Exploring the Applicability of Transfer Learning and Feature Engineering in Epilepsy Prediction Using Hybrid Transformer Model

Shuaicong Hu, Jian Liu, Rui Yang, Ya'nan Wang, Aiguo Wang, Kuanzheng Li, Wenxin Liu, and Cuiwei Yang, *Member, IEEE*

***Abstract*— *Objective:* Epilepsy prediction algorithms offer patients with drug-resistant epilepsy a way to reduce unintended harm from sudden seizures. The purpose of this study is to investigate the applicability of transfer learning (TL) technique and model inputs for different deep learning (DL) model structures, which may provide a reference for researchers to design algorithms. Moreover, we also attempt to provide a novel and precise Transformer-based algorithm. *Methods:* Two classical feature engineering methods and the proposed method which consists of various EEG rhythms are explored, then a hybrid Transformer model is designed to evaluate the advantages over pure convolutional neural networks (CNN)-based models. Finally, the performances of two model structures are analyzed utilizing patient-independent approach and two TL strategies. *Results:* We tested our method on the CHB-MIT scalp EEG database, the results showed that our feature engineering method gains a significant improvement in model performance and is more suitable for Transformer-based model. In addition, the performance improvement of Transformer-based model utilizing fine-tuning strategies is more robust than that of pure CNN-based model, and our model achieved an optimal sensitivity of 91.7% with false positive rate (FPR) of 0.00/h. *Conclusion:* Our epilepsy prediction method achieves excellent performance and demonstrates its advantage over pure CNN-based structure in TL. Moreover, we find that the information contained in the gamma (γ) rhythm is helpful for epilepsy prediction. *Significance:* We propose a precise hybrid Transformer model for epilepsy prediction. The applicability of TL and model inputs is also explored for customizing personalized models in clinical application scenarios.**

***Index Terms*— Epilepsy prediction, feature engineering, scalp electroencephalogram (sEEG), hybrid transformer, transfer learning (TL).**

Manuscript received 24 October 2022; revised 14 January 2023; accepted 7 February 2023. Date of publication 16 February 2023; date of current version 21 February 2023. This work was supported in part by the Shanghai Municipal Science and Technology Major Project under Grant 2017SHZDZX01 and in part the Shanghai Science and Technology Support Project under Grant 22S31906200. *(Corresponding author: Cuiwei Yang.)*

Shuaicong Hu, Jian Liu, Rui Yang, and Ya'nan Wang are with the Center for Biomedical Engineering, School of Information Science and Technology, Fudan University, Shanghai 200433, China (e-mail: 22110720104@m.fudan.edu.cn; 22110720117@m.fudan.edu.cn; ryang20@fudan.edu.cn; wangyn20@fudan.edu.cn).

Aiguo Wang, Kuanzheng Li, and Wenxin Liu are with the Xinghua City People's Hospital, Jiangsu 225700, China (e-mail: xhwag@163.com; stronger1831@126.com; 450930173@qq.com).

Cuiwei Yang is with the Center for Biomedical Engineering, School of Information Science and Technology, Fudan University, Shanghai 200433, China, and also with the Key Laboratory of Medical Imaging Computing and Computer Assisted Intervention of Shanghai, Shanghai 200093, China (e-mail: yangcw@fudan.edu.cn).

Digital Object Identifier 10.1109/TNSRE.2023.3244045

I. Introduction

EPILEPSY is a chronic brain disease caused by the sudden abnormal discharge of neurons in the brain resulting in temporary impairment of brain function [1], [2]. The disease affects the normal life of approximately 1% of the world's population, where about 20-30% of patients are drug-resistant, known as intractable patients [3], [4]. For these people, it is a feasible scheme to alert them before a coming seizure, which will take care of the self-esteem of patients and avoid the serious consequences caused by a sudden seizure when they go out for activities [5], [6]. At present, Electroencephalography (EEG)-based epilepsy-related tasks are mainly divided into seizure detection and seizure prediction. For intractable patients, the clinical significance of seizure prediction is more significant than seizure detection, which can reduce the emotional stress of patients caused by seizures and allow doctors enough time to intervene clinically.

During the period preceding the onset of one seizure, the EEG signals will undergo a phase transition [7], and the EEG signals in this region will differ from the inter-ictal state. Finally, the onset of one seizure is predicted by detecting the corresponding features in pre-ictal state, which divides the EEG signals into inter-ictal, pre-ictal and ictal states. Unlike the large difference between ictal and non-ictal states, the difference between the inter-ictal and pre-ictal states is not significant and the differences are also different among various patients. This poses serious challenges to manually distinguish a large number of EEG signals.

In the early stages, Maiwald et al. clearly defined the term of epilepsy prediction in 2003 [8], and many studies have emerged since then. With the development of machine learning technology, utilizing EEG signals to automatic or semi-automatic predict epilepsy has become a research hotspot. Traditional feature extraction techniques such as

dynamical similarity index [8], mean phase coherence [9], phase-locking value [10], [11], zero-crossings [12] are widely used, and these features are continuously fed into classifiers such as Gaussian mixture models, adaptive boosting (AdaBoost), support vector machine (SVM) for epilepsy prediction. However, these methods cannot achieve high sensitivity and low false prediction rate (FPR) simultaneously.

Recent studies have shown the implementation of deep learning (DL) techniques [13] for EEG-based epilepsy prediction outperforming traditional machine learning (ML) methods. Especially, convolutional neural networks (CNNs) have gained maximum attention in epilepsy prediction [14], [15]. However, many DL models including CNNs ignore the key factor of attention, which will result in each feature of the input being an equal competitor, and the neural network must additionally learn the weights corresponding to the features to achieve the purpose of distinguishing the importance of features, which will result in a complex and heavy model [16], [17].

In addition to the limitations of algorithms, the disorder of EEG signals is another challenge in current research. Since the high complexity of signals and the irregular changes over time, it is very necessary to analyze the frequency domain of EEGs, which can further mine the hidden patterns of different frequency bands in EEGs. Many researchers have combined the time domain and frequency domain for EEG analyzing [14], [15], [18], [19], but the adaptability between model inputs and model structures is lack of exploration.

Moreover, in most of the existing epilepsy prediction methods without introduction of transfer learning (TL), the training strategies are primarily divided into subject-dependent approach and subject-independent approach. Due to subject-dependent approach requires a large portion of a subject's EEG recording to train the model before reaching its optimal performance, many researchers utilize a general model trained from multiple subjects through subject-independent approach, then the model is utilized for an unseen subject. Although these studies do not require individual training of the model for each subject, this will pose a more intractable challenge to the generalization of the model due to the high inter-subject variability. To make the epilepsy prediction model more reliable, training patient-specific models with a small amount of labeled data has become a very valuable research direction [20], [21]. Nevertheless, there is still a lack of investigation into the applicability between TL and model structures.

For this purpose, we employ TL technique to transfer the knowledge of the general model to a patient-specific model by fine-tune it with a small part of recording from an individual subject. In this paper, we mainly focus on the relationship between model inputs, model structures, and TL. The main contributions of this work are as follows:

1) We propose a novel hybrid Transformer model that can analyze EEG features from multiscale resolution and applies channel attention. Experimental results show that the model has a better ability to model feature sequence patterns than pure CNN-based models.

2) A feature engineering method specially designed to the hybrid Transformer model is proposed. Through the self-attention mechanism, the model can learn the correlation patterns between different rhythms by extracting different EEG rhythm signals. The comparison of two classical feature engineering methods shows that the proposed method is more suitable for the Transformer-based model.

3) A TL approach is introduced to optimize the general model based on subject-independent approach. We conduct a comprehensive evaluation of fine-tuning with only negative samples and fine-tuning with negative-positive mixed samples, respectively. It helps to reveal the impact of the availability of annotated data and fine-tuning strategies on model performance in clinic.

The rest of the paper is organized as follows: Section II describes the material and methodology. Section III describes the experimental details. Section IV presents the experimental details and results. Section V presents the discussions. Section VI concludes this work.

II. Material and Methodology

A. Database

The CHB-MIT scalp EEG database collected by Children's Hospital Boston [22] is utilized for algorithm evaluation. It includes 24 cases (*Chb01-Chb24*) from 23 pediatric patients with intractable epilepsy, where case *Chb01* and case *Chb21* are from the same female patient with an interval of 18 months and case *Chb24* had no relevant patient information. The subjects of these cases are 17 females and 5 males aged from 1.5 to 19, 3 to 22, respectively. The electrode position system is under the international 10-20 system standard [23]. The scalp EEG signals are sampled at 256 Hz with 16-bit resolution. Each case contains 9-24 proprietary EDF format files that store EEG signals. Every EDF file contains 1/2/4-hour continuous EEG recordings, and the EDF files with seizures have detailed annotations of the start and end time for each seizure. Details are shown in Table I.

To eliminate the difficulty of data analysis caused by the difference of acquisition electrodes, we select 18 channels shared by most EDF files, including 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 and CZ-PZ. FP, F, T, P, C, and O denote frontopolar, frontal, temporal, parietal, central, and occipital, respectively. Odd number, even number and Z denote left side, right side and midline of the brain, respectively.

B. EEG Preprocessing

The Raw EEG signal is standardized to alleviate the differences between subjects and help the deep learning model to converge quickly. Due to the possible existence of false spikes in the long-segment EEG signals, the maximum-minimum standardization may be unpredictably affected. For a more robust standardization method, this study adopts Z-score normalization $X'(t) = (X(t) - \mu)/\sigma$, where μ, σ are the mean value and standard deviation of original signal $X(t)$,

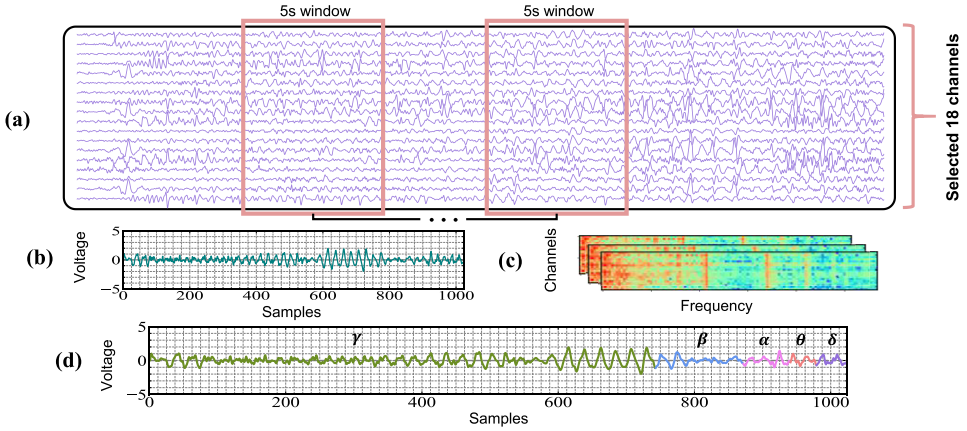

Fig. 1. Schematic diagram of three types of feature engineering. (a) Multi-lead EEG signals. (b) Raw EEG signals (Filtered). (c) EEG spectrogram. (d) Proposed mixed rhythm signals. An example of the selected 18-channel EEG recordings is visualized, and the 5 s sliding window is utilized to intercept EEG segments. (b)-(d) present the input forms based on different feature engineering methods corresponding to inter-ictal EEG signals.

TABLE I
DETAIL OF THE CHB-MIT EEG DATASET

Subjects Identifier	*Age*	*Gender*	*Record time (h)*	*Duration of seizures (s)*	*Number of Seizures (no merge)*	*Number of Seizures (merge)*
Chb01	11	Female	40.6	499	7	7
Chb02	11	Male	25.3	175	3	3
Chb03	14	Female	28.0	409	7	7
Chb04	22	Male	155.9	382	4	4
Chb05	7	Female	39.0	563	5	5
Chb06	1.5	Female	66.7	147	10	10
Chb07	14.5	Female	68.1	328	3	3
Chb08	3.5	Male	20.0	924	5	5
Chb09	10	Female	67.8	280	4	4
Chb10	3	Male	50.0	454	7	7
Chb11	12	Female	34.8	809	3	2
Chb12	2	Female	23.7	1515	40	21
Chb13	3	Female	33.0	547	12	10
Chb14	9	Female	26.0	117	8	8
Chb15	16	Male	40.0	2012	20	17
Chb16	7	Female	19.0	94	10	9
Chb17	12	Female	21.0	296	3	3
Chb18	18	Female	36.0	323	6	6
Chb19	19	Female	30.0	239	3	3
Chb20	6	Female	29.0	302	8	8
Chb21	13	Female	33.0	303	4	4
Chb22	9	Female	31.0	207	3	3
Chb23	6	Female	28.0	431	7	7

#When the interval between the end of one seizure and the start of the subsequent seizure is less than 30 min, the two seizures are combined into one seizure.

respectively. $X'(t)$ is the normalized signal. For ease of analysis, the long-term EEG signal is divided into 5 s segments. In this paper, we discuss three data input forms for model training, which are filtered raw EEG signals, short-time Fourier-transform (STFT) spectrograms, and proposed mixed EEG rhythm signals. More details can be found in Fig. 1.

1) Raw EEG signal: Referring to the previous study [24], we utilize a fifth-order Butterworth bandpass filter with a frequency range of 5-50 Hz to filter the raw EEG signals. Using filtered raw signals as model input facilitates an end-to-end automatic approach. However, accurate analysis of EEG requires a joint analysis of both time and frequency domains. Compared with the following two methods with frequency domain information, the model input of filtered EEG sequence is explored.

2) STFT Spectrogram: One of the most efficient techniques to convert the EEG sequence into the frequency domain is the Fourier Transform, particularly the fast Fourier transform (FFT). In this paper, the short-time Fourier transform with sliding window is utilized as a baseline feature engineering method:

$$STFT\left(s\left(t\right)\right) = \int_{-\infty}^{+\infty} s\left(t\right)\omega\left(t-\tau\right)e^{-i\omega t}dt \tag{1}$$

where $s(t)$ is the signal to be transformed, and $\omega(\tau)$ is the Gaussian window function. To eliminate the influence of puissance line noise and the DC element, the frequency bands range of 57-63 Hz, 117-123 Hz, and 0 Hz are removed [19]. Then we obtain a spectrogram matrix $\in \mathbb{R}^{18\times114\times9}$ for each 5s EEG segment, where 18 denotes the number of EEG channels, 114 denotes the frequency domain information, and 9 denotes the number of sliding windows. To convert the input into a form that the Transformer model can train, we reshape the spectrogram matrix to a new matrix $\in \mathbb{R}^{18\times1026}$ and utilize the SciPy signal processing module to resample the matrix $\in \mathbb{R}^{18\times1024}$.

3) Mixed EEG Rhythm Signal: The vitality of EEG signals is evaluated by normalized EEG frequency bands or rhythms, which are mainly divided into five rhythms: delta (δ) ($\leqslant$3 Hz), theta (θ) (4–7 Hz), alpha (α) (8–13 Hz), beta (β) (14–30 Hz) and gamma (γ) (>30 Hz) [19]. In our work, the discrete wavelet transform (DWT) is utilized to decompose the EEG signals in the frequency domain, and the approximate rhythm-related frequency bands are obtained to achieve the purpose of frequency domain transition. EEG signals are discrete and DWT is utilized:

$$D\left(j,k\right) = \frac{1}{\sqrt{2^j}}\int_{-\infty}^{+\infty} s\left(t\right)\Psi\left(\frac{t}{\sqrt{2^j}} - k\right)dt \tag{2}$$

where $j, k \in \mathbb{Z}$, the translation and scaling factors are discretized into 2^j and $k\cdot2^j$. The EEG signal is defined as $s(t)$. Following the work of previous researchers [25], we utilize

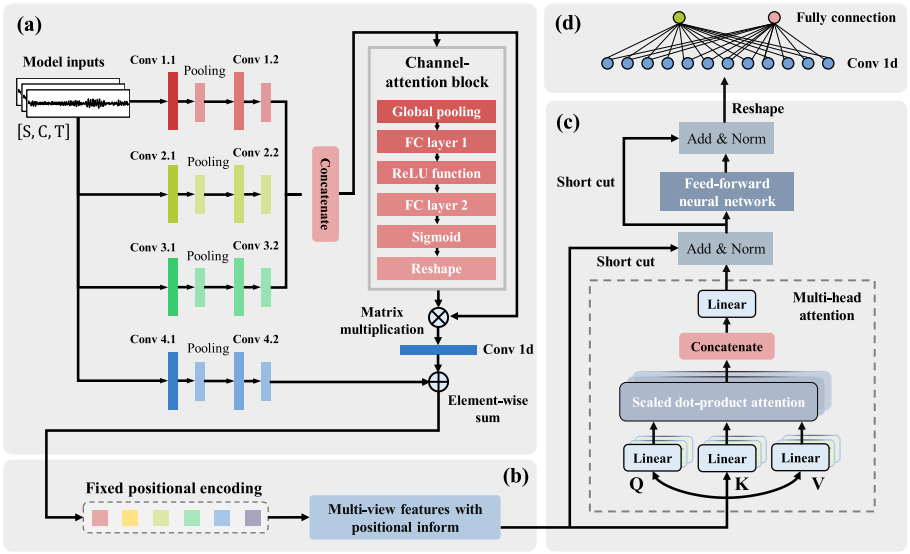

Fig. 2. The detail of the proposed hybrid Transformer model. (a) Rhythm embedding block. (b) Addition of positional encoding information. (c) Self-attention block. (d) Classifier block.

a sym6 wavelet to decompose the raw ECG sequence by six scales. Furthermore, the DWT may be interpreted in terms of a multiresolution analysis, where a hierarchy of approximations and details of the signal is constructed in nested subspaces of $\mathbf{L}^2(\mathbb{R})$. The multiresolution decomposition at level L is defined as [26]:

$$\begin{aligned} s(t) &= \sum_{k=-\infty}^{+\infty} a_{L,K} 2^{-\frac{L}{2}} \Phi\left(2^{-L}t - k\right) \\ &\quad + \sum_{j=-\infty}^{L} \sum_{k=-\infty}^{+\infty} D(j,k) 2^{-\frac{j}{2}} \Psi\left(2^{-j}t - k\right) \\ &= A_L(t) + \sum_{j=-\infty}^{L} d_j(t) \end{aligned} \tag{3}$$

where $\Phi(t)$ is a scaling function. The signal $s(t)$ is decomposed into one approximation $A_L(t)$ at level L and a succession of details $d_j(t)$ from level L down to negative infinity. Then we obtain six detailed components d_1-d_6 (64–128 Hz, 32–64 Hz, 16–32 Hz, 8–16 Hz, 4–8 Hz, 2–4 Hz) and one approximate component a_6 (0–2 Hz). We observe that an EEG rhythm corresponds to one or two wavelet transform components, and we concatenate the component signals of all scales and obtain the new signal:

$$\begin{aligned} Mix\_Rhythms_i &= Concat(\boldsymbol{\delta}, \boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}) \\ &= Concat(\mathbf{d_1}, \mathbf{d_2}, \mathbf{d_3}, \mathbf{d_4}, \mathbf{d_5}, \mathbf{d_6}, \mathbf{a_6}) \end{aligned} \tag{4}$$

where $Mix\_Rhythms_i$ is the new signal mixed with five rhythms of EEG segment sample i. In order to unify the model input, we utilize the SciPy signal processing module to resample the mixed rhythm signal to a length of 1024. Signals containing different EEG rhythms help the Transformer model to capture the correlation between different rhythms and the self-correlation of the same rhythm.

C. The Structure of Proposed Hybrid Transformer Model

The Hybrid Transformer model primarily consists of four parts: rhythm embedding block, addition of positional encoding, self-attention block, and classifier block. Fig. 2 illustrates the structure of the model, and more details can be found in Table II.

TABLE II
DETAIL OF HYBRID TRANSFORMER MODEL

Module	*Layer*	*Kernel*	*Output*
Rhythm embedding block	Input	-	(S,C,T)
	Conv1D 1.1	13, stride 1	(S,32,T/2)
	Conv1D 1.2	7, stride 1	(S,64,T/4)
	Conv1D 2.1	11, stride 1	(S,32,T/2)
	Conv1D 2.2	5, stride 1	(S,64,T/4)
	Conv1D 3.1	7, stride 1	(S,32,T/2)
	Conv1D 3.2	3, stride 1	(S,64,T/4)
	Conv1D 4.1	1, stride 1	(S,32,T/2)
	Conv1D 4.2	1, stride 1	(S,64,T/4)
	Concatenate	-	(S,192,T/4)
	Global average pooling	-	(S,192)
	FC 1	-	(S,192/16)
	FC 2	-	(S,192)
	Reshape	-	(S,192,1)
	Conv1D 5.1	1, stride 1	(S,64,T/4)
Self-attention block	Input	-	(S,64,T/4)
	Multi-head attention	-	(S,64,T/4)
	FNN Conv1D	1, stride 1	(S,512,T/4)
	FNN Conv1D	1, stride 1	(S,64,T/4)
Classifier block	Input	-	(S,64,T/4)
	Reshape	-	(S,T/4,64)
	Conv1D	1, stride 1	(S,64)
	FC	-	(S,2)

1) Rhythm Embedding Block: To obtain the multi-perspective features of inputs, we design four convolutional branches, which will facilitate multi-scale analysis of EEG spatial-temporal features. The feature vectors extracted by three convolutional branches are fused and connected to a channel attention block to evaluate the importance of channels. To avoid feature weakening, a shortcut convolutional branch is utilized to strengthen the original features. Rhythm embedding block is uniquely designed for the EEG rhythms, which can focus on the importance of the channel automatically and extract multi-view spatial-temporal features.

In this block, the input data are shaped as (S,C,T), where S, C and T denote the number of samples, channels, and data

length, respectively. C, T are set to 18 and 1024, respectively. Each convolutional branch contains two convolutional layers with different kernel size and two max-pooling layers with kernel size 2. With the padding parameters of 'same', the data length from the n *th* convolutional layer becomes half of the (n−1) *th* layer. After stacking the feature vectors obtained by the first three convolutional branches along the channel dimension, we obtain feature vectors of shape (S,192,T/4). Then a Squeeze-and-excitation network (SENet) is connected to pay attention to the channel importance, and a convolutional 1D layer with a kernel size of 1 is followed to compress the channels to 64 and get the shape of (S,64,T/4). Finally, the feature vectors are summed with the feature vectors from the fourth convolutional branch to preserve the original shallow features and make preparation for subsequent input of the self-attention block.

2) Positional Encoding: The output features of rhythm embedding block are added with the positional encoding [25]:

$$PE_{(pos,2i)} = \sin\left(pos/10000^{2i/d_{rhythm}}\right) \tag{5}$$

$$PE_{(pos,2i+1)} = \cos\left(pos/10000^{2i/d_{rhythm}}\right) \tag{6}$$

where $d_{rhythm} \in \mathbb{R}^{64}$ is the embedded dimension of feature sequence. The introduction of positional encoding enables the relative position information of the features to be represented, which helps the model learn the dependencies between the features. Considering the irregular variability of EEG signals, two other cases are also explored, which are 'no positional encoding' and 'trainable positional encoding', respectively.

3) Self-Attention Block: Then, the extracted features with positional information is connected to three Transformer encoder layers for further feature calculation with self-attention mechanism. The Transformer encoder block consists of a multi-head attention layer, a point-wise feed-forward network (FFN) layer and short connections. The multi-head attention layer contains scaled dot-product attention implemented by matrices queries $Q \in \mathbb{R}^{d_{rhythm} \times d_k}$, keys $K \in \mathbb{R}^{d_{rhythm} \times d_k}$, and values $V \in \mathbb{R}^{d_{rhythm} \times d_v}$. Q and K have the same pre-dimension for the dot product operation. The attention matrix is assigned through V and activated by the Softmax function. The resulting score matrix represents the contribution value of the feature in the entire classification. To prevent the gradient of Softmax from being too small to update the parameters, the dot product of Q and K is scaled with the scaling factor $1/\sqrt{d_k}$. The scaled dot-product attention could be calculated as follows [27]:

$$Attention\,(Q, K, V) = Soft\max\left(QK^T/\sqrt{d_k}\right)V \tag{7}$$

Instead of single attention function, multi-head attention allows the model pay attention to information of different representation subspaces at different position in parallel:

$$MultiHead\,(Q, K, V) = Concat\,(head_1, \ldots, head_h)\,W^O \tag{8}$$

$$head_i = Attention\left(QW_i^Q, KW_i^K, VW_i^V\right) \tag{9}$$

where $W_i^Q \in \mathbb{R}^{d_{rhythm} \times d_k}$, $W_i^K \in \mathbb{R}^{d_{rhythm} \times d_k}$, $W_i^V \in \mathbb{R}^{d_{rhythm} \times d_v}$, and $W^O \in \mathbb{R}^{hd_v \times d_{rhythm}}$. h denotes the number of heads.

In addition to attention sub-layers, each encoder contains a fully connected FFN, which is applied to each position separately and identically. This consists of two linear transformations with a ReLU activation in between:

$$FFN\,(x) = \max\,(0, xW_1 + b_1)\,W_2 + b_2 \tag{10}$$

While the linear transformations are the same across different positions, they use different parameters from layer to layer. Another way of describing this is as two convolutions with kernel size 1. The dimensionality of input and output is $d_{rhythm} = 64$, and the inner-layer has dimensionality $d_{ff} = 2048$.

4) Classifier Block: With the implementation of the self-attention mechanism, we get the feature vectors of shape (S,64,T/4). Through the reshape layer, we get feature vectors of shape (S,T/4,64), then a convolutional 1D layer with kernel size 1 compresses the channel dimension to 1, and finally a fully connected layer with an input dimension of 64 and output dimension of 2 is utilized to output the predicted probability value of the binary classification.

D. Training Settings of General Model

In Transformer encoder block, d_q, d_k, d_v, and d_{rhythm} are both equal to 64, and the hidden layer dimension d_{ff} is 512. The number of encoder layers is set to 3, and the number of attention heads in each layer is 4. Finally, a FFN layer is connected to compress 256 channels. The model is trained for 50 epochs with a batch size of 64. More details are as follows:

1) Adam Optimizer: The Adam optimizer [28] is used to obtain the adaptive learning rate of different parameters and updates weights and biases of the model. The recommend parameter $\beta_1 = 0.9, \beta_2 = 0.999, \varepsilon = 10^{-8}$ are utilized. And we varied the learning rate over the course of training, according to the formula [27]:

$$lrate = d_{rhythm}^{-0.5} \times \min\begin{pmatrix} step_{num}^{-0.5}, step_{num} \\ \times warmup\_steps^{-1.5} \end{pmatrix} \tag{11}$$

This corresponds to increasing the learning rate linearly for the first $warmup\_steps$ training steps, and decreasing it thereafter proportionally to the inverse square root of the step number. We used $warmup\_steps = 4000$.

2) Focal Loss Function: Since positive pre-ictal samples are much less than inter-ictal samples, the class-imbalance problem will cause the model fail to learn the relevant information from fewer samples, resulting in model performance degradation. Focal loss aims to reduce the weight of classes with a large number of samples [29]. The formula is:

$$Focal\_loss\,(p_t) = -\alpha_t\,(1 - p_t)^{\gamma}\log\,(p_t) \tag{12}$$

where p_t denotes the predicted probability of belonging to the true class. γ represents the focusing parameter, which smoothly adjusts the weight ratio of easy-to-classify samples. In this paper, we set γ to 2.

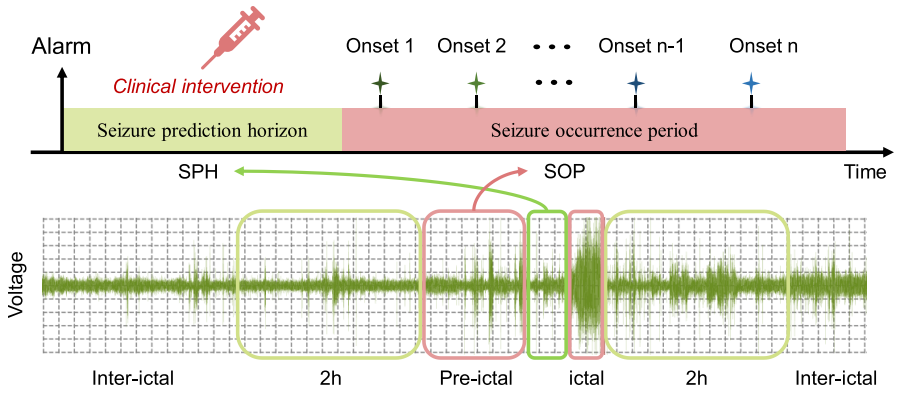

Fig. 3. Definition of inter-ictal and pre-ictal state. The corresponding relationship between SOP, SPH and labels of EEG signals is also given.

3) Early Stopping: The number of training iterations relies on the early stopping strategy so that the training process is stopped in case there is no reduction of the validation loss for five consecutive epochs. The weights from the epoch with the minimal validation loss are chosen for evaluation [30].

E. Details of Fine-Tuning Strategy

In order to further improve the performance of the general model based on patient-independent approach, the fine-tuning strategy of TL is introduced. We continue to train for 15 epochs utilizing the optimal weight parameters of the general model with a batch size of 64. The optimal weights after fine-tuning are saved by monitoring the highest accuracy on the validation set. Considering practical factors, we select the patient's first 60-min inter-ictal and 25-min pre-ictal recordings corresponding to the first seizure, which will help predict subsequent unknown data. In our work, two TL strategies are evaluated: fine-tuning based on inter-ictal samples only (three cases: 15min, 30min, 60min) and fine-tuning based on both inter-ictal and pre-ictal samples (three cases: 15min-25min, 30min-25min, 60min-25min).

III. Experimental Details

A. Definition of Inter-Ictal and Pre-Ictal State

The epilepsy prediction can be transformed into a binary classification problem (i.e., pre-ictal state and inter-ictal state). The definition of pre-ictal and inter-ictal state is crucial due to wrong labels will mislead the supervised learning algorithm and confuse the features between these two categories, which make the deep learning model unable to learn useful knowledge.

Before defining the scope of inter-ictal and pre-ictal state, we also need to consider a reality factor that patients and physicians should be given a reaction time before an epileptic seizure, known as seizure prediction horizon (SPH). The seizure occurrence period (SOP) is also defined, as shown in Fig. 3. Seizures can occur at any time in SOP. The setting of SPH should provide enough time for doctors to conduct clinical intervention, the optimal time is 3-5 min [18], and the time of SOP should not be too long, otherwise it will bring greater psychological pressure to patients. Referring to other literatures [14], [15], we set the SPH to 5 min and the SOP to 25 min in this paper. Since the location of the phase transition from the inter-ictal to pre-ictal state cannot be determined. To avoid interference, we defined the period of 2h before the pre-ictal period and 2h after the end of the seizure as the inter-ictal period. Fig. 4 illustrates the division method in detail.

B. Experimental Grouping and Data Construction

The CHB-MIT scalp EEG database contains data on 24 cases. Since case *Chb24* has no information about the subject, this case will not be studied. And the EDF files of cases *Chb12* and *Chb13* have frequent EEG channel changes. During the electrode changes, the EEG data may be contaminated, so we also do not study these two cases. Finally, the remaining 21 cases are obtained. To basically meet the 8:2 ratio of training set and test set, we decide to select 17 cases as training data and 4 cases as test data. To further investigate the generalization performance of epilepsy prediction algorithms across different subjects, we increase the variability between subjects of the test set as much as possible. Finally, we select 4 cases, including a pair of male subjects (*Chb04* and *Chb10*) and a pair of female subjects (*Chb06* and *Chb19*) with the largest difference of age, respectively.

According to the pre-ictal range defined in this paper, pre-ictal EEG data can only be obtained from the range of 5 to 30 min before an epileptic seizure, which is much less than data of the inter-ictal state. At the same time, due to the discontinuity of EDF files, pre-ictal samples are often missing. We find that there are many cases where this situation occurs, which will further increase the rarity of pre-ictal samples.

In some literatures [18], [30], a sliding window is utilized for generating pre-ictal samples and alleviate class imbalance. In our work, we utilize a sliding window of 5s and a step size of 2.5s to generate pre-ictal trials. For cases in the training set, we produce 4h inter-ictal data that meet the definition of inter-ictal state for each subject without taking overlapping sliding windows to extract EEG trials.

C. Evaluation Approach

The algorithm is evaluated using six metrics: accuracy (Acc, the ratio of the number of correctly classified samples among the total number of samples), sensitivity (Sen, the ratio of correctly classified events to all true events), specificity

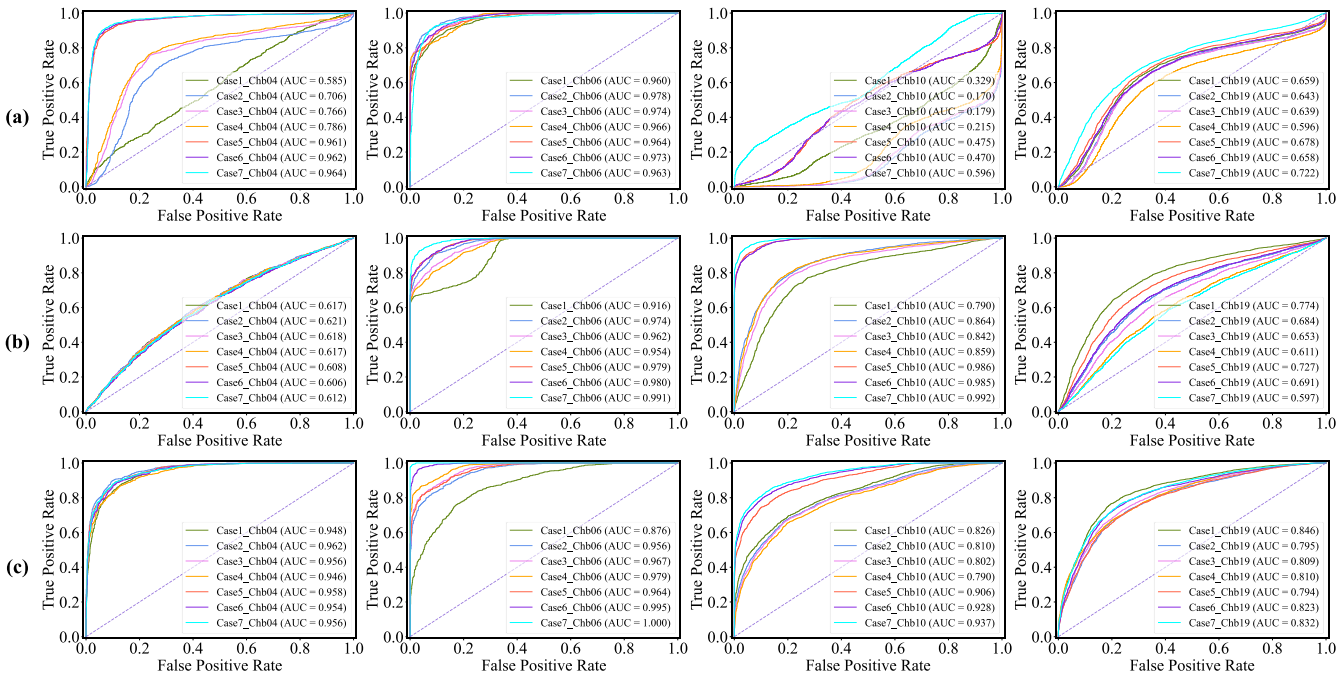

Fig. 4. ROC curves of four subjects utilizing proposed hybrid Transformer model. (a) Raw EEG signal. (b) EEG spectrogram. (c) Mixed EEG rhythm signal. Case 1 represents patient-independent approach. Case 2, Case 3, and Case 4 represent fine-tuning with 15-min, 30-min, and 60-min inter-ictal data, respectively. Case 5, Case 6, and Case 7 represent fine-tuning with both 15-min inter-ictal mixed 25-min pre-ictal, 30-min inter-ictal mixed 25-min pre-ictal, and 60-min inter-ictal mixed 25-min pre-ictal data, respectively. The first to fourth columns are case Chb04, Chb06, Chb10, and Chb19, respectively.

(Spe, the ratio between correctly classified non-events and all non-events), positive predictive rate (Ppr, the ratio of correctly classified events in all recognized events), F1 value ($F1_{Binary}$, the harmonic mean of Sen and Ppr), and weighted F1 value ($F1_{Weighted}$, denote the F1 value that takes into account the sample imbalance problem) [31]. The calculation formulas are as follows:

$$Accuracy\,(\%) = \frac{TP + TN}{TP + TN + FP + FN} \times 100 \tag{13}$$

$$Sensitivity\,(\%) = \frac{TP}{TP + FN} \times 100 \tag{14}$$

$$Ppr\,(\%) = \frac{TP}{TP + FP} \times 100 \tag{15}$$

$$F1_{Binary}\,(\%) = \frac{2 \times Sen \times Ppr}{Sen + Ppr} \times 100 \tag{16}$$

$$F1_{Weighted}\,(\%) = \frac{\sum_{category} j_{category} F1_{Binary}}{Sen + Ppr} \times 100 \tag{17}$$

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

IV. Experimental Results

A. Performance of Hybrid Transformer Model

We conduct a comprehensive evaluation of our proposed hybrid Transformer, including the applicability of three different model inputs and two TL conditions. Different model input types help to reveal the data form suitable for self-attention mechanism while two TL conditions correspond to two clinic situations. The first situation represents no pre-ictal data are available for the subjects, so we fine-tune the model only using inter-ictal data, which is divided into three cases, which are 15-min, 30-min, and 60-min of inter-ictal data. This situation will help to reveal whether the utilization of easy-to-obtain inter-ictal data is beneficial for a patient-specific model. The second TL situation is based on at least one seizure of subject, and the performance improvement to the general model is evaluated. After one seizure, we will obtain a pre-ictal recording with a length of 25 min based on the definition of SOP. By combining the 25-min pre-ictal data with the 15-min, 30-min, and 60-min inter-ictal data respectively, we get three combinations. Finally, six fine-tuning strategies are evaluated for each subject.

The experimental results are present in Table II and Fig. 4. In Table II, the optimal performance is achieved based on the feature engineering method proposed in this paper, which proves that the signal containing different rhythm information is more suitable for the Transformer-based model. For the six fine-tuning strategies, we can conclude that fine-tuning based only on the inter-ictal data is beneficial to the model performance. By comparing *Case* 3 and *Case* 4, the accuracy decreases as the amount of inter-ictal data increases, proving that fine-tuning with too much inter-ictal data may lead to overfitting and degradation of model performance. We also find that *Case* 7 achieve the best performance, which will guide the fine-tuning of the patient-specific classifier when the clinically labeled data meets the condition with at least one seizure.

From Fig. 4 we can see that the proposed model input is more robust to different subjects compared with two classical methods. We also observe that the performance of *Chb19* subjects is not satisfactory under three model inputs, probably

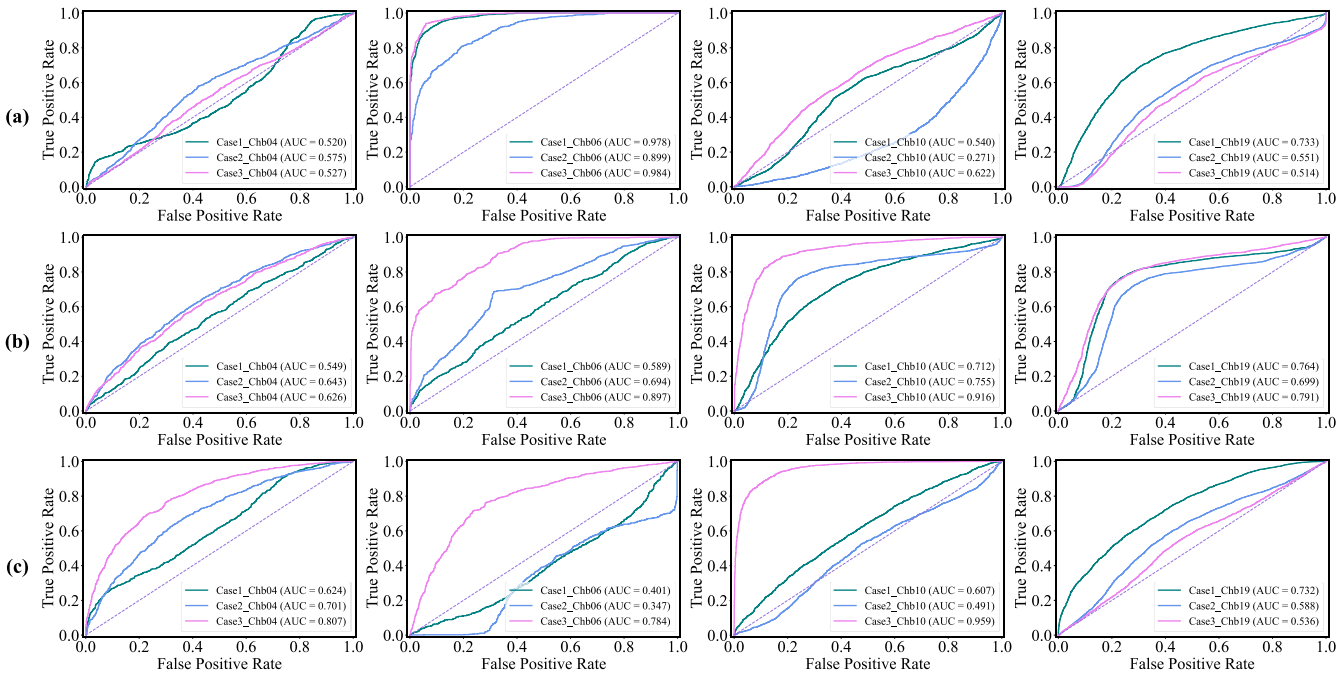

Fig. 5. ROC curves of four subjects utilizing pure CNN-based model. (a) Raw EEG signal. (b) EEG spectrogram. (c) Mixed EEG rhythm signal. Case 1 represents patient-independent approach. Case 2 represents fine-tuning with 30-min inter-ictal data. Case 3 represents fine-tuning with both 60-min inter-ictal data and 25-min pre-ictal data. The first to fourth columns are case Chb04, Chb06, Chb10, and Chb19, respectively. The pure CNN-based model shows weaker overall performances than the proposed hybrid Transformer model under three feature engineering methods, and reflects poor robustness to TL.

because its distribution is far from the distribution of the training set, which makes the model unable to generalize well.

B. Impact on Performance With Different Model Structures

To explore the effect of different model inputs on pure CNN-based networks and Transformer-based structures, we reproduce a recent state-of-the-art multi-scale network based on pure CNNs [31], it has the same sample inputs size of 1024, then we compare the model performance utilizing three model inputs, which will help reveal the ability of pure CNN-based models to learn different inter-rhythmic patterns.

The experimental results are shown in Fig. 5. We compare patient-independent approach with two fine-tuning strategies (*Case* 3 and *Case* 7) in Table II. By comparison with Fig. 4, it can be seen that the pure CNN-based model is less effective for three model inputs, and the results after fine-tuning are also unsatisfactory. For the model input of mixed rhythm signal, the CNN-based model does not seem to have learned relevant knowledge due to its locality [32]. With the limited receptive field of the convolution kernel, the signals consist of multiple rhythms seem to be irregular. However, the Transformer-based model can learn the pattern between rhythms through global modeling, which confirms that the mixed rhythms signal proposed in this paper is effective for the hybrid Transformer model.

C. Effectiveness of Three Transformer Variants

Since the input length of EEG is 1024, utilizing the self-attention mechanism to model the feature sequence will lead to a huge consumption of computing power, so the dimension reduction operation is necessary. In this paper, a rhythm embedding block based on multi-scale CNN networks and SENet is utilized to reduce the dimension of the feature sequence. This block converts the original feature sequence into an embedded representation and has the ability of capturing multi-scale information and channel importance. To demonstrate the effectiveness of this block, we compare a simple embedding block based on two CNN layers and two max-pooling layers. Then we compare other two forms of positional encoding ('without positional encoding' and 'trainable positional encoding'). Finally, three variant Transformer models are formed and compared with the proposed model.

The results in Table IV show that the hybrid Transformer model achieves the best average performance for three cases under the proposed rhythm embedding block with untrainable positional encoding. We also find that *Case 7* achieves the highest accuracy when utilizing trainable positional encoding, which may indicate that it may be more appropriate to utilize trainable positional encoding with sufficient fine-tuned data.

D. Comparison With Other State-of-the-Art Works

We compared the model performance with the state-of-the-art works in recent years (Table V). Traditional feature engineering methods such as phase/amplitude locking value, zero crossing, and similarity/dissimilarity index were considered. We can see that satisfactory sensitivity and FPR cannot be achieved simultaneously with these traditional feature methods. Machine learning-based methods such as SVM and CNN

TABLE III
COMPARISON OF THREE MODEL INPUTS AND SEVEN FINE-TUNE STATEGIES

Fine-tune Strategy	*Subject ID*	*Filtered Raw EEG*						*EEG Spectrogram*						*Mixed Rhythm signal*					
		ACC	*Sen*	*Ppr*	*F1*	*F1$_W$*	*AUC*	*ACC*	*Sen*	*Ppr*	*F1*	*F1$_W$*	*AUC*	*ACC*	*Sen*	*Ppr*	*F1*	*F1$_W$*	*AUC*
Case 1	Chb04[1]	54.5	49.1	80.1	60.9	52.4	0.585	59.1	57.8	28.5	38.2	55.6	0.617	87.2	81.4	92.3	86.5	87.3	0.948
	Chb06[1]	87.8	89.2	66.4	76.2	87.2	0.960	76.6	96.9	20.8	34.3	70.7	0.916	79.1	69.8	50.3	58.5	78.0	0.876
	Chb10[1]	38.6	56.6	54.6	55.6	39.3	0.329	74.4	97.0	65.7	78.3	75.5	0.790	73.9	99.5	63.2	77.3	74.9	0.826
	Chb19[1]	66.3	86.5	67.8	76.0	69.0	0.659	72.3	84.8	79.0	81.8	73.3	0.774	78.2	80.2	96.0	87.4	72.9	0.846
	Average[1]	61.8	70.4	67.2	67.2	62.0	0.633	70.6	84.1	48.5	58.2	66.1	0.774	79.6	82.7	75.5	77.4	78.3	0.874
Case 2	Chb04[2]	70.9	80.1	45.6	58.1	69.0	0.706	59.1	55.9	35.5	43.4	57.1	0.621	89.9	86.0	90.6	88.3	89.8	0.962
	Chb06[2]	91.6	91.4	78.5	84.5	91.4	0.978	90.5	93.9	72.1	81.6	90.1	0.974	86.9	88.0	64.1	74.2	86.3	0.956
	Chb10[2]	30.0	51.8	0.09	15.4	22.8	0.170	79.4	93.0	76.5	83.9	80.2	0.864	72.9	99.5	61.7	76.2	73.9	0.810
	Chb19[2]	65.5	87.8	65.2	74.8	68.5	0.643	65.4	87.9	65.0	74.7	68.4	0.684	73.3	80.7	86.9	83.7	71.6	0.795
	Average[2]	64.5	77.8	47.3	58.2	62.9	0.624	73.6	82.7	62.3	70.9	70.6	0.786	80.8	88.6	75.8	80.6	80.4	0.881
Case 3	Chb04[3]	76.0	88.3	52.8	66.1	74.6	0.766	58.7	56.6	28.5	37.9	55.2	0.618	88.5	89.8	83.4	86.5	88.4	0.956
	Chb06[3]	89.6	94.0	68.8	79.5	89.1	0.974	87.4	94.7	60.4	73.8	86.5	0.962	88.5	89.9	68.5	77.7	88.0	0.970
	Chb10[3]	31.5	58.6	0.09	15.8	23.7	0.179	78.5	96.2	72.3	82.6	79.4	0.842	73.0	99.6	61.8	76.3	74.0	0.802
	Chb19[3]	65.4	88.3	64.7	74.7	68.5	0.639	63.1	88.5	61.0	72.2	66.4	0.653	74.3	87.1	81.5	84.2	72.8	0.809
	Average[3]	65.6	82.3	46.6	59.0	64.0	0.640	71.9	84.0	55.6	66.6	68.5	0.769	81.1	91.6	73.8	81.2	80.8	0.884
Case 4	Chb04[4]	76.2	80.3	61.4	69.6	75.7	0.786	58.7	58.2	23.6	33.6	53.9	0.617	86.9	91.4	77.8	84.0	86.8	0.946
	Chb06[4]	86.5	95.0	57.0	71.3	85.4	0.966	85.5	95.2	53.0	68.1	84.0	0.954	90.6	98.6	68.8	81.0	90.0	0.979
	Chb10[4]	35.4	78.9	11.1	19.5	27.3	0.215	79.8	95.6	74.7	83.9	80.6	0.859	71.5	99.6	59.7	74.7	72.5	0.790
	Chb19[4]	61.7	92.0	56.3	69.9	65.2	0.596	59.2	88.4	55.5	68.2	62.9	0.611	72.9	78.8	85.6	82.1	74.1	0.810
	Average[4]	65.0	86.6	46.5	57.6	63.4	0.641	70.8	84.4	51.7	63.5	67.1	0.760	80.5	92.1	73.0	80.5	80.9	0.881
Case 5	Chb04[5]	90.7	95.0	83.4	88.8	90.6	0.961	57.7	56.3	20.3	29.8	52.1	0.608	89.0	90.1	84.4	87.2	89.0	0.958
	Chb06[5]	89.3	83.4	79.2	81.2	89.2	0.964	90.9	94.8	72.8	82.4	90.5	0.979	87.9	92.3	64.1	75.6	87.2	0.964
	Chb10[5]	52.4	84.7	39.4	53.8	52.9	0.475	92.5	99.9	89.4	94.4	92.7	0.986	82.1	99.9	74.6	85.4	82.9	0.906
	Chb19[5]	67.0	84.8	70.7	77.1	69.4	0.678	68.9	84.7	73.8	78.9	70.8	0.727	72.9	81.4	85.2	83.2	72.0	0.794
	Average[5]	74.9	87.0	68.2	75.2	75.5	0.770	77.5	83.9	64.1	71.4	71.6	0.825	83.0	90.9	77.1	82.9	82.8	0.906
Case 6	Chb04[6]	91.2	97.3	82.5	89.3	91.1	0.962	57.4	55.2	19.6	28.9	51.5	0.606	88.5	89.0	84.6	86.7	88.5	0.954
	Chb06[6]	90.0	85.8	78.9	82.2	89.9	0.973	91.1	96.8	71.8	82.5	90.6	0.980	96.7	94.9	93.6	94.3	96.7	0.995
	Chb10[6]	50.9	92.3	32.9	48.6	49.9	0.470	92.2	99.9	89.1	94.2	92.4	0.985	84.3	99.9	77.8	87.5	85.0	0.928
	Chb19[6]	65.5	85.2	68.0	75.6	68.3	0.658	66.0	85.0	69.1	76.2	68.7	0.691	76.1	81.2	90.5	85.6	73.5	0.823
	Average[6]	74.4	90.2	65.6	73.9	74.8	0.766	76.7	84.2	62.4	70.5	70.9	0.816	86.4	91.3	86.6	88.5	85.9	0.925
Case 7	Chb04[7]	91.8	96.4	84.6	90.1	91.7	0.964	58.1	57.4	21.0	30.7	52.6	0.612	88.9	93.5	80.6	86.6	88.8	0.956
	Chb06[7]	91.6	93.8	76.2	84.1	91.3	0.963	94.4	99.2	81.5	89.5	94.2	0.991	99.0	99.0	97.7	98.3	99.0	1.000
	Chb10[7]	52.2	99.6	32.1	48.6	50.6	0.596	94.8	99.9	92.7	96.2	94.9	0.992	85.6	99.9	79.6	88.6	86.2	0.937
	Chb19[7]	68.6	96.1	62.7	75.9	71.4	0.722	58.7	86.7	56.2	68.2	62.4	0.597	76.5	83.2	87.9	85.5	75.5	0.832
	Average[7]	76.1	96.5	63.9	74.7	76.3	0.811	76.5	85.8	62.9	71.2	71.3	0.798	87.5	93.9	86.5	89.8	87.4	0.931

TABLE IV
COMPARISON OF FOUR TRANSFORMER-BASED VARIANTS

Model	*Training Time (s/epoch)*	*Fine-tune Strategy*	*Accuracy* (%)*	*Sensitivity* (%)*	*Positive Predictive Rate* (%)*	*F1 Score* (%)*	*Weighted F1 Score* (%)*	*AUC**	*Fine-tune Time* (s/epoch)*
Transformer$_{\mathrm{I}}$	110.04	Case 1	65.4	64.6	83.9	72.2	61.2	0.793	-
		Case 3	72.1	86.6	58.0	72.3	66.5	0.811	1.14
		Case 7	81.5	92.9	73.1	80.3	81.4	0.887	1.60
Transformer$_{\mathrm{II}}$	107.58	Case 1	73.7	72.1	71.8	71.9	73.7	0.825	-
		Case 3	73.7	74.8	65.8	69.8	73.8	0.822	1.14
		Case 7	86.5	92.6	78.9	85.0	86.5	0.938	1.60
Transformer$_{\mathrm{III}}$	105.60	Case 1	77.7	74.1	88.1	80.2	76.2	0.806	-
		Case 3	79.9	88.1	76.1	79.8	80.2	0.835	1.11
		Case 7	89.0	94.8	86.4	90.0	89.4	0.953	1.68
Transformer$_{\mathrm{IV}}$	116.28	Case 1	79.6	82.7	75.5	77.4	78.3	0.874	-
		Case 3	81.1	91.6	73.8	81.2	80.8	0.884	1.14
		Case 7	87.5	93.9	86.5	89.8	87.4	0.931	1.56

X* denote the average performance of four subjects.

have become a trend, which can obtain better comprehensive indicators. In our proposed method, the FPR is 0.00% in all three cases, which significantly reduces the occurrence of false alarms. In the complete patient-independence paradigm, the average sensitivity is 77.0% (Note that the sensitivity is evaluated by seizures, which is different from the sensitivity in Table III). When the model is fine-tuned with 30-min inter-ictal data, the sensitivity reaches 82.0%. When fine-tuned with both the 25-min pre-ictal and 60-min inter-ictal data based on one seizure, the model achieves the best sensitivity of 91.7%.

V. DISCUSSION

As shown in Table V, compared with sEEG, iEEG has a higher signal-to-noise ratio and is less affected by power frequency interference, baseline drift and other noises, which will help the seizure prediction algorithm to achieve better performance. However, to collect iEEG data, craniotomy would cause unnecessary complications, sEEG is more suitable for non-invasive real-world scenarios. Now, more and more researchers are evaluating algorithms based on the CHB-MIT database [10], [11], [12], [14], [24], [33].

TABLE V
COMPARISON TO PRIOR WORKS

Works	*Datasets*	*Training Method*	*Feature Engineering*	*Classifier*	*Total Seizures*	*Sen(%)*	*FPR(/h)*	*SOP/SPH*
Truong et al. [14]	CHB-MIT	LOO/13 subjects	Short-time Fourier transform	CNN	64	81.2	0.16	30/5 min
Ozcan et al. [33]	CHB-MIT	LOO/16 subjects	Spectral power, Statistical moments, Hjorth parameters	3D CNN	77	85.7	0.10	30/1 min
Zhang et al. [24]	CHB-MIT	LOO/23 subjects	Combination of common spatial pattern statistics	CNN	156	92.2	0.12	30/- min
Myers et al. [10]	CHB-MIT	Validation on each subject/10 subjects	Phase/Amplitude locking value	-	31	77.0	0.17	60/- min
Zandi et al. [12]	CHB-MIT	Validation on each subject/3 subjects	Zero crossing, similarity/ dissimilarity index	-	18	83.8	0.17	40/2 min
Cho et al. [11]	CHB-MIT	10-fold cross-validation with random selection for each subject/21 subjects	Phase locking value	SVM	65	82.4	-	30/3 min
Aarabi et al. [34]	Freiburg Hospital	Random subsampling cross-validation of each subject /10 subjects	Univariate and bivariate features	SVM	28	86.7	0.13	30/$\frac{1}{6}$ min
Wang et al. [15]	Freiburg Hospital	K-fold cross-validation of each subject/19 subjects	DTF channel-frequency maps	CNN	82	90.8	0.08	30/5 min
Proposed[1]	CHB-MIT	17 subjects for training/ 4 subjects for testing	Wavelet Transform	Hybrid Transformer	24	*Chb04*[1]	100.0	0.00	25/5 min
						Chb06[1]	70.0	0.00	
						Chb10[1]	71.4	0.00	
						Chb19[1]	66.7	0.00	
						Average[1]	**77.0**	**0.00**	
Proposed[2]						*Chb04*[2]	100.0	0.00	
						Chb06[2]	90.0	0.00	
						Chb10[2]	71.4	0.00	
						Chb19[2]	66.7	0.00	
						Average[2]	**82.0**	**0.00**	
Proposed[3]						*Chb04*[3]	100.0	0.00	
						Chb06[3]	100.0	0.00	
						Chb10[3]	100.0	0.00	
						Chb19[3]	66.7	0.00	
						Average[3]	**91.7**	**0.00**	

#CHB-MIT database: scalp EEG (sEEG). Freiburg Hospital: intracranial EEG (iEEG). LOO: Leave-One-Out method.

It can be seen in Fig. 4 that the model performance for *Chb10* based on filtered raw EEG signals is not well, but the model input based on spectrogram and rhythm can be well recognized. When processing the raw EEG signal, we utilize a band-pass filter to limit the frequency of the signal to 5-50 Hz. Due to the absence of high frequency signal, the information that would enable the model to distinguish between inter-ictal and pre-ictal periods may be lost, which proves the gamma frequency band is useful for discriminating between inter-ictal and pre-ictal stages [11].

In the evaluation of TL based on pure CNNs and Transformer structures, we find that the network based purely on CNN cannot transfer the existing knowledge well, because the innate inductive bias of the CNN structure is easy to overfitting the data. As can be seen from Fig. 5, the performance improvement of pure CNN-based networks is not robust after the introduction of TL, and sometimes it will produce harmful effects to the model. By comparing the structure of different deep learning models, it is revealed that the model based on the Transformer structure is more suitable for the introduction of TL.

The attention mechanism helps to reveal regions that contribute more to the label, which is more in line with the recognition process of human experts [35]. To show how the multi-head attention mechanism works more intuitively, we show the attention maps from four attention heads and the corresponding rhythm signal in Fig. 6. The results show that different attention heads learned in parallel will pay attention to patterns between different rhythms, increasing interpretability of the detection process.

In addition, it is worth noting that in Table V, many works [14], [24], [33] have utilized the leave-one-out (LOO) method to validate algorithms, where the arithmetic mean of the model for each subject yields a more robustness evaluation of the proposed model. However, because the weights of each model are inconsistent, we cannot evaluate the performance by fine-tuning one general model for multiple subjects. This paper changes the grouping idea, we utilize 17 subjects to

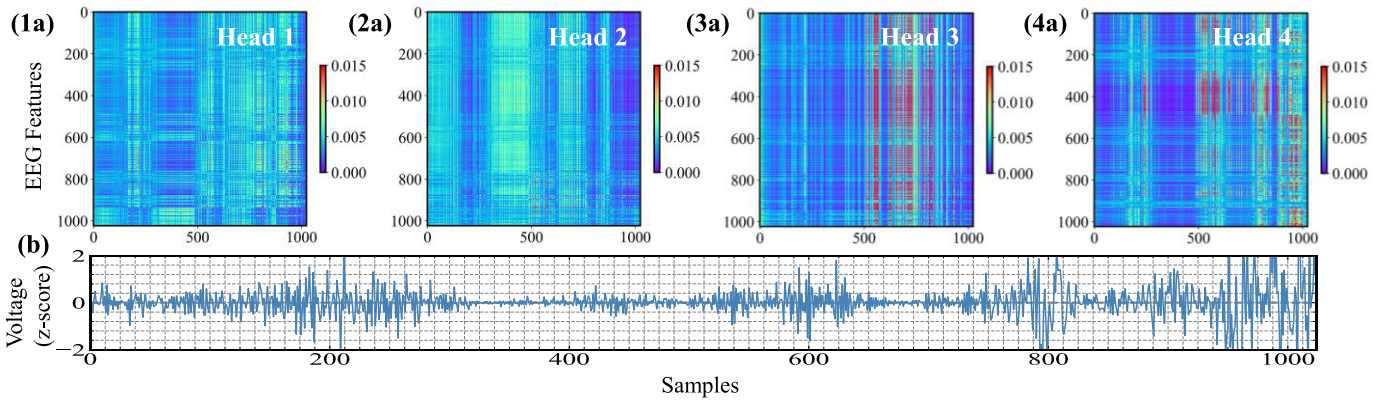

Fig. 6. Visualization of self-attention mechanism. (1a)-(4a) show the attention maps of four attention heads and (b) is the corresponding mixed rhythm signal. The abscissa of the attention map is stretched to make it the same length as the 5 s EEG segments. It can be observed that not all attention heads pay attention to the pre-ictal signals because different attention heads focus on different patterns. The redundant attention heads together are able to help recognize complicated patterns and then yield better feature representations of inter-ictal and pre-ictal segments.

train the general model and evaluate the 4 subjects with the largest physiological differences, which is more in line with the application scenario in clinic. Moreover, the good robustness of Transformer-based model is demonstrated during the process of TL optimization in the case that the difference of test set is large. Meanwhile, the self-attention mechanism and multi-rhythms analysis are beneficial to capture the correlation patterns among different rhythms.

However, this paper still has some limitations. Since many EEG epilepsy databases are not freely available, they come from different intracranial EEG data and scalp EEG data, and the electrode standards are not consistent, so the cross-database validation of fine-tuning strategies is limited. Moreover, the frequency band of wavelet transform decomposition has a characteristic of two-fold reduction, which will result in some minor errors at the boundary of different rhythm frequency band, and often losing a small portion of useful rhythm information.

Through our investigation, there are some interesting feature engineering methods to distinguish ictal and non-ictal EEGs in epilepsy detection tasks, which are of significantly morphological difference [36], [37], [38], [39]. For example, the work of [36] utilized rhythms obtained with Fourier–Bessel series expansion (FBSE) of EEG signals to further classify. The work of [37] explored the ability of the second-order difference plot (SODP) of intrinsic mode functions (IMFs) for classification of ictal and non-ictal EEG signals. The empirical mode decomposition (EMD) is a promising method which helps to develop feature space using ellipse area parameters of two IMFs. The work of [38] presented a fractional-order calculus-based method to model ictal and non-ictal EEG signals. It is found that the modeling error energy for ictal EEG signals is substantially higher than that for inter-ictal EEG signals. Moreover, the work of [39] proposed a novel time–frequency representation (TFR) which is termed as improved eigenvalue decomposition of Hankel matrix and Hilbert transform (IEVDHM–HT). The IEVDHM–HT method has provided better TF resolution compared with the existing methods in terms of Rényi entropy measure (REM) values. The above methods have the advantages of high computational efficiency than DL algorithms due to low dimensions of features, and show good performance in distinguishing different classes with large morphological differences. Moreover, signal frequency band decomposition based on FBSE may help more accurate rhythm band estimation and gain better performance compared with DWT utilizing in this paper.

Future research can focus on the investigation of these methods' improved versions and apply to inter-ictal and pre-ictal states, which have less significant differences in waveform morphology. The applicability of these feature engineering methods and traditional classifiers, e.g., SVM, random forest (RF), and k nearest neighbors (KNN) for epilepsy prediction task and their advantages for DL models in the condition of low data availability is also a direction worth exploring for further research.

VI. Conclusion

In this paper, we propose a novel hybrid Transformer network with a specially designed feature engineering method, the effectiveness of TL and different model input for pure CNN-based and Transformer-based models is also discussed. The experimental results show that the proposed model and feature engineering method have good adaptability. In this work, four patients with the largest difference in physiological parameters are selected from the CHB-MIT database for experiments and to provide three model paradigms. The experimental results show that the model achieves significant performance under the patient-independent paradigm, and the performance can be further improved by the introduction of TL, especially when fine-tuning the model utilizing the data after one seizure of the patient can achieve the state-of-the-art performance.

Compared with the purely CNN-based model structure, we find that the Transformer-based model structure has better robustness to TL, which will provide ideas for future researchers to customize personalized models for patients. Moreover, in the discussion of different model inputs, we find that gamma rhythm band corresponding to high frequency information is helpful to distinguish between pre-ictal and inter-ictal state.

REFERENCES

[1] F. Mormann, R. G. Andrzejak, C. E. Elger, and K. Lehnertz, "Seizure prediction: The long and winding road," *Brain*, vol. 139, pp. 1625–1627, Feb. 2016.

[2] J. Yang and M. Sawan, "From seizure detection to smart and fully embedded seizure prediction engine: A review," *IEEE Trans. Biomed. Circuits Syst.*, vol. 14, no. 5, pp. 1008–1023, Oct. 2020.

[3] B. Maimaiti et al., "An overview of EEG-based machine learning methods in seizure prediction and opportunities for neurologists in this field," *Neuroscience*, vol. 481, pp. 197–218, Jan. 2022.

[4] Y. Xu, J. Yang, and M. Sawan, "Multichannel synthetic preictal EEG signals to enhance the prediction of epileptic seizures," *IEEE Trans. Biomed. Eng.*, vol. 69, no. 11, pp. 3516–3525, Nov. 2022.

[5] L. D. Iasemidis et al., "Adaptive epileptic seizure prediction system," *IEEE Trans. Biomed. Eng.*, vol. 50, no. 5, pp. 616–627, May 2003.

[6] 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.

[7] M. Scheffer et al., "Early-warning signals for critical transitions," *Nature*, vol. 461, no. 7260, pp. 53–59, Sep. 2009.

[8] 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, 2004.

[9] Y. Zheng, G. Wang, K. Li, G. Bao, and J. Wang, "Epileptic seizure prediction using phase synchronization based on bivariate empirical mode decomposition," *Clin. Neurophysiol.*, vol. 125, no. 6, pp. 1104–1111, Jun. 2014.

[10] M. H. Myers, A. Padmanabha, G. Hossain, A. L. De Jongh Curry, and C. D. Blaha, "Seizure prediction and detection via phase and amplitude lock values," *Frontiers Hum. Neurosci.*, vol. 10, p. 80, Mar. 2016.

[11] 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.

[12] 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.

[13] Y. LeCun, Y. Bengio, and G. E. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, Dec. 2015.

[14] 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.

[15] G. Wang et al., "Seizure prediction using directed transfer function and convolution neural network on intracranial EEG," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 28, no. 12, pp. 2711–2720, Dec. 2020.

[16] S. Hu, W. Cai, T. Gao, and M. Wang, "An automatic residual-constrained and clustering-boosting architecture for differentiated heartbeat classification," *Biomed. Signal Process. Control*, vol. 77, Aug. 2022, Art. no. 103690.

[17] Y. Wang, G. Zhou, and C. Yang, "Interpatient heartbeat classification using modified residual attention network with two-phase training and assistant decision," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–15, 2023.

[18] J. Yan, J. Li, H. Xu, Y. Yu, and T. Xu, "Seizure prediction based on transformer using scalp electroencephalogram," *Appl. Sci.*, vol. 12, no. 9, p. 4158, Apr. 2022.

[19] A. Bhattacharya, T. Baweja, and S. P. K. Karri, "Epileptic seizure prediction using deep transformer model," *Int. J. Neural Syst.*, vol. 32, no. 2, Feb. 2022, Art. no. 2150058.

[20] B. S. Zargar, M. R. K. Mollaei, F. Ebrahimi, and J. Rasekhi, "Generalizable epileptic seizures prediction based on deep transfer learning," *Cognit. Neurodynamics*, vol. 17, no. 1, pp. 119–131, Feb. 2023.

[21] Z. Yu et al., "Epileptic seizure prediction using deep neural networks via transfer learning and multi-feature fusion," *Int. J. Neural Syst.*, vol. 32, no. 7, Jul. 2022, Art. no. 2250032.

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

[23] V. Jurcak, D. Tsuzuki, and I. Dan, "10/20, 10/10, and 10/5 systems revisited: Their validity as relative head-surface-based positioning systems," *Neuroimage*, vol. 34, no. 4, pp. 1600–1611, 2007.

[24] 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.

[25] D. Chen, S. Wan, and F. S. Bao, "Epileptic focus localization using discrete wavelet transform based on interictal intracranial EEG," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 25, no. 5, pp. 413–425, May 2017.

[26] S. G. Mallat, "A theory for multiresolution signal decomposition: The wavelet representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 11, no. 7, pp. 674–693, Jul. 1989.

[27] A. Vaswani et al., "Attention is all you need," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 6000–6010.

[28] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[29] S. Hu, W. Cai, T. Gao, J. Zhou, and M. Wang, "Robust wave-feature adaptive heartbeat classification based on self-attention mechanism using a transformer model," *Physiological Meas.*, vol. 42, no. 12, Dec. 2021, Art. no. 125001.

[30] 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.

[31] P. Thuwajit et al., "EEGWaveNet: Multiscale CNN-based spatiotemporal feature extraction for EEG seizure detection," *IEEE Trans. Ind. Informat.*, vol. 18, no. 8, pp. 5547–5557, Aug. 2022.

[32] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[33] 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.

[34] A. Aarabi and B. He, "Seizure prediction in patients with focal hippocampal epilepsy," *Clin. Neurophysiol.*, vol. 128, no. 7, pp. 1299–1307, Jul. 2017.

[35] S. Hu, W. Cai, T. Gao, and M. Wang, "A hybrid transformer model for obstructive sleep apnea detection based on self-attention mechanism using single-lead ECG," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–11, 2022.

[36] V. Gupta and R. B. Pachori, "Epileptic seizure identification using entropy of FBSE based EEG rhythms," *Biomed. Signal Process. Control*, vol. 53, Aug. 2019, Art. no. 101569.

[37] R. B. Pachori and S. Patidar, "Epileptic seizure classification in EEG signals using second-order difference plot of intrinsic mode functions," *Comput. Methods Programs Biomed.*, vol. 113, no. 2, pp. 494–502, Feb. 2014.

[38] V. Joshi, R. B. Pachori, and A. Vijesh, "Classification of ictal and seizure-free EEG signals using fractional linear prediction," *Biomed. Signal Process. Control*, vol. 9, pp. 1–5, Jan. 2014.

[39] R. R. Sharma and R. B. Pachori, "Time–frequency representation using IEVDHM–HT with application to classification of epileptic EEG signals," *IET Sci., Meas. Technol.*, vol. 12, no. 1, pp. 72–82, Jan. 2018.