.000	<0.001	0.765	100.0	0.294	<0.001	0.817	83.3	0.038	<0.001
Pat 13	0.968	80.0	0.140	<0.001	0.786	85.7	0.219	0.001	0.920	100.0	0.143	<0.001	0.996	100.0	0.143	<0.001	0.899	100.0	0.044	<0.001	0.969	100.0	0.143	<0.001
Pat 14	0.662	80.0	0.400	0.004	0.710	100.0	0.313	0.031	0.640	80.0	0.400	0.005	0.764	71.4	0.400	0.003	0.705	87.5	0.159	<0.001	0.731	100.0	0.400	<0.001
Pat 18	0.935	100.0	0.280	<0.001	0.955	50.0	0.208	0.036	0.825	100.0	0.083	<0.001	1.000	100.0	0.000	<0.001	0.811	75.0	0.000	<0.001	0.961	100.0	0.042	<0.001
Pat 19	0.999	100.0	0.000	<0.001	0.990	0.0	0.038	0.302	0.989	100.0	0.040	<0.001	0.993	100.0	0.040	<0.001	0.990	100.0	0.000	0.001	0.997	100.0	0.000	<0.001
Pat 20	0.984	100.0	0.250	<0.001	0.960	100.0	0.184	<0.001	0.982	100.0	0.100	<0.001	0.979	83.3	0.050	<0.001	0.996	100.0	0.000	<0.001	0.997	100.0	0.050	<0.001
Pat 21	0.903	100.0	0.230	<0.001	0.648	100.0	0.521	0.101	0.864	100.0	0.091	<0.001	0.969	100.0	0.125	<0.001	0.867	100.0	0.000	<0.001	0.947	100.0	0.000	<0.001
Pat 23	0.999	100.0	0.330	<0.001	0.839	85.7	0.095	<0.001	0.930	100.0	0.000	<0.001	0.992	100.0	0.078	<0.001	0.932	100.0	0.172	<0.001	0.998	100.0	0.000	<0.001
Average	0.834	80.18	0.182	9/13	0.824	75.12	0.253	2/13	0.861	80.2	0.085	11/13	0.917	85.8	0.071	11/13	0.843	93.27	0.059	12/13	**0.923**	**93.33**	**0.057**	**13/13**

TABLE IV

OVERALL PERFORMANCE COMPARISON ON KAGGLE DATASET

Participant	CNN [9]				DCNN+Bi-LSTM [13]				Vision Transformer [20]				AdderNet [8]				Multi-scale ProtoPNet [26]				B2-ViT			
	AUC	$S_n(\%)$	FPR/h	p-value	AUC	$S_n(\%)$	FPR/h	p-value	AUC	$S_n(\%)$	FPR/h	p-value	AUC	$S_n(\%)$	FPR/h	p-value	AUC	$S_n(\%)$	FPR/h	p-value	AUC	$S_n(\%)$	FPR/h	p-value
Dog 1	0.498	50.0	0.19	0.053	0.566	50.0	0.352	0.59	0.577	25.0	0.000	<0.001	0.603	75.0	0.161	0.010	0.568	50.0	0.164	<0.001	0.608	50.0	0.000	<0.001
Dog 2	0.941	100.0	0.04	<0.001	0.953	100.0	0.048	<0.001	0.838	57.1	0.012	<0.001	0.917	85.7	0.024	<0.001	0.870	100.0	0.048	<0.001	0.898	100.0	0.024	<0.001
Dog 3	0.824	58.3	0.14	<0.001	0.831	91.7	0.189	<0.001	0.800	41.7	0.013	<0.001	0.830	75.0	0.050	<0.001	0.811	100.0	0.180	<0.001	0.887	83.3	0.004	<0.001
Dog 4	0.775	78.6	0.48	<0.001	0.754	85.7	0.191	<0.001	0.693	50.0	0.052	<0.001	0.726	92.9	0.313	<0.001	0.700	92.9	0.244	<0.001	0.773	92.9	0.037	<0.001
Dog 5	0.922	80.0	0.08	<0.001	0.928	80.0	0.134	<0.001	0.885	60.0	0.013	<0.001	0.892	80.0	0.040	<0.001	0.872	100.0	0.094	<0.001	0.913	100.0	0.000	<0.001
Average	0.792	73.38	0.19	4/5	0.806	81.5	0.183	4/5	0.759	47.8	0.02	5/5	0.794	81.7	0.118	4/5	0.764	88.6	0.146	5/5	**0.816**	85.2	**0.013**	**5/5**

nodes can be defined as $H^m \triangleq [H_1, H_2, \ldots, H_m]$, ξ_j is the tansig function here, tansig is a hyperbolic tangent s-type nonlinear function, which is defined as:

$$\operatorname{tansig}(x) = \frac{2}{1+e^{-2x}} - 1 \tag{12}$$

Therefore, the output Y_P of the improved algorithm with BLS can be constructed by the following formula:

$$\begin{aligned} Y_P &= [Z^n | H^m] w_2 \\ &= A w_2 \end{aligned} \tag{13}$$

w_2 can be obtained by solving the ridge regression problem:

$$w_2 = (\lambda_2 I + A A^T)^{-1} A^T Y_T \tag{14}$$

where λ_2 is the regularization coefficient.

Our proposed model B2-ViT is a novel bilevel programming problem, the goal of the model is shown in Eq. (9), shown at the bottom of the page, where x_1 is the input data, Y_T is the true label, $w_{1,r}$ is the corresponding weight of the front r layer of our proposed model, W_z and β_z are the corresponding weight and offset of the feature nodes, W_h and β_h are the corresponding weight and offset of the enhancement nodes, ϕ and ξ are random mappings used to generate feature nodes and enhancement nodes, L is the length of x_1, $f(x_1; w_{1,r})$ is a BViT function of input x_1, which is parameterized by a weight vector $w_{1,r}$, l is the loss function of BViT, Out_{DB} can be denoted as $f(x_1, w_{1,r-1})$, $\lambda_1 \|w_{1,r}\|_2^2$ is the regularization term that penalizes the complexity of weights, softmax is a classification function.

D. Postprocessing

In this work, the k-of-n method is used to predict seizure as in [9], [11], an alarm is set only when at least k of the n predictions are positive, we set k to 4 and n to 5. In addition, to avoid the increase of False Prediction Rate (FPR) caused by multiple alarms in a short time, we set the refractory period to 30 min, that is, the reoccurring alarm within 30 min after the alarm occurs will be ignored.

IV. EXPERIMENTAL RESULTS

A. Experimental Settings and Evaluation Metrics

In this work, Area Under Curve (AUC), Sensitivity (S_n), FPR, p-value are choosen as the evaluation metrics of the proposed method. AUC is a performance metric to measure the quality of the classifier. The closer to 1, the better the effect. S_n is the ratio of correctly predicted seizures to all seizures. FPR is the number of mispredictions per hour. p-value is the probability of predicting at least m of M seizures, which can be obtained by the following

$$\begin{aligned} &\min_{w_2} \; \|[\phi(x_1, w_{1,r-1}; W_z, \beta_z), \xi(x_1, w_{1,r-1}; W_z, \beta_z, W_h, \beta_h)] w_2 - Y_T\|_2^2 + \lambda_2 \|w_2\|_2^2 \\ &\begin{cases} w_{1,r} = \arg\min_{w_{1,r}} \sum_{i=1}^{L} l(Y_{T,i}, f(x_{1,i}, w_{1,r})) + \lambda_1 \|w_{1,r}\|_2^2 \\ w_{1,r} = [w_{1,r-1}, \text{softmax}] \end{cases} \end{aligned} \tag{9}$$

TABLE V
PERFORMANCE COMPARISON OF VIT, B-VIT AND B2-VIT ON CHB-MIT DATASETS

Patient	ViT				BViT				B2-ViT			
	AUC	$S_n(\%)$	FPR/h	p-value	AUC	$S_n(\%)$	FPR/h	p-value	AUC	$S_n(\%)$	FPR/h	p-value
Pat 1	0.997	85.7	0.000	<0.001	0.998	100.0	0.059	<0.001	0.999	100.0	0.000	<0.001
Pat 2	0.745	33.3	0.000	<0.001	0.763	66.7	0.000	<0.001	0.856	100.0	0.043	<0.001
Pat 3	0.975	66.7	0.091	<0.001	0.977	83.3	0.182	<0.001	0.979	100.0	0.000	<0.001
Pat 5	0.885	60.0	0.000	<0.001	0.934	60.0	0.000	<0.001	0.939	80.0	0.000	<0.001
Pat 9	0.753	50.0	0.043	0.003	0.754	50.0	0.022	<0.001	0.807	50.0	0.022	<0.001
Pat 10	0.692	66.7	0.115	<0.001	0.534	66.7	0.115	<0.001	0.817	83.3	0.038	<0.001
Pat 13	0.920	100.0	0.143	<0.001	0.949	100.0	0.143	<0.001	0.969	100.0	0.143	<0.001
Pat 14	0.640	80.0	0.400	0.005	0.662	80.0	0.400	0.005	0.731	100.0	0.400	<0.001
Pat 18	0.825	100.0	0.083	<0.001	0.897	100.0	0.042	<0.001	0.961	100.0	0.042	<0.001
Pat 19	0.989	100.0	0.040	<0.001	0.991	100.0	0.000	<0.001	0.997	100.0	0.000	<0.001
Pat 20	0.982	100.0	0.100	<0.001	0.987	100.0	0.050	<0.001	0.997	100.0	0.050	<0.001
Pat 21	0.864	100.0	0.091	<0.001	0.875	100.0	0.046	<0.001	0.947	100.0	0.000	<0.001
Pat 23	0.930	100.0	0.000	<0.001	0.998	100.0	0.000	<0.001	0.998	100.0	0.000	<0.001
Average	0.861	80.2	0.085	11/13	0.871	85.1	0.081	12/13	**0.923**	**93.3**	**0.057**	**13/13**

TABLE VI
PERFORMANCE COMPARISON OF VIT, B-VIT AND B2-VIT ON KAGGLE DATASETS

Participant	ViT				BViT				B2-ViT			
	AUC	$S_n(\%)$	FPR/h	p-value	AUC	$S_n(\%)$	FPR/h	p-value	AUC	$S_n(\%)$	FPR/h	p-value
Dog 1	0.577	25.0	0.000	<0.001	0.584	25.0	0.000	<0.001	0.608	50.0	0.000	<0.001
Dog 2	0.838	57.1	0.012	<0.001	0.865	71.4	0.012	<0.001	0.898	100.0	0.024	<0.001
Dog 3	0.800	41.7	0.013	<0.001	0.836	58.3	0.013	<0.001	0.887	83.3	0.004	<0.001
Dog 4	0.693	50.0	0.052	<0.001	0.716	78.6	0.045	<0.001	0.773	92.9	0.037	<0.001
Dog 5	0.885	60.0	0.013	<0.001	0.886	80.0	0.000	<0.001	0.913	100.0	0.000	<0.001
Average	0.759	47.8	0.02	5/5	0.777	62.7	0.014	5/5	**0.816**	**85.2**	**0.013**	**5/5**

TABLE VII
COMPUTATIONAL COMPLEXITY EVALUATION METRICS OF VIT, B-VIT AND B2-VIT ON CHB-MIT AND KAGGLE DATASETS

Model	CHB-MIT		Kaggle	
	Params (M)	Training time (s)	Params (M)	Training time (s)
ViT	19.08	1798.64	19.07	3717.13
BViT	19.08	1860.67	19.07	3825.45
B2-ViT	19.08	1862.61	19.07	3826.62

formula [9], [11]:

$$p = \sum_{i \geqslant m} \binom{i}{M} P^{i}(1-P)^{M-i} \tag{15}$$

where $P \approx 1 - e^{-\text{FPR}\times\text{SOP}}$, SOP is the seizure occurrence period, is set to 30 min. If $p < 0.001$, it can be considered that our model is superior to random prediction at the 0.001 significance level.

To make the results more reliable, the leave-one-out cross-validation (LOOCV) method is used for each subject. If the subject has M seizures, $M - 1$ seizures will be used for training, and the rest seizure will be used for testing. Each seizure will be taken as the testing set in turn. In addition, to monitor whether the model is overfitting in real-time and adjust the parameters of the model, $M - 1$ seizures data for training are divided into training set and validation set, the proportion of validation set is set to 25% as in most seizure prediction studies. The number of transformer blocks l is 6, the number of self-attention heads h is 8, the dimension of one self-attention head is 64, the hidden layer size is 512.

Our experiments are based on PyTorch 1.11.0, which is implemented using Python 3.8 and Cuda 11.3.0. The loss is optimized by the Stochastic Gradient Descent (SGD) optimizer (learning rate = 0.001, momentum = 0.9, weight decay = $5e^{-5}$), the cosine scheduler is used to optimize the learning rate, the epoch of training is set to 100, the loss function is Cross Entropy Loss. The early stopping method with patience 10 is used to obtain better generalization performance and avoid over-fitting. Two GeForce RTX 3090 Ti are used that approximately need 64 GB GPU memory in total.

B. Overall Performance Comparison

The overall performance of B2-ViT Net is evaluated in the following methods. CNN [9] is a forward neural network with deep structure and convolution calculation, which is one of the representative algorithms of deep learning and has achieved good results in computer vision and natural language processing in recent years. It is one of the most popular in deep learning methods currently designed for EEG decoding [41]. DCNN+Bi-LSTM [13] used DCNN to extract spatial features, and Bi-LSTM was used as a classifier to improve classification accuracy, which is typically designed to predict EEG seizures. Vision Transformer [20] is the application of transformer in the field of computer vision, achieving performance beyond CNN in most visual tasks. AdderNet [8] proposed a simple and effective end-to-end adder network and supervised contrastive learning, used addition instead of multiplication significantly reduces computational costs. Multi-scale ProtoPNet [26] proposed a deep learning model for patient-specific seizure prediction, it attempted to measure the similarity between the inputs and prototypes (learned during training) as evidence to make final predictions.

From Table III, Table IV and Fig 3, the proposed B2-ViT scheme yields an average AUC of 0.923 while other methods only achieve an average AUC of 0.834, 0.824, 0.861, 0.917, 0.843 on the CHB-MIT datasets and an average AUC of 0.816 while other baseline methods only achieve an average

TABLE VIII
EXPERIMENTAL SETUP AND PERFORMANCE RESULTS OF EXISTING METHODS ON CHB-MIT AND KAGGLE DATASETS

Years	Methods	Datasets	No. of patients-seizures	Validation scheme	Intericatal distance-Preictal length (min)	No. of patients over chance	Average AUC-Sn(%) -FDR/h
2013	Zero-croassing intervals+ Bayesian Gaussian mixture [34]	CHB-MIT	3-18	no CV	40-40	3/3	NR-83.8-0.165
2017	EMD, PLV + SVM [35]	CHB-MIT	21-65	10-fold CV	30-5	NR	NR-82.4-NR
2018	STFT + CNN [9]	CHB-MIT Kaggle2014	13-64 7-48	LOOCV	240-30 10080-30	12/13 5/7	0.834-80.2-0.182 NR-75.0-0.210
2023	Multi-scale ProtoPNet [26]	CHB-MIT Kaggle2014	13-71 5-42	LOOCV	60-30 240-60	13/13 5/5	0.843-93.27-0.059 0.764-88.6-0.146
2018	Wavelet Transform + CNN [36]	CHB-MIT	15-18	10-fold CV	10-10	NR	0.866-87.8-0.147
2019	Spectral power, statistical moments, Hjorth + 3D CNN [11]	CHB-MIT	16-77	LOOCV	60-60 120-60 240-60	13/16 14/16 15/16	NR-86.8-0.292 NR-87.0-0.186 NR-85.7-0.096
2020	Common spatial pattern statistics, Butterworth band-pass filter + CNNs [29]	CHB-MIT	23-156	LOOCV	30-30	NR	0.90-92.0-0.120
2018	Temporal, frequency, channels correlation, graph theoretic features + LSTM [37]	CHB-MIT	24-185	10-fold CV	NR-15 NR-30 NR-60 NR-120	NR NR NR NR	NR-99.3-0.110 NR-99.4-0.060 NR-99.6-0.030 NR-99.8-0.020
2021	STFT + STCNN [27]	Kaggle2014	7-64	LOOCV	NR	NR	0.746-82.0-0.380
2021	STFT + Residual Network-Self Attention [38]	CHB-MIT	13-64	LOOCV	240-30	NR	0.913-89.3-NR
2022	Raw data + AdderNet [8]	CHB-MIT Kaggle2014	13-68 5-42	LOOCV	240-60 10080-60	NR	0.917-85.8-0.071 0.794-81.7-0.118
2023	Temporal feature + PCA-SVM [39]	CHB-MIT	23-173	NR	NR	NR	0.900-91.8-NR
2023	Tangent space features + GSFDA [28]	CHB-MIT Kaggle2014	15-89 5-42	LOOCV	240-60 10080-60	NR	0.683-70.5-0.381 0.622-61.9-0.422
2023	SVM [40] +STFT	CHB-MIT Kaggle2014	13-64 5-42	LOOCV	240-30 10080-60	9/13 3/5	0.663-60.4-0.347 0.654-69.4-0.208
This work	**STFT + B2-ViT Net**	**CHB-MIT Kaggle2014**	**13-64 5-42**	**LOOCV**	**240-30 10080-60**	**13/13 5/5**	**0.923-93.3-0.057 0.816-85.2-0.013**

AUC of 0.792, 0.806, 0.759, 0.794, 0.764 on the Kaggle datasets, which shows that our proposed method has good classification ability. In particular, patients 1, 19, 20 and 23 of CHB-MIT reach an AUC greater than 0.99, which proves the effectiveness of our method in distinguishing preictal EEG signal from interictal EEG signal. In addition, our seizure prediction method is superior to other compared methods by successfully warning 60 seizures out of 64 on the CHB-MIT dataset, 37 seizures out of 42 on the Kaggle dataset. Meanwhile, our method achieves a remarkably low FPR.

As a result, the bi-level programming model B2-ViT Net obtains the promising AUC, S_n and FPR, which indicates the effectiveness of our proposed method in automatic seizure prediction. In addition, for all subjects in CHB-MIT and Kaggle datasets, the p-value is less than 0.001, this shows that our seizure predictor is significantly better than the random predictor under 99.9% confidence interval (0.001 significance level), which is statistically significant, providing significantly excellent performance in automatic seizure prediction of our proposed B2-ViT framework.

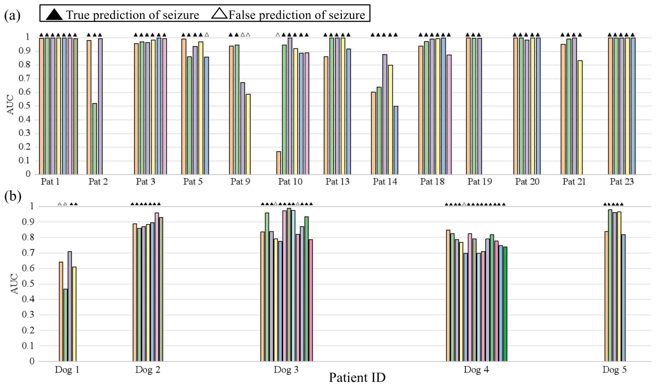

Fig. 3. The AUC for each seizure prediction on the (a) CHB-MIT and (b) Kaggle datasets. Each bar represents one seizure. Correct and incorrect predictions of seizure are given with ▲ and △, respectively.

V. DISCUSSIONS

A. Ablation Studies

To verify the effectiveness of our proposed B2-ViT model, we conducted further ablation experiments, and the results are shown in Table V, VI and Fig. 4. It can be seen that on the CHB-MIT dataset, all the evaluation metrics of BViT model are higher than ViT, AUC is increased by 1%, S_n is increased by 4.9%, FPR is decreased by 0.004, and the p-value under the significance level of 0.001 is increased from 11/13 to 12/13. The evaluation metrics of B2-ViT are significantly higher than those of BViT, with AUC increased by 4.2%, S_n increased by 8.2%, FPR decreased by 0.024, and the p-value under the significant level of 0.001 increased from 12/13 to 13/13. Similar results can be obtained on the Kaggle dataset. As can be observed, the results prove the effectiveness of B2-ViT model, B2-ViT model consistently outperforms ViT and BViT models. Moreover, Table VII shows the params and training time of different algorithms, which indicates that our method achieves high performance improvement in a small increment of training time without increasing trainable parameters.

B. Effects of Different Window Lengths of EEG Signals

An appropriate window length is expected to achieve better performance. We evaluate the effect of different window

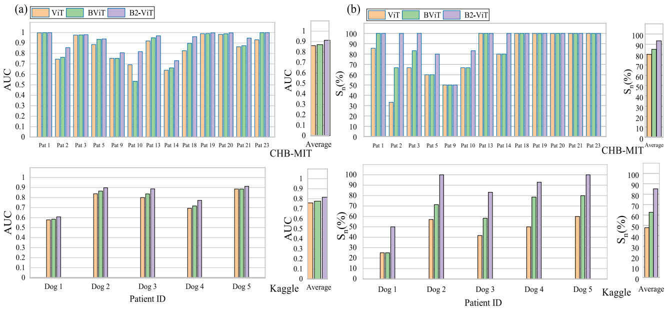

Fig. 4. (a) AUC comparison (left column) and (b) S_n comparison (right column) of ViT, B-ViT and B2-ViT on CHB-MIT and Kaggle Datasets.

TABLE IX
PERFORMANCE COMPARISON OF DIFFERENT WINDOW LENGTHS ON THE CHB-MIT DATASET

Model	Window	AUC	$S_n(\%)$	FDR/h	p-value
ViT	10s	0.843	77.6	0.090	1/13
	20s	0.859	78.6	0.090	1/13
	30s	0.861	80.2	0.085	1/13
	40s	0.860	79.5	0.095	2/13

lengths on the experimental results using the baseline method ViT, and find that the window length of 30s is more appropriate. The results are shown in table IX. Within 30s, with the increase of window length, ViT contains more and more distinctive feature information, and its performance is getting better and better. When the window length exceeds 30s, all evaluation metrics decline, and the classification performance reaches the bottleneck. This shows that the window length of 30s contains enough feature information for classification, so the window length of 30s is chosen for seizure prediction.

C. Effects of Parameter Settings in BLS

Relevant parameter settings in BLS may affect the experimental results of our proposed model, the number of feature nodes and enhancement nodes can be adjusted according to different experimental scenarios. To verify the robustness of our proposed model, the influence of important experimental parameters of BLS on AUC is analyzed. Fig. 5 shows the corresponding AUC under different mapping feature nodes and enhancement nodes. The range of the feature nodes' groups is set to 10-15, and the number of enhancement nodes is set to 1, 100, 500, 1000, 5000. It can be seen that the best experimental results can be obtained when the mapping feature nodes and enhancement feature nodes are 165 and 100 respectively on patient 1 of the CHB-MIT dataset. The AUC of B2-ViT is relatively stable, and good experimental results are obtained. Therefore, the automatic seizure prediction performance of B2-ViT does not fluctuate obviously due to the change of parameters of BLS, which shows that our proposed method has good robustness in BLS module.

D. Performance Comparison of the Existing State-of-the Art Methods

Table VIII shows the experimental settings and performance results of the existing state-of-the-art methods on CHB-MIT and Kaggle datasets, where NR is not reported values. It is necessary to point out that it is difficult to compare our method directly with the existing methods due to the different experimental settings, such as the Interictal distance-Preictal length and validation scheme. Compared with our LOOCV strategy, the no-cv and k-fold cv in [34], [35], [36], and [37] are much less challenging and stable, and the intrapatient variation of seizures is ignored. In addition, although statistical significance research has always been emphasized, only [9], [11], [26], [34], and [42] have statistical evaluation. Moreover, deep learning method is usually a black box, and the interpretability of the model is an important research direction at present, but few studies give interpretability in specific scenarios.

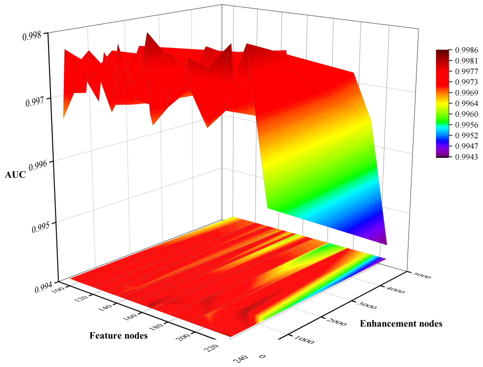

Fig. 5. The impact of feature nodes and enhancement nodes on the B2-ViT Net (chb01).

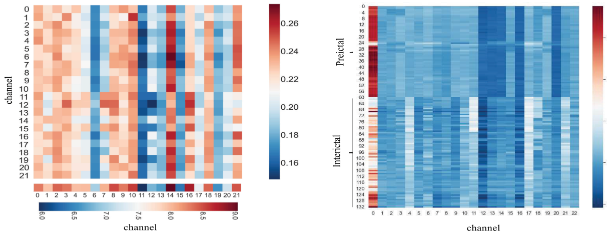

Fig. 6. Left: attention weights among channels. Channel 15 show strong abnormal connections with other channels, so channel 15 may be located in the seizure zone, where 0 represents channel 1. The heatmap below represents the sum of attention for each channel. Right: channels' attention weights of preictal and interictal. In preictal, channel 12, 13, 14, 16, 20 are assigned lower attention weights and other channels are assigned higher attention weights, where 0 represents the classfication token, 1 represents channel 1 (This figure originates from chb01).

As a result, compared with other methods, our bi-level programming model B2-ViT Net yields a competitive AUC, S_n, FPR and p-value. The AUC, FPR and p-value has reached SOTA, only the S_n lower than [37]. Although [37] achieved very high sensitivity on the CHB-MIT dataset, they adopted a complex time-consuming feature extraction method and 10-fold CV instead of LOOCV. Because each seizure is independent in LOOCV, it is more realistic and useful in clinical application. What's more, our model uses the attention mechanism to explain the global spatial interactions among channels and long-range temporal dependencies required for seizure prediction, so that the model has a certain degree of interpretability.

E. Limitations and Future Directions

Although our proposed seizure prediction algorithm achieves strong prediction performance, some limitations still remain in the current work. On the one hand, due to the lack of detailed information on the patient's epileptogenic zone and corresponding biomarkers, the results were not validated through neuroscience experiments. For example, channels located in the epileptogenic zone may show strong abnormal connections with other channels, channels located in the epileptogenic zone are assigned attention weights higher than other channels. Besides, the neural links between brain regions assigned high attention weights were not captured. Fig. 6 shows some of our conjectures.

On the other hand, our method is based on patient-dependent, meaning that both the training and test sets come from the same patient. It cannot be directly used for patient-independent seizure early warning tasks, i.e., the model trained by one patient cannot be applied to another patient. This is mainly because our method lacks the ability to handle the different distribution between the training and test sets. Therefore, transfer learning strategies [43], [44] will be considered to improve the performance of patients-independent seizure prediction tasks in our future work. In addition, we will try to cooperate with medical institutions, further explore the biomarkers of the epileptogenic zone, the neural links between the brain regions assigned high attention weights, and apply our proposed method to the realistic seizure prediction tasks in the future.

VI. Conclusion

Based on neuroscience mechanisms, we consider the global channel interactions in spatial, long-range dependencies in temporal together, and explore the generalized spatio-temporal long-range correlation features required for seizure prediction in a vast space. A novel bilevel programming model B2-ViT Net is proposed for extracting generalized spatio-temporal long-range correlation features for automatic seizure prediction. The proposed model has strong generalized feature search capability, which can comprehensively learn generalized spatio-temporal long-range correlation features that are conducive to automatic seizure prediction in a vast space, improving feature representation ability. In addition, the attention mechanism of our proposed model can calculate the interaction weights among channels, and evaluate the importance of each channel at any time. We evaluated the performance of B2-ViT model on the CHB-MIT and Kaggle datasets, the model yields promising results in terms of AUC, Sn, FPR and p-value, where the AUC, FPR and p-value have reached SOTA. Experimental results illustrate that our proposed method can predict seizures efficiently, help patients prevent or control the impending seizure, and improve their quality of life.

References

[1] P. Kwan, S. C. Schachter, and M. J. Brodie, "Drug-resistant epilepsy," *New England J. Med.*, vol. 365, no. 10, pp. 919–926, 2011.

[2] L. Ridsdale, J. Charlton, M. Ashworth, M. P. Richardson, and M. C. Gulliford, "Epilepsy mortality and risk factors for death in epilepsy: A population-based study," *Brit. J. Gen. Pract.*, vol. 61, no. 586, pp. e271–e278, May 2011.

[3] L. Kuhlmann, K. Lehnertz, M. P. Richardson, B. Schelter, and H. P. Zaveri, "Seizure prediction—Ready for a new era," *Nature Rev. Neurol.*, vol. 14, no. 10, pp. 618–630, Oct. 2018.

[4] S. L. Moshé, E. Perucca, P. Ryvlin, and T. Tomson, "Epilepsy: New advances," *Lancet*, vol. 385, no. 9971, pp. 884–898, Mar. 2015.

[5] C. Baumgartner and J. P. Koren, "Seizure detection using scalp-EEG," *Epilepsia*, vol. 59, no. S1, pp. 14–22, Jun. 2018.

[6] Y. Sun et al., "Continuous seizure detection based on transformer and long-term iEEG," *IEEE J. Biomed. Health Informat.*, vol. 26, no. 11, pp. 5418–5427, Nov. 2022.

[7] E. van Dellen et al., "Long-term effects of temporal lobe epilepsy on local neural networks: A graph theoretical analysis of corticography recordings," *PLoS ONE*, vol. 4, no. 11, Nov. 2009, Art. no. e8081.

[8] Y. Zhao, C. Li, X. Liu, R. Qian, R. Song, and X. Chen, "Patient-specific seizure prediction via adder network and supervised contrastive learning," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 30, pp. 1536–1547, 2022.

[9] N. D. Truong et al., "Convolutional neural networks for seizure prediction using intracranial and scalp electroencephalogram," *Neural Netw.*, vol. 105, pp. 104–111, Sep. 2018.

[10] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[11] A. R. Ozcan and S. Erturk, "Seizure prediction in scalp EEG using 3D convolutional neural networks with an image-based approach," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 27, no. 11, pp. 2284–2293, Nov. 2019.

[12] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 221–231, Jan. 2013.

[13] H. Daoud and M. A. Bayoumi, "Efficient epileptic seizure prediction based on deep learning," *IEEE Trans. Biomed. Circuits Syst.*, vol. 13, no. 5, pp. 804–813, Oct. 2019.

[14] D. J. Englot et al., "Global and regional functional connectivity maps of neural oscillations in focal epilepsy," *Brain*, vol. 138, no. 8, pp. 2249–2262, Aug. 2015.

[15] Y. Chen et al., "Adversarial learning based node-edge graph attention networks for autism spectrum disorder identification," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Mar. 14, 2022, doi: 10.1109/TNNLS.2022.3154755.

[16] P. Li et al., "EEG based emotion recognition by combining functional connectivity network and local activations," *IEEE Trans. Biomed. Eng.*, vol. 66, no. 10, pp. 2869–2881, Oct. 2019.

[17] T. Zhang, X. Wang, X. Xu, and C. L. P. Chen, "GCB-Net: Graph convolutional broad network and its application in emotion recognition," *IEEE Trans. Affect. Comput.*, vol. 13, no. 1, pp. 379–388, Jan. 2022.

[18] M. J. Cook et al., "The dynamics of the epileptic brain reveal long-memory processes," *Frontiers Neurol.*, vol. 5, p. 217, Oct. 2014.

[19] E. J. Ngamga, S. Bialonski, N. Marwan, J. Kurths, C. Geier, and K. Lehnertz, "Evaluation of selected recurrence measures in discriminating pre-ictal and inter-ictal periods from epileptic EEG data," *Phys. Lett. A*, vol. 380, no. 16, pp. 1419–1425, Apr. 2016.

[20] A. Dosovitskiy et al., "An image is worth 16 × 16 words: Transformers for image recognition at scale," in *Proc. 9th Int. Conf. Learn. Represent.*, May 2021, pp. 1–22.

[21] N. Li, Y. Chen, W. Li, Z. Ding, D. Zhao, and S. Nie, "BViT: Broad attention-based vision transformer," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, May 1, 2023, doi: 10.1109/TNNLS.2023.3264730.

[22] T. Zhang, X. Gong, and C. L. P. Chen, "BMT-Net: Broad multitask transformer network for sentiment analysis," *IEEE Trans. Cybern.*, vol. 52, no. 7, pp. 6232–6243, Jul. 2022.

[23] C. L. P. Chen and Z. Liu, "Broad learning system: An effective and efficient incremental learning system without the need for deep architecture," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 1, pp. 10–24, Jan. 2018.

[24] A. H. Shoeb, "Application of machine learning to epileptic seizure onset detection and treatment," Ph.D. dissertation, Dept. Health Sci. Technol., Massachusetts Inst. Technol., Cambridge, MA, USA, 2009.

[25] B. H. Brinkmann et al., "Crowdsourcing reproducible seizure forecasting in human and canine epilepsy," *Brain*, vol. 139, no. 6, pp. 1713–1722, Jun. 2016.

[26] Y. Gao, A. Liu, L. Wang, R. Qian, and X. Chen, "A self-interpretable deep learning model for seizure prediction using a multi-scale prototypical part network," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 31, pp. 1847–1856, 2023.

[27] R. Chen and K. K. Parhi, "Seizure prediction using convolutional neural networks and sequence transformer networks," in *Proc. 43rd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Nov. 2021, pp. 6483–6486.

[28] Y. Zhao, S. Feng, C. Li, R. Song, D. Liang, and X. Chen, "Source-free domain adaptation for privacy-preserving seizure prediction," *IEEE Trans. Ind. Informat.*, early access, Aug. 1, 2023, doi: 10.1109/TII.2023.3297323.

[29] Y. Zhang, Y. Guo, P. Yang, W. Chen, and B. Lo, "Epilepsy seizure prediction on EEG using common spatial pattern and convolutional neural network," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 2, pp. 465–474, Feb. 2020.

[30] T. Maiwald, M. Winterhalder, R. Aschenbrenner-Scheibe, H. U. Voss, A. Schulze-Bonhage, and J. Timmer, "Comparison of three nonlinear seizure prediction methods by means of the seizure prediction characteristic," *Phys. D, Nonlinear Phenomena*, vol. 194, nos. 3–4, pp. 357–368, Jul. 2004.

[31] L. Durak and O. Arikan, "Short-time Fourier transform: Two fundamental properties and an optimal implementation," *IEEE Trans. Signal Process.*, vol. 51, no. 5, pp. 1231–1242, May 2003.

[32] M. Farge, "Wavelet transforms and their applications to turbulence," *Annu. Rev. Fluid Mech.*, vol. 24, no. 1, pp. 395–458, Jan. 1992.

[33] K. Samiee, P. Kovács, and M. Gabbouj, "Epileptic seizure classification of EEG time-series using rational discrete short-time Fourier transform," *IEEE Trans. Biomed. Eng.*, vol. 62, no. 2, pp. 541–552, Feb. 2015.

[34] A. S. Zandi, R. Tafreshi, M. Javidan, and G. A. Dumont, "Predicting epileptic seizures in scalp EEG based on a variational Bayesian Gaussian mixture model of zero-crossing intervals," *IEEE Trans. Biomed. Eng.*, vol. 60, no. 5, pp. 1401–1413, May 2013.

[35] D. Cho, B. Min, J. Kim, and B. Lee, "EEG-based prediction of epileptic seizures using phase synchronization elicited from noise-assisted multivariate empirical mode decomposition," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 25, no. 8, pp. 1309–1318, Aug. 2017.

[36] H. Khan, L. Marcuse, M. Fields, K. Swann, and B. Yener, "Focal onset seizure prediction using convolutional networks," *IEEE Trans. Biomed. Eng.*, vol. 65, no. 9, pp. 2109–2118, Sep. 2018.

[37] K. M. Tsiouris, V. C. Pezoulas, M. Zervakis, S. Konitsiotis, D. D. Koutsouris, and D. I. Fotiadis, "A long short-term memory deep learning network for the prediction of epileptic seizures using EEG signals," *Comput. Biol. Med.*, vol. 99, pp. 24–37, Aug. 2018.

[38] X. Yang, J. Zhao, Q. Sun, J. Lu, and X. Ma, "An effective dual self-attention residual network for seizure prediction," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 29, pp. 1604–1613, 2021.

[39] L. Jiang, J. He, H. Pan, D. Wu, T. Jiang, and J. Liu, "Seizure detection algorithm based on improved functional brain network structure feature extraction," *Biomed. Signal Process. Control*, vol. 79, Jan. 2023, Art. no. 104053.

[40] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf, "Support vector machines," *IEEE Intell. Syst.*, vol. 13, no. 4, pp. 18–28, Jul./Aug. 1998.

[41] R. T. Schirrmeister et al., "Deep learning with convolutional neural networks for EEG decoding and visualization," *Hum. Brain Mapping*, vol. 38, no. 11, pp. 5391–5420, Nov. 2017.

[42] C. Li, X. Huang, R. Song, R. Qian, X. Liu, and X. Chen, "EEG-based seizure prediction via transformer guided CNN," *Measurement*, vol. 203, Nov. 2022, Art. no. 111948.

[43] T. Zhang, G. Su, C. Qing, X. Xu, B. Cai, and X. Xing, "Hierarchical lifelong learning by sharing representations and integrating hypothesis," *IEEE Trans. Syst. Man, Cybern. Syst.*, vol. 51, no. 2, pp. 1004–1014, Feb. 2021.

[44] X. Li et al., "EEG-based mild depression recognition using convolutional neural network," *Med. Biol. Eng. Comput.*, vol. 57, no. 6, pp. 1341–1352, Jun. 2019.