ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-4485-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICASSP48485.2024.10447845

TEMPORAL-SPATIAL PREDICTION: PRE-TRAINING ON DIVERSE DATASETS FOR EEG CLASSIFICATION

Ziyi Li[⋆†] *Li-Ming Zhao*[†] *Wei-Long Zheng*[⋆] *Bao-Liang Lu* [⋆†‡]

[⋆] Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China
[†] Emotion Helper, Shanghai, China

ABSTRACT

Electroencephalogram (EEG) classification tasks have received increasing attention because its high application value. Meanwhile, the great success of general pre-training models in language processing areas inspires us to excavate the potential of an EEG pre-trained model. This model is expected to adapt to diverse downstream tasks. However, current studies either ignore the temporal or spatial domain in EEG signals, or only use single datasets in pre-training. The proposed Temporal-Spatial Prediction (TSP) model effectively solve these issues. Specifically, the output of the TSP encoder serves as the input of two tasks: spatial prediction, i.e., masked autoencoder, and temporal prediction, i.e, contrastive predictive coding. In addition, in order to provide more diverse information and thus benefit the downstream fine-tuning, we pre-train TSP on six large EEG datasets with four different numbers of channels. Results on three public downstream datasets SEED, SEED-IV, TUEV demonstrate that TSP achieves the state-of-the-art performance on different EEG classification tasks. In addition, according to the ablation experiments, TSP performs better than the single-domain method, i.e. Temporal Prediction (TP) model and Spatial Prediction (SP) model.

***Index Terms*—** EEG, Self-supervised Learning, Emotion Recognition, Transformer, CNN

1. INTRODUCTION

As an indispensable part of Brain Computer Interface (BCI), EEG plays an essential role in many fields such as depression diagnosis, emotion recognition and so on. However, for an extended period, these areas have progressed separately with inadequate integration. Given that large general language or vision models have achieved great success. The demand of a general EEG self-supervised model has emerged. By excavating information from numerous EEG signals without the need of labels at scale, this kind of model is expected to learn a generalized representation that can adapt to diverse downstream tasks regardless of the EEG recording devices, the number of EEG channels and EEG sampling rates of different tasks. The central problem of it lies on the design of pre-training tasks.

Some approaches focus on the temporal information. For example, Contrastive Predictive Coding (CPC) provides an effective way to predict features in future timesteps [1]. However, the CPC-based methods only focus on temporal domain and fail to integrate spatial or frequency information. Some studies focus on spatial domain. Li *et al* [2] adapted masked autoencoder (MAE) [3] for masked channel predicting but failed to predict time information.

Biosignal Transformer (BIOT) [4] cleverly flattens the time and channel dimensions as a series of sequences and convey them to Transformer, through which BIOT is able to deal with different number of channels or time steps. In BIOT, the same processing strategy is used for both channel and time. It is indeed a feasible and effective direction. However, Considering that EEG signal has spatial smearing characteristic [5] ,i.e. adjacent channels contain similar information and EEG signals are non-stationary, the inter-channel characteristics and inter-time characteristics are apparently different. This prompts us to consider whether we ought to use different strategies for the two domains.

Therefore, we propose the Temporal-Spatial prediction (TSP) algorithm. TSP effectively solves the following problems: 1) TSP can adapt to different EEG recording devices and number of channels, which means TSP is able to be fine-tuned on different downstream tasks. This places high demands not only on the model structure but also on the pre-training data as the pre-training data needs to include as many kinds of EEG recording devices as possible. 2) TSP takes both time and spatial (channel) information into consideration, i.e., CPC for time prediction and MAE for spatial prediction, which is proved to have improved performance on three public datasets.

‡ Corresponding author

This work was supported in part by grants from STI 2030-Major Projects+2022ZD0208500, National Natural Science Foundation of China (Grant No. 62376158), Shanghai Municipal Science and Technology Major Project (Grant No. 2021SHZD ZX), Shanghai Pujiang Program (Grant No. 22PJ1408600), Medical-Engineering Interdisciplinary Research Foundation of Shanghai Jiao Tong University "Jiao Tong Star" Program (YG2023ZD25), and GuangCi Professorship Program of RuiJin Hospital Shanghai Jiao Tong University School of Medicine.

2. METHODOLOGY

2.1. Overview

We design the Temporal-Spatial prediction (TSP) based on spatial masked autoencoder and temporal contrastive predictive coding. The illustration of TSP pre-training is shown in Figure 1. We formulate the input signal as $X \in \mathbb{R}^{C \times T}$, where C represents the number of channels and T is the number of samples in time series. We first randomly mask and remove some channels of X with variable masking ratios to get $X_v \in \mathbb{R}^{C_v \times T}$, where C_v is the number of visible channels. Then we divided X_v into two part in chronological order: past signal $X_v^p \in \mathbb{R}^{C_v \times T_p}$ and future signal $X_v^f \in \mathbb{R}^{C_v \times T_f}$, where T_p and T_f are the number of samples in past and future time series, respectively. The past visible signal X_v^p is conveyed to an encoder to get the embedding $E_v^p \in \mathbb{R}^{C_v \times D_e}$, where D_e is the embedding dimension of the encoder. E_v^p serves as the input of two different tasks: a decoder of spatial MAE to predict the BIOT embedding of $X_m \in \mathbb{R}^{C_m \times T_p}$, where C_m is the number of masked channels and a temporal CPC to predict the BIOT embedding of future signal X_v^f. Note that only the encoder is utilized in fine-tuning process. By combing the spatial and temporal prediction, we aim to pre-train a strong encoder that is able to deal with variable sampling rates, number of channels or time steps.

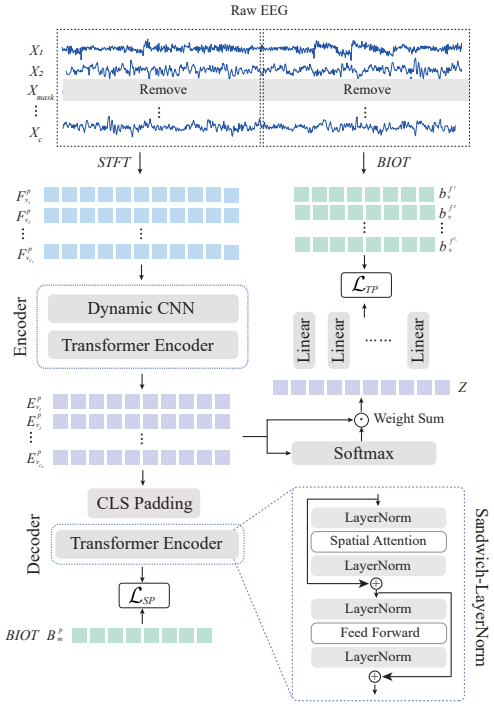

Fig. 1. Illustration of TSP structure.

Data Preprocessing: EEG signals from multiple datasets are first resampled to 200 Hz. Then a 0-75 Hz band-pass filter is applied. As in BIOT, we also normalize each channel by the 95-percentile of its absolute amplitude.

Random Masking: As mentioned above, EEG signals have spatial smearing characteristic, which means adjacent channels contain similar information. Masked channels can be easily recovered from neighboring channels, leading insufficient training of encoder. Image MAE encountered the same problem and solved it by a simple trick: a high mask rate. Gaining experience from it, we set a relatively high lowest mask rate, i.e., 0.5. Meanwhile, to avoid excessive hyperparameter selection and encourage the model to handle more diverse situations, we employ variable masking ratios from 0.5 to 0.9 for different batches as in the work [6]. Specifically, a Gaussian distribution $\mathcal{N}(0.55, 0.25^2)$ with truncated interval $(0.5, 0.9)$ is applied.

BIOT Embedding: BIOT embeds the Short Time Fourier Transform (STFT) values of raw EEG signals in a more abstract and high-level semantics space. Previous works [6] have proved that reconstructing high-level semantics significantly improves the performance. So we utilize the BIOT embedding of X_m^p and X_v^f as the spatial and temporal prediction goals. Specifically, BIOT embeds X_m^p to $B_m^p \in \mathbb{R}^{(C_m \times T_b) \times D_b}$ and X_v^f to $B_v^f \in \mathbb{R}^{(C_v \times T_b) \times D_b}$, where T_b and D_b are the number of windows in STFT and the embedding dimension of BIOT, respectively. The T_b dimension in B_m^p and C_v in B_v^f need to be squeezed as we only focus on the recovery of masked channels (C_m in B_m^p) and prediction of future time steps (T_b in B_v^f). An attention-based weight sum method is employed. Finally, B_m^p can be updated by $\sum_{t=0}^{T_b-1} B_m^p \odot weight$, where $weight = softmax(B_m^p)$ and $weight$ is computed along T_b. B_v^f follows similar steps except that the $weight$ is computed along C_v.

2.2. Encoder

After masking, the past visible signal X_v^p is conveyed to the encoder. We first segment the input with window size W and overlap O to get $S_v^p \in \mathbb{R}^{C_v \times T_s \times W}$, where $T_s = (T_p - W)/O + 1$. For each segment, we embed it by the frequency amplitude computed by STFT to get $F_v^p \in \mathbb{R}^{C_v \times (T_s \times D_f)}$, where D_f is the number of frequency components. F_v^p serves as the input of the dynamic CNN layer.

Dynamic CNN: We adopt a dynamic CNN layer inspired by [7] to enrich the EEG representations. It can extract multiscale time-frequency features by parallelly applying I convolutional layers with different kernel sizes $(1, k_i)$ and the same number of kernels C_1 to F_v^p. As F_v^p is first reshaped to $1 \times C_v \times (T_s \times D_f)$, the 2-D convolutional operations are only employed on the last dimension, i.e., time-frequency dimension. Here, we set $I = 3$ and $k_i = 0.5^i D_f, i = 1, 2, 3$. With a total of T_s time steps, longer kernel length can integrate more frequency components in one convolutional operation for each time step. The I output are then concatenated along the last dimension and layer normalized (LN) to get

$G_1 \in \mathbb{R}^{C_1 \times C_v \times D_1}$.

$$G_1 = LN([G^1, G^2, ..., G^I]), \tag{1}$$

$$G^i = AP(\Phi(Conv2d(F_v^p, (1, k_i), C_1))), \tag{2}$$

where AP represents an average pooling layer with the pooling size of $(1, p)$ and Φ represents $LeakyReLU$. Then, to integrate information from the output of C_1 kernels, we convey G_1 to a pointwise convolution layer and get $G_2 = AP(\Phi(Conv2d(G_1, (1, 1), C_2)))$. Finally, after reshaping and flattening $G_2 \in \mathbb{R}^{C_v \times (C_2 \times D_2)}$, we can get the final output $G_v^p = LN(G_2)W_g$, where $W_g \in \mathbb{R}^{(C_2 \times D_2) \times D_e}$ is a linear projection.

Spatial Transformer: Added with *sin-cos* channel embedding [8] and concatenated with a class token along the channel dimension, $G_v^p \in \mathbb{R}^{(C_v+1) \times D_e}$ is fed into L_e multi-head spatial transformer (ST) encoders, through which the similarity among all the channels is computed to get the attention matrices. We first start with the vanilla transformer encoder [8]. However, A NaN loss is frequently observed when we increase the number of pre-training datasets. Gaining experience from Ding *et al.*[9], we apply a Sandwich-LayerNorm structure as in Figure 1, which is effective to avoid overflow and thus eliminates the NaN loss. Finally, the encoder embedding of X_v^p is formulated as $E_v^p \in \mathbb{R}^{C_v \times D_e}$ and we also get a learnable class token C_0.

2.3. Spatial Prediction

The aim of SP task is to predict the BIOT embedding $B_m^p \in \mathbb{R}^{C_m \times D_b}$ of masked channels from visible channels. Before decoding, we first pad E_v^p to full channels using the class token C_0 learned by the encoder, i.e., C_0 serves as the initial values of masked channels. The *sin-cos* channel embedding is also added to all channels to provide the channel location information. We stacked L_d ST as in the encoder and a linear projection with dimension $D_e \times D_b$ to form the decoder. The mean squared error (MSE) between the predicted BIOT embedding of masked channels and B_m^p is our SP loss function $\mathcal{L}_{SP}$. Note that as in the image MAE, we normalize the B_m^p for each channel before the loss computation.

2.4. Time Prediction

The aim of TP task is to predict the BIOT embedding $B_v^f \in \mathbb{R}^{T_b \times D_b}$ of the visible channels' future time steps using the visible channels' past time steps. We first need to summarize all the past time steps to one high-level representation Z. The same weight sum operation mentioned above is applied. Specifically, $Z = \sum_{t=0}^{C_v-1} E_v^p \odot weight$ where $weight = softmax(E_v^p)$ and $weight$ is computed along C_v. To predict B_v^f, we utilize a log-bilinear model $f(Z, b_v^{f^j}) = exp(b_v^{f^j} W_j Z^T)$ to evaluate the mutual information between $Z \in \mathbb{R}^{1 \times D_e}$ and each future time step $b_v^{f^j} \in B_v^f, j \in [1, T_b]$. $W_j \in \mathbb{R}^{D_b \times D_e}$ is a linear projection that maps Z to the same feature space as $b_v^{f^j}$. The prediction goal is to maximally preserve the mutual information between the predicted representation $W_j Z^T$ and the true corresponding time step $b_v^{f^j}$ while minimize the mutual information with the other time steps. We apply the InfoNCE loss as follow:

$$\mathcal{L}_{TP} = -\frac{1}{T_b} \sum_{j=0}^{T_b-1} \log \frac{f(Z, b_v^{f^j})}{\sum_{k \in T_b} f(Z, b_v^{f^k})}. \tag{3}$$

The final loss of TSP is formulated as $\mathcal{L}_{TSP} = \mu_1 \mathcal{L}_{SP} + \mu_2 \mathcal{L}_{TP}$, where μ_1 and μ_2 are hyperparameters.

3. RESULT ANALYSES

3.1. Datasets and Implementation Details

As shown in Table 1, in pre-training, we collect 6 datasets with different number of channels and paradigms to enrich the data type. The SEED-series dataset[1] contains SEED-V [10], SEED-GER [11], SEED-FRA [11] and some unreleased datasets. The EmotionHelper (EH) is a private dataset that collects EEG data from depression and healthy subjects. For HBN dataset [12],we select the 19 named-channels according to the provided location file. For PRED+CT [13], we select 3 datasets named Depression RL, Depression Rest and OCD Flankers.

Table 1. Dataset Information

Name	Channels	Rate	Files	Duration
SEED series[1]	62	200 Hz	295	$\approx$ 60min
EP	18	300 Hz	983	[30min, 60min]
TUAB [14]	19	250 Hz	2717	>15 min
HBN [12]	19	500 Hz	8277	[2min, 5min]
TDBRAIN [15]	26	500 Hz	2690	2min
PRED+CT [13]	62	500 Hz	271	[8min, 20min]

As for the downstream tasks, the selected datasets are different from those in pre-training. we select two datasets for emotion recognition (SEED, SEED-IV) and one for event detection (TUEV [16]). Different from previous studies on SEED and SEED-IV, we introduce the validation set. For SEED (3-class), the total 15 clips of all subjects are divided into 9, 3, 3 clips for training, validation and test sets. For SEED-IV (4-class), the total 24 clips are divided into 4, 4, 16 for test, validation and training sets. TUEV is a 6-class task and provides the test set. We randomly split the training patients into training and validation sets by 80% and 20%.

Simialrly as BIOT, the results are averaged on three seeds $[42, 0, 10]$. The hyperparameter selection is based on the validation set under seed 42. We set the time steps $T = 2000$ and $T_p = T_f = 1000$. The parameters of BIOT embedding follow the default settings. As for the encoder, we set $W = 200$ and

[1]https://bcmi.sjtu.edu.cn/home/seed/index.html

Table 2. Performance (%) of different algorithms

	SEED		SEED-IV		TUEV	
	Balanced Acc.	Coken's Kappa	Balanced Acc.	Coken's Kappa	Balanced Acc.	Coken's Kappa
ST-Trans [17]	54.15/0.54	28.60/0.36	36.69/1.34	15.33/1.68	39.84/2.28	37.65/3.06
SPaRCNet [18]	57.23/0.20	32.89/0.20	40.25/0.71	20.33/0.42	41.61/2.62	42.33/1.81
CNN-Trans [19]	59.02/0.60	35.33/1.00	43.33/0.53	23.93/0.81	40.87/1.61	38.15/1.34
ContraWR [20]	59.24/0.59	36.18/1.16	45.61/0.36	26.81/0.08	43.84/3.49	39.12/2.37
FFCL [21]	59.88/0.56	36.90/1.01	43.44/0.31	23.90/1.15	39.79/1.04	37.32/1.88
BIOT [4]	64.34/0.50	44.20/0.62	44.25/0.01	25.14/0.81	52.81/2.25	**52.73/2.49**
Scratch	58.61/0.69	34.55/1.22	43.47/0.88	23.15/0.99	42.90/1.69	37.15/3.46
Spatial Prediction (SP)	59.37/0.24	35.21/0.43	44.51/1.11	23.82/1.66	47.87/1.86	41.54/2.58
Temporal Prediction (TP)	63.17/0.63	41.96/0.76	45.97/0.58	26.57/0.91	51.76/1.56	43.22/2.35
TSP ($\mu_1 = 0.5, \mu_2 = 0.5$)	**65.33/0.49**	**44.98/0.88**	**46.40/0.47**	**27.37/0.82**	**53.37/1.10**	52.61/2.44

Table 3. Performance of different μ_1, μ_2 (%)

	SEED	
	Balanced Acc.	Coken's Kappa
TSP ($\mu_1 = 0.8, \mu_2 = 0.2$)	64.38/1.02	43.35/2.09
TSP ($\mu_1 = 0.2, \mu_2 = 0.8$)	64.09/1.15	43.01/2.14

$O = 100$. The CNN embedding dimension $D_e = 512$ and kernel length p of pooling layer is 8. The number of kernels $C_1 = 8, C_2 = 4$. We stacked 12 ST blocks in encoder and 6 ST blocks in decoder. For pre-training, we use a half-cycle cosine learning rate schedule with 40 warm-up epochs and base learning rate of $1e^{-5}$. We set $\mu_1 = 0.5, \mu_2 = 0.5$. The batch size is set to 2048 and epoch is 70. For the fine-tuning, learning rate is selected from $1e-4$ to $5e-5$ and weight decay is in $[0.05, 0.0005, 0.1]$. The Coken's Kappa is set as the monitor index.

3.2. Results Analysis and Comparison

Table 2 lists the the Balanced Accuracy and Coken's Kappa and corresponding standard deviation. As TUEV has imbalanced number of samples, we provide Balanced Accuracy instead of the conventional accuracy. All the baseline methods are CNN-based or Transformer-based algorithms designed for raw-EEG classification tasks. For BIOT, we load the provided pre-trained model for SEED and SEED-IV. Other methods are training from scratch. All the approaches follow the same hyperparameter settings.

Comparison with Baseline: As shown in Table 2, the pre-trained BIOT obviously performs better than other baselines on SEED and TUEV datasets. Even the minimum improvements reach 4% on SEED and 9% on TUEV, which proves the superiority of pre-training. However, ContraWR surpasses BIOT on SEED-IV around 1% for both indexes. The proposed TSP surpasses all the baselines for both indexes except that the Coken's Kappa on TUEV has a slightly decrease of 0.12% compared with BIOT. It should be noted that the pre-training of BIOT utilized the training set of TUEV, which may explain the decrease. The results show the superiority of TSP and thus suggest that using high-level representations as the prediction objectives is beneficial and prove the effectiveness of using diverse datasets for pre-training.

Ablation Study: The second part of Table 2 presents the ablation study results. We first training the TSP model from scratch to demonstrate the effectiveness of pre-training. The pre-trianed TSP significantly surpasses scratch with a maximum improvement of 10% on TUEV. We then provide the results of single-domain approaches, Temporal Prediction (TP) model and Spatial Prediction (SP) model. The structures of TP and SP are totally the same as those in TSP. Compared with single-domain approaches, TSP achieves the highest results for all the datasets. This verifies that separately processing the temporal and spatial domains indeed improve the model performance. Although both TP and SP perform better than scratch, SP only improve 0.76% on SEED. The improvement is much smaller than TP especially on SEED and TUEV. So we test if increasing the proportion of $\mathcal{L}_{TP}$, i.e. μ_2 can improve the performance. Table 3 presents the results on SEED. Unexpectedly, Larger μ_1 is slightly better than larger μ_2, and balanced μ_1, μ_2 is the best.

4. CONCLUSION

We propose the Temporal-Spatial prediction (TSP) self-supervised method based on spatial masked autoencoder and temporal contrastive predictive coding, and using six kinds of datasets for pre-training. TSP can adapt to different number of channels and different EEG classification tasks. Results on three public datasets show that TSP surpasses the SOTA self-supervised method BIOT and the single-domain methods TP and SP. We demonstrate the effectiveness of separately processing the temporal and spatial domains and the feasibility of pre-training on diverse datasets.

5. REFERENCES

[1] Aäron van den Oord, Yazhe Li, and Oriol Vinyals, "Representation learning with contrastive predictive coding," *ArXiv*, vol. abs/1807.03748, 2018.

[2] Rui Li, Yiting Wang, Wei-Long Zheng, and Bao-Liang Lu, "A multi-view spectral-spatial-temporal masked autoencoder for decoding emotions with self-supervised learning," *Proceedings of the 30th ACM International Conference on Multimedia*, 2022.

[3] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Doll'ar, and Ross B. Girshick, "Masked autoencoders are scalable vision learners," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15979–15988, 2021.

[4] Chaoqi Yang, Michael Brandon Westover, and Jimeng Sun, "BIOT: Cross-data biosignal learning in the wild," *ArXiv*, vol. abs/2305.10351, 2023.

[5] Alice F. Jackson and Donald J. Bolger, "The neurophysiological bases of EEG and EEG measurement: A review for the rest of us," *Psychophysiology*, vol. 51, no. 11, pp. 1061–1071, 2014.

[6] Tianhong Li, Huiwen Chang, Shlok Kumar Mishra, Han Zhang, Dina Katabi, and Dilip Krishnan, "Mage: Masked generative encoder to unify representation learning and image synthesis," *ArXiv*, vol. abs/2211.09117, 2022.

[7] Yi Ding, Neethu Robinson, Qiuhao Zeng, and Cuntai Guan, "Tsception: Capturing temporal dynamics and spatial asymmetry from EEG for emotion recognition," *ArXiv*, vol. abs/2104.02935, 2021.

[8] Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *NIPS*, 2017.

[9] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, and Jie Tang, "CogView: Mastering text-to-image generation via transformers," in *Neural Information Processing Systems*, 2021.

[10] Wei Liu, Jielin Qiu, Wei-Long Zheng, and Bao-Liang Lu, "Comparing recognition performance and robustness of multimodal deep learning models for multimodal emotion recognition," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 14, pp. 715–729, 2021.

[11] W. Liu, Wei-Long Zheng, Ziyi Li, Si-Yuan Wu, Lu Gan, and Bao-Liang Lu, "Identifying similarities and differences in emotion recognition with EEG and eye movements among chinese, german, and french people," *Journal of Neural Engineering*, vol. 19, 2022.

[12] Lindsay M. Alexander and Jasmine Escalera et al., "An open resource for transdiagnostic research in pediatric mental health and learning disorders," *Scientific Data*, vol. 4, 2017.

[13] James F. Cavanagh, Arthur Napolitano, Christopher Wu, and Abdullah Al Mueen, "The patient repository for EEG data + computational tools (pred+ct)," *Frontiers in Neuroinformatics*, vol. 11, 2017.

[14] Iyad Obeid and Joseph W. Picone, "The temple university hospital EEG data corpus," *Frontiers in Neuroscience*, vol. 10, 2016.

[15] Hanneke van Dijk, Guido A. van Wingen, Damiaan A.J.P. Denys, Sebastian Olbrich, Rosalinde van Ruth, and Martijn Arns, "The two decades brainclinics research archive for insights in neurophysiology (td-brain) database," *Scientific Data*, vol. 9, 2022.

[16] Amir Hossein Harati Nejad Torbati, Meysam Golmohammadi, Silvia Lopez de Diego, Iyad Obeid, and Joseph W. Picone, "Improved EEG event classification using differential energy," *2015 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, pp. 1–4, 2015.

[17] Yonghao Song, Xueyu Jia, Lie Yang, and Longhan Xie, "Transformer-based spatial-temporal feature learning for EEG decoding," *ArXiv*, vol. abs/2106.11170, 2021.

[18] Jin Jing and Wendong Ge et al., "Development of expert-level classification of seizures and rhythmic and periodic patterns during EEG interpretation," *Neurology*, vol. 100, pp. e1750 – e1762, 2023.

[19] Wei Yan Peh, Yu Yao, and Justin Dauwels, "Transformer convolutional neural networks for automated artifact detection in scalp EEG," *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pp. 3599–3602, 2022.

[20] Chaoqi Yang, Danica Xiao, Michael Brandon Westover, and Jimeng Sun, "Self-supervised EEG representation learning for automatic sleep staging," 2021.

[21] Hongli Li, Man Ding, Ronghua Zhang, and Chunbo Xiu, "Motor imagery EEG classification algorithm based on cnn-lstm feature fusion network," *Biomedical Signal Processing and Control*, 2022.