MBrain: A Multi-channel Self-Supervised Learning Framework for Brain Signals

Donghong Cai*
Zhejiang University
donghongcai@zju.edu.cn

Junru Chen*
Zhejiang University
jrchen_cali@zju.edu.cn

Yang Yang†
Zhejiang University
yangya@zju.edu.cn

Teng Liu
Zhejiang University
liuteng_27@zju.edu.cn

Yafeng Li
Nuozhu Technology Co., Ltd.
yafeng.li@neurox.cn

ABSTRACT

Brain signals are important quantitative data for understanding physiological activities and diseases of human brain. Meanwhile, rapidly developing deep learning methods offer a wide range of opportunities for better modeling brain signals, which has attracted considerable research efforts recently. Most existing studies pay attention to supervised learning methods, which, however, require high-cost clinical labels. In addition, the huge difference in the clinical patterns of brain signals measured by invasive (*e.g.*, SEEG) and non-invasive (*e.g.*, EEG) methods leads to the lack of a unified method. To handle the above issues, in this paper, we propose to study the self-supervised learning (SSL) framework for brain signals that can be applied to pre-train either SEEG or EEG data. Intuitively, brain signals, generated by the firing of neurons, are transmitted among different connecting structures in human brain. Inspired by this, we propose *MBrain* to learn implicit spatial and temporal correlations between different channels (*i.e.*, contacts of the electrode, corresponding to different brain areas) as the cornerstone for uniformly modeling different types of brain signals. Specifically, we represent the spatial correlation by a *graph* structure, which is built with proposed multi-channel CPC. We theoretically prove that optimizing the goal of multi-channel CPC can lead to a better predictive representation and apply the *instantaneou-time-shift prediction* task based on it. Then we capture the temporal correlation by designing the *delayed-time-shift prediction* task. Finally, *replace-discriminative-learning* task is proposed to preserve the characteristics of each channel. Extensive experiments of seizure detection on both EEG and SEEG large-scale real-world datasets demonstrate that our model outperforms several state-of-the-art time series SSL and unsupervised models, and has the ability to be deployed to clinical practice.

*Both authors contributed equally to this research.
†Corresponding author.

KDD '23, August 6–10, 2023, Long Beach, CA, USA

ACM ISBN 979-8-4007-0103-0/23/08…$15.00
https://doi.org/10.1145/3580305.3599426

CCS CONCEPTS

• **Applied computing** → **Health care information systems**.

KEYWORDS

brain signals, self-supervised learning, multi-channel time series, seizure detection

ACM Reference Format:
Donghong Cai, Junru Chen, Yang Yang, Teng Liu, and Yafeng Li. 2023. MBrain: A Multi-channel Self-Supervised Learning Framework for Brain Signals. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '23), August 6–10, 2023, Long Beach, CA, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3580305.3599426

1 INTRODUCTION

Brain signals are foundational quantitative data for the study of human brain in the field of neuroscience. The patterns of brain signals can greatly help us to understand the normal physiological function of the brain and the mechanism of related diseases. There are many applications of brain signals, such as cognitive research [20, 22], emotion recognition [6, 38], neurological disorders [1, 42] and so on. Brain signals can be measured by noninvasive or invasive methods [29]. The noninvasive methods, like *electroencephalography* (EEG), cannot simultaneously consider temporal and spatial resolution along with the deep brain information, but they are easier to implement without any surgery. As for invasive methods like *stereoelectroencephalography* (SEEG), they require extra surgeries to insert the recording devices, but have access to more precise and higher signal-to-noise data. For both EEG and SEEG data, there are multiple *electrodes* with several contacts (also called *channels*) that are sampled at a fixed frequency to record brain signals.

Recently, discoveries in the field of neuroscience have inspired advances of deep learning techniques, which in turn promotes neuroscience research. According to the literature, most deep learning-based studies of brain signals focus on supervised learning [10, 33, 37, 46], which relies on a large number of clinical labels. However, obtaining accurate and reliable clinical labels requires a high cost. In the meantime, the emergence of self-supervised learning (SSL) and its great success [4, 9, 13, 28] makes it a predominant learning paradigm in the absence of labels. Therefore, some recent studies have introduced the means of SSL to extract the representations of brain signal data. For example, Banville et al. [3] directly applies general SSL tasks to pre-train EEG data, including relative position prediction [14], temporal shuffling [25] and contrastive

predictive coding [28]. Mohsenvand et al. [26] designs data augmentation methods, and extends the self-supervised model SimCLR [7] in computer vision to EEG data. In contrast to numerous works investigating EEG, few studies focus on SEEG data. Martini et al. [24] proposes an SSL model for real-time epilepsy monitoring in multimodal scenarios with SEEG data and video recordings.

Despite the advances on representation learning of brain signals, two main issues remain to be overcome. **Firstly, almost all existing methods are designed for a particular type of brain signal data, and there is a lack of a unified method for handling both EEG and SEEG data.** The challenge mainly lies in the different clinical patterns of brain signals that need to be measured in different ways. On the one hand, EEG collects noisy and rough brain signals on the scalp; differently, SEEG collects deeper signals with more stereo spatial information, which indicates more significant differences of different brain areas [31]. On the other hand, in contrast to EEG with a gold-standard collection location, the monitoring areas of SEEG vary greatly between subjects, leading to different number and position of channels. Therefore, how to find the commonalities of EEG and SEEG data to design a unified framework is challenging.

Another issue is the gap between existing methods and the real-world applications. In clinical scenarios, doctors typically locate brain lesions by analyzing signal patterns of *each* channel and their *holistic* correlations. A straight-forward way for this goal is to model each of the channels separately by single-channel time series models, which, however, cannot exploit correlations between brain areas [11, 23]. As for the existing multivariable time series models, most of them can only capture implicit correlation patterns [8, 45], whereas explicit correlations are required by doctors for identifying lesions. Moreover, although some graph-based methods have been proposed to explicitly learn correlations, they focus on giving an overall prediction for all channels at a time but overlook the prediction on one specific channel [36, 47]. Therefore, how to explicitly capture the spatial and temporal correlations while giving channel-wise prediction is another issue to be overcome.

To address the challenges above, we propose a multi-channel self-supervised learning framework *MBrain*, which can be generally applied for learning representations of both EEG and SEEG data. Specifically, based on domain knowledge and data observations, we propose to learn the correlation graph between channels as the common cornerstone for both two types of brain signals. In particular, we employ Contrastive Predictive Coding (CPC) [28] as the backbone model of our framework by extending it to handle multi-channel data. We theoretically prove that the optimization objective of the proposed multi-channel CPC is to maximize the mutual information of each channel and its correlated ones, so as to obtain better predictive representations. Based on the multi-channel CPC, we propose the instantaneous time shift task to explicitly learn the spatial correlations between channels, and the delayed time shift task and the replace discriminative task are designed to capture the temporal correlation patterns and to preserve the characteristics of each channel respectively. To validate the effectiveness of our model, we pay special attention to its application in seizure detection. Extensive experiments show that *MBrain* outperforms several state-of-the-art baselines on large-scale real-world EEG and SEEG datasets for the seizure detection task. Overall, the main contributions of this work can be summarized as follows:

- We are the first work to design a generalized self-supervised learning framework, which can be applied to pre-train both EEG and SEEG signals.
- We propose *MBrain* to explicitly capture the spatial and temporal correlations of brain signals to learn a unique representation for each channel.
- We validate the effectiveness and clinical value of the proposed framework through extensive experiments on large-scale real-world EEG and SEEG datasets.

2 PRELIMINARY: THEORETICAL ANALYSIS OF MULTI-CHANNEL CPC

We employ Contrastive Predictive Coding (CPC) [28] as the basis of our framework. The pretext task of CPC is to predict low-level local representations by high-level global contextual representations c_t at the t-th time step. Theoretically, the optimal InfoNCE loss proposed by CPC with $N-1$ negative samples $\mathcal{L}_N^{\text{opt}}$ is a lower bound of the mutual information between contextual semantic distribution $p(c_t)$ and raw data distribution $p(x_{t+k})$, *i.e.*, $\mathcal{L}_N^{\text{opt}} \geq -I(x_{t+k}; c_t) + \log N$, where k is the prediction step size. CPC is originally designed for single-channel sequence data only, and there are two natural ways to extend single channel CPC to multi-channel version. The first one is to use CNNs with multiple kernels to encode all channels simultaneously, which cannot offer explicit correlation patterns for doctors to identify lesions. The second one is to train a shared CPC regarding all channels as one, which has no ability to capture the correlation patterns. Taking a comprehensive consideration, we propose *multi-channel CPC* in this paper. Our motivation is to explicitly aggregate the semantic information of multiple channels to predict the local representations of one channel. Formally, we propose the following proposition as our basic starting point.

Proposition 1. *Introducing the contextual information of the correlated channels increases the amount of mutual information with the raw data of the target channel.*

$$I(x^i_{t+k}; \Phi(c_t)) = I(x^i_{t+k}; c^i_t, \Phi(\{c^j_t\}_{j\neq i})) \geq I(x^i_{t+k}; c^i_t), \tag{1}$$

where i and j are indexes of the channels. $\Phi(\cdot)$ represents some kinds of aggregate function, which has no additional formal constraints other than the need to retain information of the target channel.

Proof. We use the linear operation of mutual information to obtain: $I(x^i_{t+k}; c^i_t, \Phi(\{c^j_t\}_{j\neq i})) = I(x^i_{t+k}; c^i_t) + I(x^i_{t+k}; \Phi(\{c^j_t\}_{j\neq i})|c^i_t)$. According to the non-negativity of the conditional mutual information, we complete the proof. □

It seems natural that the predictive ability of multiple channels is stronger than that of a single channel, which is also consistent with the assumption of Granger causality [18] to some extent. Therefore, we choose to approximate the more informative $I(x^i_{t+k}; \Phi(c_t))$ to obtain more expressive representations. Specifically, followed by InfoNCE, we define our loss function $\mathcal{L}_N$ as

$$\mathcal{L}_N = -\sum_i \mathbb{E}_{X^i}\left[\log \frac{f_k(x_{t+k}, \Phi(c_t))}{\sum_{x_j \in X} f_k(x_j, \Phi(c_t))}\right], \tag{2}$$

where X^i denotes the data sample set consisting of one positive sample and $N-1$ negative samples of the i-th channel. We then establish the relationship between $\mathcal{L}_N$ and $I(x^i_{t+k};\Phi(c_t))$.

Theorem 1. *Given a sample set for each channel $X^i = \{x^i_1,\dots,x^i_N\}$, $i = 1,\dots,n$ consisting of one positive sample from $p(x^i_{t+k}|\Phi(c_t))$ and $N-1$ negative samples from $\sum_j p(x^j_{t+k})/n$, where n is the number of channels. The optimal $\mathcal{L}^{opt}_N$ is the lower bound of $\sum_i I(x^i_{t+k};\Phi(c_t))$:*

$$\mathcal{L}^{opt}_N \geq \sum_i \left[-I(x^i_{t+k};\Phi(c_t)) + \log N\right]. \tag{3}$$

Proof. The optimal $f_k(x_{t+k},\Phi(c_t))$ is proportional to the division of two distributions $p(x^i_{t+k}|\Phi(c_t))/(\sum_j p(x^j_{t+k})/n)$, which is the same as single-channel CPC. And we can directly replace the data distributions in the proof of single-channel CPC (see details in Appendix B) to obtain the inequality below:

$$\mathcal{L}^{\text{opt}}_N \geq \sum_i \left[\mathbb{E}_{X^i}\log\left[\frac{\frac{1}{n}\sum_j p(x^j_{t+k})}{p(x^i_{t+k}|\Phi(c_t))}\right] + \log N\right] \tag{4}$$

$$= \mathbb{E}_{X^1,X^2,\dots,X^n}\log\left[\frac{[\frac{1}{n}\sum_j p(x^j_{t+k})]^n}{\Pi_j p(x^j_{t+k}|\Phi(c_t))}\right] + n\log N. \tag{5}$$

According to the Jensen Inequality and concavity of the logarithmic function, we obtain that $(\sum_j \log p(x^j_{t+k}))/n \leq \log(\sum_j p(x^j_{t+k})/n)$. By exponentiating the two equations, we have

$$\Pi_j p(x^j_{t+k}) \leq [\frac{1}{n}\sum_j p(x^j_{t+k})]^n. \tag{6}$$

With the help of equation 6, we can further obtain the lower bound of equation 5:

$$\mathcal{L}^{\text{opt}}_N \geq \mathbb{E}_{X^1,X^2,\dots,X^n}\log\left[\frac{\Pi_j p(x^j_{t+k})}{\Pi_j p(x^j_{t+k}|\Phi(c_t))}\right] + n\log N \tag{7}$$

$$= \sum_i \left[-I(x^i_{t+k};\Phi(c_t)) + \log N\right]. \tag{8}$$

Then we complete the proof. □

We next analyze the advantages of multi-channel CPC over single-channel CPC. Our loss function $\mathcal{L}_N$ leads to a better predictive representation because we approximate a more informative objective $I(x^i_{t+k};\Phi(c_t))$, if the optimal loss function for each channel has $\log N$ gap with $I(x^i_{t+k};\Phi(c_t))$, which is the same in single-channel CPC. Moreover, with the same GPU memory, the more channels, the smaller the batch size that can be accommodated. But we can randomly sample negative samples across all channels, which increases the diversity of negative samples. However, in order to narrow the approximation gap, equation 6 should be considered. The equality sign in this inequality holds if and only if samples from each channel follows the same distribution. In fact, for many large-scale time series data (*e.g.*, brain signal data used in this work), by normalizing each channel, they all exhibit close normal distributions, leading to small gaps in equation 6.

3 PROPOSED METHOD

In this section, we introduce the details of the novel self-supervised learning framework *MBrain*. For the commonality between EEG and SEEG, we are inspired by the synergistic effect of brain function and nerve cells, that is, different connectivity patterns correspond to different brain states [23]. In particular, for brain signals, nerve cells will spontaneously generate traveling waves and spread them out [11], maintaining some characteristics such as shape during the process. Therefore, the degree of channel similarity implies different propagation patterns of traveling waves, reflecting the differences in connectivity patterns to some extent. Both EEG and SEEG data follow the inherent physiological mechanism. Therefore, we propose to extract the correlation graph structure between channels (brain areas) as the cornerstone to unify EEG and SEEG (Section 3.1). Next, we introduce three SSL tasks to model brain signals in Section 3.2. We propose *instantaneous time shift task* based on multi-channel CPC and *delayed time shift task* to capture the spatial and temporal correlation patterns. Then *Replace discriminative task* is designed to preserve characteristics of each channel.

Notations. For both EEG and SEEG data, there are multiple electrodes with **C** channels. We use $X = \{x_l \in \mathbb{R}^{\mathbf{C}}, l = 1,\dots,\mathbf{L}\}$ to represent raw time series data with **L** time points. i and j denote the index of channels. $Y_{l,i} \in \{0,1\}$ is the label for the l-th time point of the i-th channel. We use a **W**-length window with no overlap to obtain the time segments $S = \{s_t, t = 1,\dots,|S|\}$ (see details in Appendix A). The label corresponding to the t-th time segment of the i-th channel is denoted as $Y^s_{t,i}$.

3.1 Learning Correlations between Channels

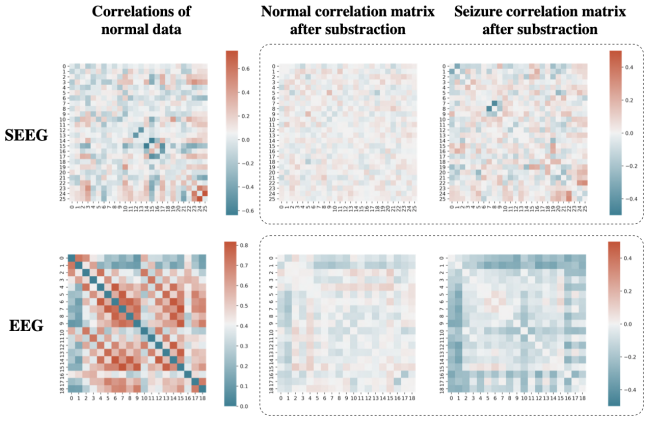


Figure 1: The normal and seizure correlation matrices of EEG and SEEG brain signals. The top row is for SEEG and the bottom row is for EEG. For clear presentation, we sample some channels in SEEG data. The leftmost two figures are the base correlation matrices on normal data. The two figures in the middle column represent the matrices after subtracting another normal correlation matrices from the base matrices, and the rightmost column includes matrices after subtracting seizure correlation matrices from the base matrices.

As mentioned above, the correlation patterns between different brain areas can help us to distinguish brain activities in downstream

tasks to a large extent. Taking the seizure detection task as an example, when seizures occur, more rapid and significant propagation of spike-and-wave discharges will appear [32], which greatly enhances the correlation between channels. This phenomenon is also verified by data observations in Figure 1. As Figure 1 shows, for both EEG and SEEG data, we can observe that the correlation matrices are nearly identical on two normal segments without overlap in the same subject. In contrast, the correlation matrix in the epileptic states differs greatly from the normal ones. These data observations verify the conclusion that correlation patterns can help us to distinguish different brain states, and support us to treat correlation graph structure learning as the common cornerstone of our framework. However, correlations between brain regions are difficult to be observed and recorded directly. Therefore, for each time step t, our goal is to learn the structure of the correlation graph, whose adjacency matrix is $\mathbf{A}_t$, where nodes in the graph indicate channels and weighted edges denote the correlations between channels.

Considering that the brain is in normal and stable state most of the time, we first define the *coarse-grained* correlation graph as the prior graph for a particular individual as

$$\mathbf{A}^{\text{coarse}}(i,j) = \mathbb{E}_{s_t}[\text{Cosine}(s_{t,i}, s_{t,j})], \tag{9}$$

where the expectation operation averages over all the correlation matrices computed in only one time segment s_t, and $\text{Cosine}(\cdot,\cdot)$ denotes the cosine similarity function.

Next, based on $\mathbf{A}^{\text{coarse}}$, for each pair of channels, we further model their *fine-grained* short-term correlation within each time segment. We assume that the fine-grained correlations follow a Gaussian distribution element-wise, whose location parameters are elements of $\mathbf{A}^{\text{coarse}}$ and scale parameters will be learned from the data. By means of the reparameterization trick, the short-term correlation matrix of the t-th time segment is sampled from the learned Gaussian distribution:

$$\sigma_t(i,j) = \text{SoftPlus}(\text{MLP}(c^{\text{self}}_{t,\tau,i}, c^{\text{self}}_{t,\tau,j})), \tag{10}$$

$$n_t(i,j) \sim \mathcal{N}(0,1), \tag{11}$$

$$\mathbf{A}^{\text{fine}}_t(i,j) = \mathbf{A}^{\text{coarse}}(i,j) + \sigma_t(i,j) \times n_t(i,j). \tag{12}$$

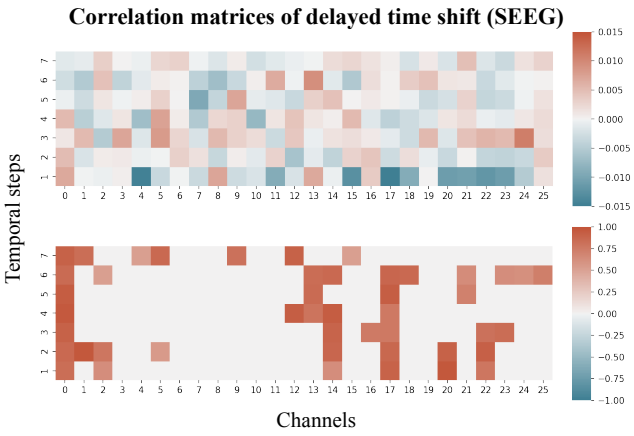
$\text{SoftPlus}(\cdot)$ is a commonly used activation function to ensure the learned standard deviation is positive. $c^{\text{self}}_{t,\tau}$ is the contextual representation of raw time segments extracted by encoders (see details in Section 3.2). To remove the spurious correlations caused by low frequency signals and enhance the sparsity, which is a common assumption in neuroscience [41], we filter the edges by a threshold-based function to obtain the final correlation graph structure $\mathbf{A}_t$:

$$\mathbf{A}_t(i,j) = \begin{cases} \mathbf{A}^{\text{fine}}_t(i,j), & \mathbf{A}^{\text{fine}}_t(i,j) \geq \theta_1, \\ 0, & \mathbf{A}^{\text{fine}}_t(i,j) < \theta_1. \end{cases} \tag{13}$$

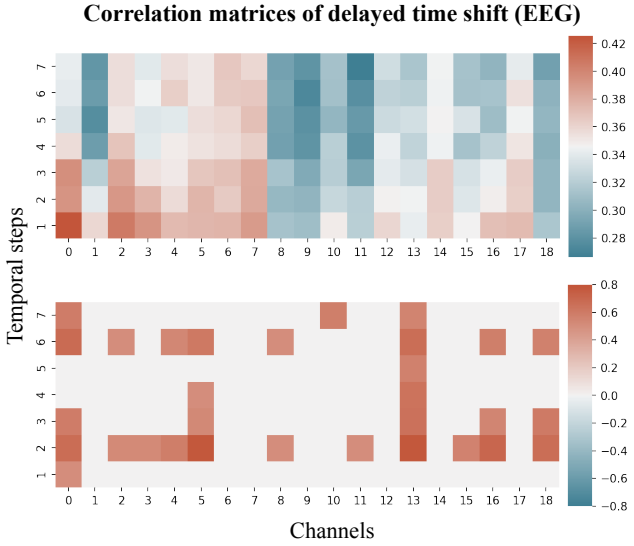
3.2 Self-supervised Learning for Brain Signals

To capture the correlation patterns in space and time, we propose two self-supervised tasks: *instantaneous time shift* that is based on multi-channel CPC and captures the short-term correlations focusing on spatial patterns; and *delayed time shift* for temporal patterns in broader time scales. *Replace discriminative learning* is designed to preserve the unique characteristics of each channel so as to achieve accurate channel-wise prediction.

Instantaneous Time Shift. For spatial patterns, we aim to leverage the contextual information of correlated channels to better predict future data of the target channel. Therefore, we apply multi-channel CPC and utilize the fine-grained graph structure $\mathbf{A}_t$ obtained in Section 3.1 as the correlations between channels.



(a) The correlation matrices of delayed time shift of SEEG data.



(b) The correlation matrices of delayed time shift of EEG data.

Figure 2: The correlation matrices of delayed time shift of SEEG and EEG. For each subfigure, the top figure shows the average correlation matrix over all clips. And the bottom figure represents the correlation matrix of one particular sampled clip. We compute cosine similarity between the first time segment of the first channel and the time segments of other channels in the next consecutive 7 time steps. For clear presentation, we sample 26 channels for SEEG data and set correlations below 0.5 to 0 for the bottom figure.

We first use a non-linear encoder g_{enc} (1D-CNN with d kernels) mapping the observed time segments to the local latent d-dimensional representations $z_t = g_{\text{enc}}(s_t) \in \mathbb{R}^{\mathcal{T}\times \text{C}\times d}$ for each channel separately. $\mathcal{T}$ is the sequential length after down sampling

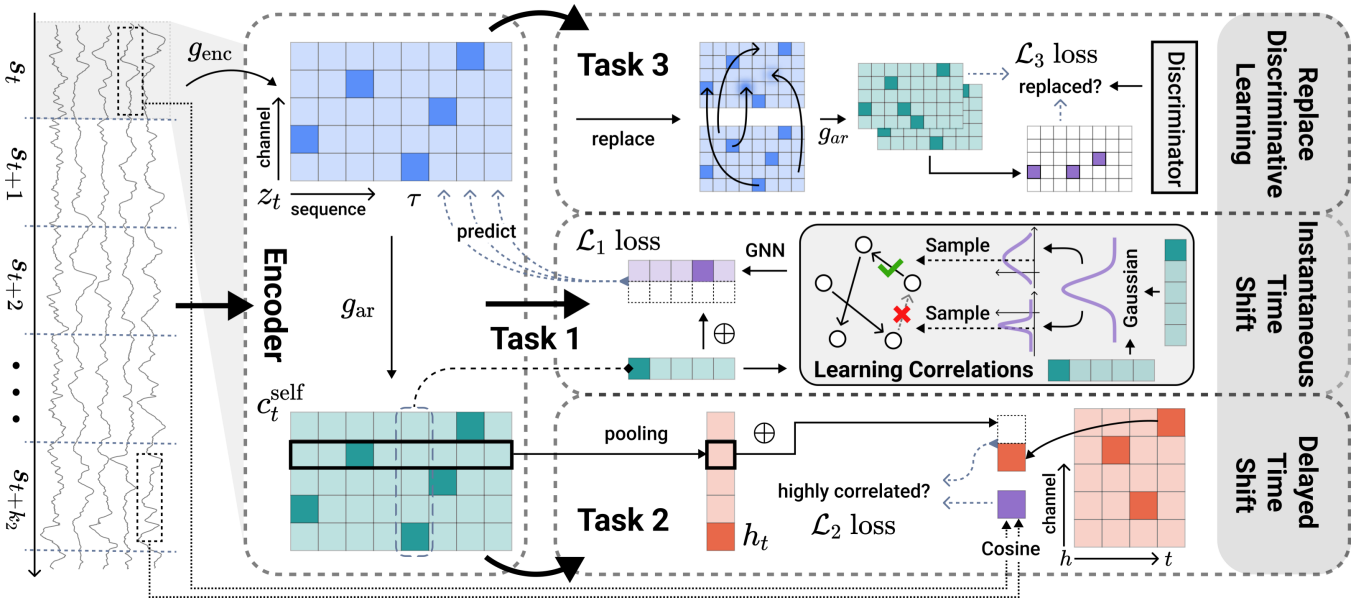


Figure 3: Overview of *MBrain*. The leftmost is the raw multi-channel brain signals. We use an encoder to map the raw data into a low-dimensional representation space. To capture the spatial and temporal correlation patterns, we propose three SSL tasks to guide the encoder to learn informative and distinguishable representations.

by g_{enc}. Then an autoregressive model g_{ar} is utilized to summarize the historical τ-length local information of each channel itself to obtain the respective contextual representations:

$$c_{t,\tau}^{\text{self}} = g_{\text{ar}}(z_{t,1}, \cdots, z_{t,\tau}). \tag{14}$$

In this step, we only extract the contextual information of all channels independently. Based on the graph structure $\mathbf{A}_t$, we instantiate the aggregate function $\Phi(\cdot)$ in equation 5 as GNNs due to their natural message-passing ability on a graph. Here we use a one-layer directed GCN [44] to show the process:

$$c_{t,\tau,i}^{\text{other}} = \text{ReLU}\left(\frac{\sum_{j\neq i} \mathbf{A}_t(i,j) \cdot c_{t,\tau,j}^{\text{self}}}{\sum_{j\neq i} \mathbf{A}_t(i,j)} \cdot \Theta\right), \tag{15}$$

where Θ is the learnable matrix. Considering that we only aggregate other channels' information, the self-loop in GCN is removed here. Finally, by combining both $c_{t,\tau}^{\text{self}}$ and $c_{t,\tau}^{\text{other}}$ to obtain the global representations $c_{t,\tau}$, the model can predict the local representations k_1-step away $z_{t,\tau+k_1}$ based on the multi-channel CPC loss:

$$c_{t,\tau} = \text{Concat}(c_{t,\tau}^{\text{self}}, c_{t,\tau}^{\text{other}}), \tag{16}$$

$$\mathcal{L}_1 = \mathcal{L}_N = -\mathbb{E}_{t,i,k_1}\left[\log \frac{c_{t,\tau,i}^{\top} W_{k_1} z_{t,\tau+k_1,i}}{\sum_{z_j \in X_t^i} c_{t,\tau,i}^{\top} W_{k_1} z_j}\right], \tag{17}$$

where X_t^i denotes the random noise set including one positive sample $z_{t,\tau+k_1,i}$ and $N-1$ negative samples. W_{k_1} is the learnable bilinear score matrix of the k_1-th step prediction.

Delayed Time Shift. For brain areas far apart, there exists delayed brain signal propagation, which is confirmed by the data observations showed in Figure 2. Figure 2 confirms that there still exist significant correlations between time segments across several time steps. Unlike instantaneous time shift, delayed correlations are not stable. This can be concluded from the numerical difference between the averaged correlation matrix and the sampled correlation matrix in both figures. Therefore, we design a more flexible self-supervised task to learn the delayed correlations.

Our motivation is that if a simple classifier can easily predict whether two time segments are highly correlated, the segment representations will be significantly different from those with weaker correlations. We thus define the delayed time shift task to encourage more distinguishable segment representations. Similar with instantaneous time shift, we first compute the cosine similarity matrix based on raw data between time segments across several time steps. For the t-th time segment of the i-th channel, the long-term correlation matrix $\mathbf{B}_t^i$ is computed as

$$\mathbf{B}_t^i(k_2, j) = \text{Cosine}(s_{t,i}, s_{t+k_2,j}), \tag{18}$$

where j traverses all channels including the i-th target channel and k_2 traverses at most K_2 prediction steps. Then we construct pseudo labels Y_t^i according to $\mathbf{B}_t^i$ to encourage the segment representations with higher correlations to be closer. A predefined threshold θ_2 is set to assign pseudo labels:

$$Y_t^i(k_2, j) = \begin{cases} 1, & \mathbf{B}_t^i(k_2, j) \geq \theta_2, \\ 0, & \mathbf{B}_t^i(k_2, j) < \theta_2. \end{cases} \tag{19}$$

With the pseudo labels, we define the cross entropy loss of the delayed time shift prediction task:

$$h_t = \text{Pooling}(c_{t,1}^{\text{self}}, \cdots, c_{t,\mathcal{T}}^{\text{self}}), \tag{20}$$

$$\hat{p} = \text{Softmax}(\text{MLP}(\text{Concat}(h_{t,i}, h_{t+k_2,j}))), \tag{21}$$

$$\mathcal{L}_2 = -\mathbb{E}_{t,i,k_2,j}\left[Y_t^i(k_2, j)\log \hat{p} + (1 - Y_t^i(k_2, j))\log(1-\hat{p})\right] \tag{22}$$

where $\hat{p}$ is the predicted probability that the two segments are highly correlated. In practical application, we randomly choose 50% labels from each Y_t^i for efficient training.

Replace Discriminative Learning. Consistently exploiting correlation for all channels will weaken the specificity between channels. However, there are significant differences in the physiological signal patterns of different brain areas recorded by channels. Therefore, retaining the characteristics of each channel cannot be ignored for the modeling of brain signals. For this purpose, we further design the replace discriminative learning task.

Following BERT [13], we randomly replace r% local representations throughout z_t by $\hat{z}_t$, which is sampled from any $\mathcal{T}$ sequences and any C channels in z_t. We use the notation $\mathcal{I}(\hat{z}_t)$ to represent the new local representations after replacement and the corresponding channel indexes of $\hat{z}_t$ in the original sequence. We generate pseudo labels Y_t of the task as below:

$$Y_t(\tau, i) = \begin{cases} 1, & \mathcal{I}(\hat{z}_{t,\tau,i}) \neq i, \\ 0, & \mathcal{I}(\hat{z}_{t,\tau,i}) = i. \end{cases} \tag{23}$$

τ and i traverse $\mathcal{T}$ sequences and C channels of $\hat{z}_t$. After obtaining $\hat{z}_t$, we put it into the autoregressive model to get the new contextual representations $\hat{c}_t = g_{\text{ar}}(\hat{z}_t)$. Finally, a simple discriminator implemented by an MLP is utilized to classify whether $\hat{c}_t$ are replaced by other channels or not:

$$\mathcal{L}_3 = -\mathbb{E}_{t,\tau,i}\left[Y_t(\tau,i)\log\hat{q} + (1 - Y_t(\tau,i))\log(1-\hat{q})\right], \tag{24}$$

where $\hat{q}$ is the predicted probability that $\hat{c}_{t,\tau,i}$ is replaced. When the accuracy of discrimination increases, different channel representations output by the autoregressive model are easier to distinguish. Therefore, the task encourages the model to preserve the unique characteristics of each channel.

Combining the multi-task loss functions equation 17, equation 22 and equation 24, we jointly train *MBrain* with $\mathcal{L} = (1-\lambda_1-\lambda_2)\mathcal{L}_1 + \lambda_1\mathcal{L}_2 + \lambda_2\mathcal{L}_3$. After the SSL stage, the segment representations h_t obtained from equation 20 are used for downstream tasks.

4 EXPERIMENTS

4.1 Datasets and Baselines

SEEG dataset. The SEEG dataset used in our experiment is anonymous and provided by a first-class hospital we cooperate with. For a subject suffering from epilepsy, 4 to 10 invasive electrodes with 52 to 124 channels are used for recording signals. It is worth noting that since SEEG data are collected in a high frequency (1,000Hz or 2,000Hz) through multiple channels for several days, our data is massive. In total, we have collected 470 hours of SEEG signals with a total capacity of 550GB. Professional neurosurgeons help us label the epileptic segments for **each channel**.

We obtain the samples for each subject respectively. For the i-th subject, we first sample a dataset for self-supervised learning which is denoted as SS_i (80% for training and 20% for validation), then sample training set T_i, validation set V_i and testing set E_i for the downstream stage. SS_i, T_i and V_i contain 1,000, 800 and 200 10-second SEEG clips respectively, while E_i contains 510 10-second SEEG clips with positive-negative sample ratio of 1:50. There is no overlap among the samples of the three sets. We use a 1-second window to segment each clip without overlap and our target is to make predictions for **all channels** in each **1-second segment**.

EEG dataset. We use the Temple University Hospital EEG Seizure Corpus (TUSZ) v1.5.2 [35] as our EEG dataset. It is the largest public EEG seizure database, containing 5,612 EEG recordings, 3,050 annotated seizures from clinical recordings, and eight seizure types. We include 19 EEG channels in the standard 10-20 system. We randomly split the official TUSZ train set by subjects into training and validation sets at a ratio of 90/10 for model training and hyperparameter tuning respectively, and we keep out the official TUSZ test set for model evaluation. Therefore, the training, validation and testing sets consist of distinct subjects. After dividing the dataset by subjects, we start to sample EEG clips. For the self-supervised learning, we randomly sample 3,000 12-second unlabeled EEG clips for training and validation, with ratios of 90% and 10% respectively. As for the downstream task, we first obtain 3,000 sampled 12-second labeled EEG clips (80% for training and 20% for validation). Then, we sample another 3,900 12-second labeled EEG clips with positive-negative sample ratio of 1:10 for testing. It is worth noting that the labels of EEG data are coarse-grained, which means we only have the label of whether epilepsy occurs in a whole EEG clip.

Baselines. We compare *MBrain* with state-of-the-art models including one supervised classification model **MiniRocket** [12] and several self-supervised and unsupervised models: **CPC** [28], **SimCLR** [7], **Triplet-Loss (T-Loss)** [17], **Time Series Transformer (TST)** [45], **GTS** [36], **TS-TCC** [16] and **TS2Vec** [43].

4.2 Experimental Setup

For EEG data, as the number of subjects is large while the number of samples for each subject is very small, we follow the standard experimental setting to divide the training, validation and testing sets by subjects. As for SEEG data, since every subject includes many samples, it is accessible to sample training, validation and testing sets for each subject respectively. To demonstrate the effectiveness of *MBrain*, we first formally define the seizure detection task. Then we perform three experiments to show that our model outperforms the state-of-the-art baselines and has the ability to be deployed to clinical practice. We also show the ablation study and case study of the correlation graph in Section 4.6 and 4.7. The hyperparameter analysis is showed in Appendix E. We report the results of another downstream task of emotion recognition in Appendix F. In order to ensure the reliability of the experimental results, we repeat all the experiments five times with five different random seeds in the fine-tuning stage and report standard deviation in all tables.

Task 1 (Seizure Detection). ***Given*** *a time-ordered set including $I_{\mathcal{S}}$ consecutive time segments with the index of the first segment being t_0: $\mathcal{S} = \{s_{t_0}, \ldots, s_{t_0+I_{\mathcal{S}}}\}$, models* ***predict*** *the labels $\hat{Y}^s_{t,i}$ for all time segments in $\mathcal{S}$ (*i.e.*, $t = t_0, \ldots, t_0 + I_{\mathcal{S}}$) and all channels in each segment (*i.e.*, $i = 1, \ldots, \text{C}$).*

Subject dependent experiment [5]. Due to the larger difference between subjects in SEEG dataset than that in EEG dataset, we first perform the subject dependent experiment to obtain the upper bound of model performance on SEEG dataset. More specifically, for the i-th subject, we first perform self-supervised learning of the model on unlabeled data sampled from itself (*i.e.*, SS_i). When

training the downstream task, the encoder of SSL models will be **fine-tuned with a very low learning rate** on labeled data sampled from itself (*i.e.*, T_i and V_i). Finally, we test the models on E_i and report the average performance over all subjects. For fair comparison, we use **the same downstream model and experimental setup** for all models (see details in Appendix C).

Subject independent experiments. To meet practical clinical needs, we design two clinically feasible experiments. **The first is the domain generalization experiment**, that is, training the model on data of existing subjects and directly predicting data of unknown subjects. This is the standard experimental setting on EEG dataset. As for SEEG dataset, we follow the 3-1-1 setting, where 3 subjects are used for training (*i.e.*, SSL on SS_i, SS_j, SS_k; fine-tuning on T_i, T_j, T_k), 1 subject is used for validation (*i.e.*, V_m) and 1 subject is used for testing (*i.e.*, E_n). Note that i, j, k, m and n are indexes for different subjects. We conduct the experiments for random combinations, pick up the best result for each subject, and report the average results over all subjects.

The second is the domain adaptation experiment [27]. Different from the ideal domain generalization experiment which does not use the labeled data of target subjects at all, domain adaptation experiment allows using a small amount of the data to achieve better clinical performance of our model. This is because of the large data size due to the long-time records of the subjects in the SEEG dataset, and even if the model is fine-tuned with partially labeled data, it is clinically valuable to predict the large amount of remaining data in the target subjects. In this experiment, we first perform SSL on one subject (*i.e.*, source domain SS_i) and then fine-tuning is performed using partially labeled data from another subject (*i.e.*, target domain T_j and V_j). Finally, we perform seizure detection on the testing set of the target subject (*i.e.*, E_j). We pick up four subjects with typical seizure patterns in the SEEG dataset, and report the results of all one-to-one combinations.

Table 1: The average performance of the subject dependent experiment on SEEG dataset.

Models	Pre.	Rec.	F_1	F_2
MiniRocket	22.98±0.15	**66.24±0.26**	31.79±0.19	43.58±0.22
CPC	27.65±4.49	55.07±3.52	34.20±3.40	42.73±2.57
SimCLR	11.06±3.95	51.54±5.87	16.60±4.68	25.41±4.95
T-Loss	29.29±2.65	51.55±2.53	36.00±1.97	43.13±1.57
TST	13.60±3.48	44.65±4.21	19.80±3.73	28.41±3.29
GTS	24.29±4.26	40.39±5.80	29.16±2.97	34.17±2.36
TS-TCC	22.10±7.65	49.94±5.41	25.32±8.02	32.74±7.95
TS2Vec	30.56±2.17	52.83±2.89	36.03±1.72	43.35±1.59
MBrain	**37.97±2.75**	65.07±2.68	**46.45±2.25**	**55.28±1.77**

4.3 Subject Dependent Experiment

The average performance of the subject dependent experiment on the SEEG dataset is presented in Table 1. Since the positive-negative sample ratio of SEEG dataset is imbalanced, F-score is a more appropriate metric to evaluate the performance of models than only considering precision or recall. Especially in clinical applications, doctors pay more attention to finding as much seizures as possible, we thus choose F_1 and F_2 scores in the experiment. Overall, *MBrain* improves the F_1-score by 28.92% and the F_2-score by 26.85% on SEEG dataset, compared to the best baseline, demonstrating that *MBrain* can learn more informative representations from SEEG data. Through this experiment, we obtain the upper bound of the performance of models on SEEG dataset. We can find that it is still difficult to achieve high performance even if models are trained, verified and tested on the same subject. Combined with the analysis of subsequent experimental results, this reflects that seizure detection on SEEG data is much more difficult than that on EEG.

4.4 Domain Generalization Experiment

In this experiment, we validate and compare the generalization ability of all models under a strict setting, in which the models are trained on source subjects and then directly perform seizure detection on the unseen target subjects. This is an ideal scenario for clinical applications and the results are shown in Table 2. For SEEG dataset, in general, the performance of models under the domain generalization setting decreases significantly (41.73% on average in terms of F_2-score) compared with that in subject dependent experiment. The drop for recall metric is more pronounced, confirming that the distribution shift of subjects in SEEG data is more significant than that in EEG. This results from the fact that different brain regions and different types of epileptic waves have different physiological properties and patterns. Nonetheless, *MBrain* still improves F_1 and F_2 scores by 21.77% and 27.83% respectively, compared to the best baseline. The results prove that *MBrain* has a superior generalization ability benefiting from rational inductive assumption of model design. We point out that although GTS is also graph-based model, it directly learns the graph structure for each segment and ignores the stable and long-term correlations between different channels. This implies that our proposed graph structure learning strategy based on the stable correlations is reasonable and effective.

Table 2 also shows the results of domain generalization experiment on EEG dataset. Following the common evaluation scheme on EEG dataset [39], we add Area Under the Receiver Operating Characteristic (AUROC) metric in our experiment. Our model is designed to learn the representation for each channel, while there is only one label for an EEG clip. Therefore, it requires the pooling operation to aggregate representations output by our model over channels and time segments for seizure detection. This setting makes the performance improvement of our model not as significant as that in the SEEG experiment. Nevertheless, *MBrain* still outperforms all baselines on F_1-score, F_2-score and AUROC with an increase of 2.26%, 9.23% and 2.74%, respectively. SimCLR gets the highest recall but the lowest precision and AUROC, indicating that it may be not reasonable to regard time segments as independent samples without considering the contextual data. The worst performance for TST shows that mask-prediction SSL paradigm may not be suitable for non-stationary time series data.

4.5 Domain Adaptation Experiment

According to the results of domain generalization experiment, it is difficult for *MBrain* to achieve competitive performance as shown

Table 2: The average performance of the domain generalization experiment on SEEG and EEG datasets.

Models	SEEG				EEG				
	Pre.	Rec.	F_1	F_2	Pre.	Rec.	F_1	F_2	AUROC
MiniRocket	5.85±0.20	**39.18±0.59**	9.93±0.29	17.24±0.37	**22.86±0.84**	63.08±1.47	33.56±1.11	46.66±1.33	75.30±0.77
CPC	22.88±5.06	23.92±3.90	20.11±3.27	21.23±2.49	22.81±2.04	58.31±7.55	32.50±1.24	44.02±2.43	74.53±1.00
SimCLR	14.02±3.71	26.36±4.99	11.07±3.49	13.47±4.01	12.63±1.62	74.88±16.77	21.33±1.95	36.78±2.61	55.86±5.36
T-Loss	21.38±4.25	28.50±4.07	23.48±3.30	25.90±3.06	20.72±1.26	69.25±3.99	31.82±1.08	47.00±0.50	75.88±0.49
TST	8.37±3.96	32.48±8.25	11.80±3.91	15.67±3.69	15.65±1.54	28.59±12.93	19.65±4.36	23.87±8.09	58.20±4.27
GTS	24.16±5.91	27.99±4.98	22.77±2.69	24.15±2.79	18.86±1.09	62.51±5.04	28.88±0.88	42.54±1.48	71.69±1.88
TS-TCC	24.24±4.51	26.61±5.96	19.89±5.23	22.11±5.08	15.55±0.88	39.76±11.08	21.89±1.20	29.60±4.64	58.63±1.62
TS2Vec	27.93±5.23	29.49±3.97	26.78±3.29	27.88±3.52	21.40±0.63	58.31±6.14	31.24±1.18	43.24±2.78	73.35±1.02
MBrain	**30.69±5.92**	38.94±4.34	**32.61±3.60**	**35.64±3.04**	22.13±1.03	**76.99±4.49**	**34.32±0.90**	**51.34±0.97**	**77.96±0.97**

Table 3: The performance of the domain adaptation experiment on SEEG dataset in terms of F_2-score. DA row denotes the performance of *MBrain* in the domain adaptation experiment. Max-base and Non-DA rows represent the best performance of baselines and *MBrain* in the subject dependent experiment. We bold the best result and underline the second best result.

Setting	Group A			Group B			Group C			Group D		
	$B\rightarrow A$	$C\rightarrow A$	$D\rightarrow A$	$A\rightarrow B$	$C\rightarrow B$	$D\rightarrow B$	$A\rightarrow C$	$B\rightarrow C$	$D\rightarrow C$	$A\rightarrow D$	$B\rightarrow D$	$C\rightarrow D$
DA	68.55±4.27	<u>69.14±6.54</u>	68.78±4.12	41.08±2.59	46.06±3.05	<u>46.12±2.04</u>	40.04±3.98	39.34±2.11	**48.64±5.48**	<u>80.82±0.65</u>	79.90±1.11	80.72±1.31
Max-base	62.49±2.30			39.78±2.04			33.59±2.23			75.35±0.79		
Non-DA	**70.63±1.41**			**46.62±2.42**			<u>46.09±2.35</u>			**83.27±0.95**		

in Table 1 on SEEG dataset. The results show that seizure detection on SEEG dataset is much more difficult than that on EEG dataset. Alternatively, due to the long-time record, clinical SEEG data contains tens or even hundreds of seizures, allowing us to use a small amount of labeled data to fine-tune our model and then use it to predict the remaining data. In this way, *MBrain* can still achieve great performance, showing good generalization ability and clinical application value of our model. Table 3 shows the performance of the domain adaptation (DA) experiment for four subjects with typical seizure patterns provided by doctors from SEEG dataset. More specifically, we train *MBrain* on one subject and fine-tune it on all other three subjects. $B\rightarrow A$ denotes that the SSL model is trained on Subject-B, and then fine-tuned and tested on data from Subject-A. The results of Max-base and Non-DA rows correspond to the performance of the best baseline and *MBrain* respectively in scenarios $A\rightarrow A$, $B\rightarrow B$, $C\rightarrow C$ and $D\rightarrow D$.

Compared with the results of the setting that the self-supervised model and downstream model are both trained on the same subject, the F_2-scores of all 12 cross-domain scenarios reduce by less than 15%. Additionally, it can be observed that in all cross-domain scenarios, *MBrain* beats the best baseline in the corresponding scenarios without DA. It is worth noting that $D\rightarrow C$ scenario even outperforms corresponding Non-DA result. The possible reason is that the signal patterns on Subject-D are more significant and recognizable than those on Subject-C. Therefore, the SSL model trained on higher quality source domain can better distinguish signal states when performing downstream tasks on target domain. Overall, the domain adaptation experiment makes *MBrain* achieve competitive performance as shown in Table 1 by fine-tuning it on only a small amount of labeled data from the target domain. The results suggest that *MBrain* captures the inherent features and outputs generalized representations between subjects, because we fine-tune the SSL model with a very low learning rate (1e-6). From the perspective of pre-training, the SSL model trained on the source subject gives good initial parameters for the fine-tuning stage on the target subject.

4.6 Ablation Study

Considering the complexity of our model, we conduct sufficient ablation experiments to demonstrate the effectiveness of each component in *MBrain*. Specifically, we mainly compare *MBrain* with three types of different model variants.

(1) **Replace the method to aggregate channel information.** To verify the effectiveness of our proposed graph structure learning, we have proposed two ideas on how to directly implement the multi-channel CPC in Section 2. For the second idea, we have reported the results of a shared CPC regarding all channels as one on the CPC row of Table 1. For the first idea, we design two strategies to combine multi-channel CNN or MLP into CPC respectively to learn representations for each channel. See detailed description in Appendix D.

(2) **Remove one component.** We firstly remove the correlation graph structure learning module from the instantaneous time shift task and degenerate the task to single-channel CPC while still uniformly sampling negative samples in all channels. This variant is denoted as *MBrain* -Graph. Next, we respectively remove the whole instantaneous time shift task, the delayed time shift

task and replace discriminative task. These variants are denoted as *MBrain* -Instant, *MBrain* -Delay and *MBrain* -Replace.

(3) **Preserve one SSL task.** *MBrain* -onlyInstant, *MBrain* -onlyDelay and *MBrain* -onlyReplace indicate that *MBrain* only performs instantaneous time shift task, delayed time shift task and replace discriminative task respectively.

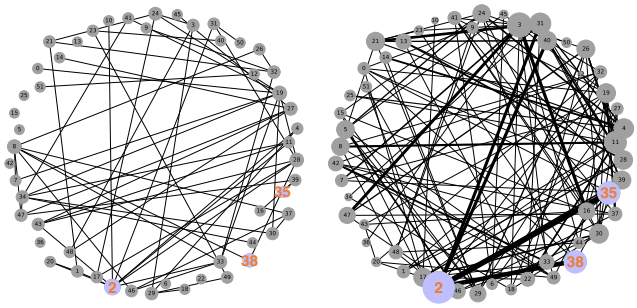
Table 4: The results of ablation study.

Models	Pre.	Rec.	F_1	F_2
CPC	27.65±4.49	55.07±3.52	34.20±3.40	42.73±2.57
CPC-Conv	6.39±0.77	33.21±4.00	10.53±1.07	17.46±1.42
CPC-MLP	25.84±3.07	52.70±3.65	32.18±2.46	40.34±2.05
MBrain -Graph	36.72±4.59	60.48±4.47	43.61±3.08	51.47±2.68
MBrain -Instant	34.49±4.37	55.41±3.90	41.57±3.48	48.38±2.52
MBrain -Delay	35.00±4.49	**65.61±2.94**	42.97±3.61	52.51±1.93
MBrain -Replace	36.08±5.35	63.67±4.24	43.66±3.66	52.49±2.32
MBrain -onlyInstant	36.43±4.44	63.66±2.12	43.35±3.83	51.82±2.67
MBrain -onlyDelay	31.59±4.24	55.03±5.26	38.56±2.84	46.05±2.26
MBrain -onlyReplace	34.13±6.84	56.06±3.68	40.02±4.47	47.44±2.40
MBrain	**37.97±2.75**	65.07±2.68	**46.45±2.25**	**55.28±1.77**

Table 4 shows the results of ablation study on SEEG dataset. It can be observed that the complete *MBrain* achieves the best performance on F_1 and F_2 scores, demonstrating the effectiveness of each component in our model design. For the first type of variants, we can observe that the performance of CPC-Conv decreases dramatically. We speculate that this is because the channels are relatively independent, and the correlation between most channels is weak or even non-existent. Direct adoption of multi-channel convolution may introduce spurious and noisy correlations. However, the graph structure learning proposed by us has a sparsity assumption, and the representation extraction of each channel is relatively independent, so it can effectively learn and aggregate more significant information. For CPC-MLP, we use an MLP to aggregate the representations of other channels, and then concatenate it with the representation of the target channel to predict future data. Unlike CPC-Conv, which adopts multi-channel convolution for the raw data to obtain the *mixed* low-level representations, CPC-MLP, like *MBrain*, learns the correlation of channels based on the *separate* high-level representations. Therefore, the performance of CPC-MLP does not drop as dramatically as that of CPC-Conv.

For *MBrain* -Instant, the significant decrease in performance illustrates that capturing the spatial and short-term patterns is quite important and is the key to learning the essential representations in multi-channel brain signals. For *MBrain* -Graph, the decrease in performance demonstrates that multi-channel CPC can greatly help learn more informative representations. Additionally, the performance in *MBrain* -Delay and *MBrain* -Replace also decreases significantly, illustrating that modeling long-term temporal patterns and preserving the characteristics of channels can help learn more distinguishable representations. For the third type of variants, it can be observed that the instantaneous time shift is the most important task, and the delayed time shift task and the replace discriminative task contribute similarly to the performance of the complete model.

4.7 Case Study



(a) Normal correlation graph. (b) Seizure correlation graph.

Figure 4: Case study on correlation graphs learned by *MBrain*.

In this section, we study the correlation graphs between the channels learned by *MBrain*. We randomly sample normal and seizure SEEG clips of one particular subject, and visualize their correlation graphs $\mathbf{A}_t$ (defined in Section 3.1) in Figure 4. In this figure, the thickness of an edge indicates its weight. And the larger the sum of weights of the edges connected to the node, the larger the size of the circle of the node. It can be observed that in the normal state, the correlation is sparser and the weights for edges are smaller, indicating a holistically weaker correlation between channels. In contrast, in the seizure state, the connection pattern between channels varies, where the correlation becomes denser and the edge weights become larger. Furthermore, in Figure 4(b), edges with larger weights are usually connected to 2 seizure channels. For example, Channel-2, Channel-35 and Channel-38 are all in seizure states and the edge weights between them are large, indicating that the brain areas recorded by the three channels have a higher probability of being the focal area. This can help neurosurgeons to better localize seizure lesions.

5 CONCLUSION

In this paper, we propose a general multi-channel SSL framework *MBrain*, which can be applied for learning representations of both EEG and SEEG brain signals. Based on domain knowledge and data observations, we succeed to use the correlation graph between channels as the cornerstone of our model. The proposed instantaneous and delayed time shift tasks help us capture the correlation patterns of brain signals spatially and temporally. The replace discriminative task helps *MBrain* learn a unique representations for each channel to achieve accurate channel-wise prediction. Extensive experiments of seizure detection on large-scale real-world datasets demonstrate the superior performance and clinical value of *MBrain*. However, there are still some limitations of our work. For example, negative sampling of multi-channel CPC consumes certain memory and time. As for the future work, we plan to collect more types of brain signals and extend *MBrain* to more downstream tasks.

Acknowledgment. This work is supported by NSFC (No.62176233), the National Key Research and Development Project of China (No.2018AAA0101900) and the Fundamental Research Funds for the Central Universities.

REFERENCES

[1] Fahd A. Alturki, Khalil AlSharabi, Akram M. Abdurraqeeb, and Majid Aljalal. 2020. EEG signal analysis for diagnosing neurological disorders using discrete wavelet transform and intelligent techniques. *Sensors* 20, 9 (2020), 2505.

[2] Anthony Bagnall, Jason Lines, Aaron Bostrom, James Large, and Eamonn Keogh. 2017. The great time series classification bake off: A review and experimental evaluation of recent algorithmic advances. *Data mining and knowledge discovery* 31, 3 (2017), 606–660.

[3] Hubert Banville, Omar Chehab, Aapo Hyvärinen, Denis-Alexander Engemann, and Alexandre Gramfort. 2021. Uncovering the structure of clinical EEG signals with self-supervised learning. *Journal of Neural Engineering* 18, 4 (2021), 046020.

[4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *NeurIPS*. 1877–1901.

[5] Junru Chen, Yang Yang, Tao Yu, Yingying Fan, Xiaolong Mo, and Carl Yang. 2022. BrainNet: Epileptic wave detection from SEEG with hierarchical graph diffusion learning. In *KDD*. 2741–2751.

[6] J. X. Chen, P. W. Zhang, Z. J. Mao, Y. F. Huang, D. M. Jiang, and Y. N. Zhang. 2019. Accurate EEG-based emotion recognition on combined features using deep convolutional neural networks. *IEEE Access* 7 (2019), 44317–44328.

[7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *ICML*. 1597–1607.

[8] Wei Chen and Ke Shi. 2021. Multi-scale attention convolutional neural network for time series classification. *Neural Networks* 136 (2021), 126–140.

[9] Xinlei Chen and Kaiming He. 2021. Exploring simple siamese representation learning. In *CVPR*. 15745–15753.

[10] Alexander Craik, Yongtian He, and Jose L Contreras-Vidal. 2019. Deep learning for electroencephalogram (EEG) classification tasks: A review. *Journal of Neural Engineering* 16, 3 (2019), 031001.

[11] Zachary W Davis, Lyle Muller, Julio Martinez-Trujillo, Terrence Sejnowski, and John H Reynolds. 2020. Spontaneous travelling cortical waves gate perception in behaving primates. *Nature* 587, 7834 (2020), 432–436.

[12] Angus Dempster, Daniel F Schmidt, and Geoffrey I Webb. 2021. Minirocket: A very fast (almost) deterministic transform for time series classification. In *KDD*. 248–257.

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional Transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[14] Carl Doersch, Abhinav Gupta, and Alexei A Efros. 2015. Unsupervised visual representation learning by context prediction. In *ICCV*. 1422–1430.

[15] Ruo-Nan Duan, Jia-Yi Zhu, and Bao-Liang Lu. 2013. Differential entropy feature for EEG-based emotion classification. In *NER*. 81–84.

[16] Emadeldeen Eldele, Mohamed Ragab, Zhenghua Chen, Min Wu, Chee Keong Kwoh, Xiaoli Li, and Cuntai Guan. 2021. Time-series representation learning via temporal and contextual contrasting. In *IJCAI*. 2352–2359.

[17] Jean-Yves Franceschi, Aymeric Dieuleveut, and Martin Jaggi. 2019. Unsupervised scalable representation learning for multivariate time series. In *NeurIPS*.

[18] Clive WJ Granger. 1969. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: journal of the Econometric Society* (1969), 424–438.

[19] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation* 9, 8 (1997), 1735–1780.

[20] Lina Elsherif Ismail and Waldemar Karwowski. 2020. Applications of EEG indices for the quantification of human cognitive performance: A systematic review and bibliometric analysis. *PloS one* 15, 12 (2020), e0242857.

[21] Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR*.

[22] Shiba Kuanar, Vassilis Athitsos, Nityananda Pradhan, Arabinda Mishra, and K.R. Rao. 2018. Cognitive analysis of working memory load from eeg, by a deep recurrent neural network. In *ICASSP*. 2576–2580.

[23] Christopher W Lynn and Danielle S Bassett. 2019. The physics of brain network structure, function and control. *Nature Reviews Physics* 1, 5 (2019), 318–332.

[24] Michael L Martini, Aly A Valliani, Claire Sun, Anthony B Costa, Shan Zhao, Fedor Panov, Saadi Ghatan, Kanaka Rajan, and Eric Karl Oermann. 2021. Deep anomaly detection of seizures with paired Stereoelectroencephalography and video recordings. *Scientific Reports* 11, 1 (2021), 1–11.

[25] Ishan Misra, C Lawrence Zitnick, and Martial Hebert. 2016. Shuffle and learn: Unsupervised learning using temporal order verification. In *ECCV*. 527–544.

[26] Mostafa Neo Mohsenvand, Mohammad Rasool Izadi, and Pattie Maes. 2020. Contrastive representation learning for Electroencephalogram classification. In *PMLR*. 238–253.

[27] Saeid Motiian, Marco Piccirilli, Donald A Adjeroh, and Gianfranco Doretto. 2017. Unified deep supervised domain adaptation and generalization. In *ICCV*. 5715–5725.

[28] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748* (2018).

[29] M. Paluszek, D. Avirovik, Y. Zhou, S. Kundu, A. Chopra, R. Montague, and S. Priya. 2015. 11 - Magnetoelectric composites for medical application. In *Composite Magnetoelectrics*. Woodhead Publishing, 297–327.

[30] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*.

[31] Piero Perucca, François Dubeau, and Jean Gotman. 2014. Intracranial electroencephalographic seizure-onset patterns: Effect of underlying pathology. *Brain* 137, 1 (2014), 183–196.

[32] Timothée Proix, Viktor K Jirsa, Fabrice Bartolomei, Maxime Guye, and Wilson Truccolo. 2018. Predicting the spatiotemporal diversity of seizure propagation and termination in human focal epilepsy. *Nature communications* 9, 1 (2018), 1–15.

[33] Khansa Rasheed, Adnan Qayyum, Junaid Qadir, Shobi Sivathamboo, Patrick Kwan, Levin Kuhlmann, Terence O'Brien, and Adeel Razi. 2020. Machine learning for predicting epileptic seizures using EEG signals: A review. *IEEE Reviews in Biomedical Engineering* 14 (2020), 139–155.

[34] Patrick Schäfer. 2015. The BOSS is concerned with time series classification in the presence of noise. *Data Mining and Knowledge Discovery* 29, 6 (2015), 1505–1530.

[35] Vinit Shah, Eva Von Weltin, Silvia Lopez, James Riley McHugh, Lillian Veloso, Meysam Golmohammadi, Iyad Obeid, and Joseph Picone. 2018. The Temple University Hospital seizure detection corpus. *Frontiers in Neuroinformatics* 12 (2018), 83.

[36] Chao Shang, Jie Chen, and Jinbo Bi. 2021. Discrete graph structure learning for forecasting multiple time series. In *ICLR*.

[37] Afshin Shoeibi, Marjane Khodatars, Navid Ghassemi, Mahboobeh Jafari, Parisa Moridian, Roohallah Alizadehsani, Maryam Panahiazar, Fahime Khozeimeh, Assef Zare, Hossein Hosseini-Nejad, et al. 2021. Epileptic seizures detection using deep learning techniques: A review. *International Journal of Environmental Research and Public Health* 18, 11 (2021), 5780.

[38] Tengfei Song, Wenming Zheng, Peng Song, and Zhen Cui. 2020. EEG emotion recognition using dynamical graph convolutional neural networks. *IEEE Transactions on Affective Computing* 11, 3 (2020), 532–541.

[39] Siyi Tang, Jared Dunnmon, Khaled Kamal Saab, Xuan Zhang, Qianying Huang, Florian Dubost, Daniel Rubin, and Christopher Lee-Messer. 2022. Self-supervised graph neural networks for improved Electroencephalographic seizure analysis. In *ICLR*.

[40] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NeurIPS*.

[41] Renping Yu, Han Zhang, Le An, Xiaobo Chen, Zhihui Wei, and Dinggang Shen. 2017. Connectivity strength-weighted sparse group representation-based brain network construction for M CI classification. *Human brain mapping* 38, 5 (2017), 2370–2383.

[42] Ye Yuan, Guangxu Xun, Kebin Jia, and Aidong Zhang. 2019. A multi-view deep learning framework for EEG seizure detection. *IEEE Journal of Biomedical and Health Informatics* 23, 1 (2019), 83–94.

[43] Zhihan Yue, Yujing Wang, Juanyong Duan, Tianmeng Yang, Congrui Huang, Yunhai Tong, and Bixiong Xu. 2021. TS2Vec: Towards universal representation of time series. *arXiv preprint arXiv:2106.10466* (2021).

[44] Seongjun Yun, Minbyul Jeong, Raehyun Kim, Jaewoo Kang, and Hyunwoo J Kim. 2019. Graph Transformer networks. In *NeurIPS*.

[45] George Zerveas, Srideepika Jayaraman, Dhaval Patel, Anuradha Bhamidipaty, and Carsten Eickhoff. 2021. A Transformer-based framework for multivariate time series representation learning. In *KDD*. 2114–2124.

[46] Xiang Zhang, Lina Yao, Xianzhi Wang, Jessica Monaghan, David Mcalpine, and Yu Zhang. 2021. A survey on deep learning-based non-invasive brain signals: Recent advances and new frontiers. *Journal of Neural Engineering* 18, 3 (2021), 031002.

[47] Xiang Zhang, Marko Zeman, Theodoros Tsiligkaridis, and Marinka Zitnik. 2022. Graph-guided network for irregularly sampled multivariate time series. In *ICLR*.

[48] Zhe Zhu. 2017. Change detection using landsat time series: A review of frequencies, preprocessing, algorithms, and applications. *ISPRS Journal of Photogrammetry and Remote Sensing* 130 (2017), 370–384.

A PRELIMINARIES

Brain signal data. For both EEG and SEEG data, there are multiple *electrodes* with $\mathbf{C}$ contacts that are sampled at a fixed frequency to record the brain signals. We also call these contacts *channels*. For every sampling point, each channel records the potential value of the brain region in which they are located, constituting abstract multi-channel time series data. A complete record file contains a total of $\mathbf{L}$ time points, for which we use the notation $X = \{x_l \in \mathbb{R}^{\mathbf{C}}\}_{l=1}^{\mathbf{L}}$ to represent. In this paper, we use i and j to denote the indexes of channels, such as $x_l = \{x_{l,i}\}_{i=1}^{\mathbf{C}}$. For every $x_{l,i}$, we assign a binary label $Y_{l,i} \in \{0, 1\}$ to it according to the start and end time of seizure signals marked by doctors. The time points are in the seizure state with positive labels ($Y_{l,i} = 1$), while zero labels ($Y_{l,i} = 0$) represent the normal data.

Preprocessing. Following the existing time series works [2, 34, 48] with the common preprocessing of segmentation, we use a $\mathbf{W}$-length window to divide the original data X into time segments $S = \{s_t \in \mathbb{R}^{\mathbf{W}\times\mathbf{C}}\}_{t=1}^{|S|}$ without overlapping. The number of segments $|S| = \lfloor \mathbf{L}/\mathbf{W} \rfloor$. The segment label is obtained from the time points of the whole segment, *i.e.*, $Y^s_{t,i} = \max\{Y_{t\times\mathbf{W}+1,i}, \ldots, Y_{(t+1)\times\mathbf{W},i}\}$.

B SINGLE-CHANNEL CPC

Contrastive Predictive Coding (CPC), a pioneering model for self-supervised contrastive learning, sets the pretext task to predict low-level local representations by high-level global contextual information c_t. In this way, the model can avoid learning too many details of the raw data and pay more attention to the contextual semantic information. The InfoNCE loss proposed in CPC has become the basic design of the contrastive learning loss function. Formally, given a raw data sample set $X = \{x_1, \ldots, x_N\}$ consisting of one positive sample from $p(x_{t+k}|c_t)$ and $N-1$ negative samples from the noisy distribution $p(x_{t+k})$, InfoNCE will optimize:

$$\mathcal{L}_N = -\mathbb{E}_X \left[\log \frac{f_k(x_{t+k}, c_t)}{\sum_{x_j \in X} f_k(x_j, c_t)} \right]. \tag{25}$$

In order to obtain the best classification probability of the positive sample with the cross entropy loss function, the optimal $f_k(x_{t+k}, c_t)$ is proportional to $p(x_{t+k}|c_t)/p(x_{t+k})$. Furthermore, the optimal loss function is also closely related to mutual information, as below:

$$\begin{aligned}
\mathcal{L}_N^{\text{opt}} &= -\mathbb{E}_X \left[\log \frac{p(x_{t+k}|c_t)/p(x_{t+k})}{p(x_{t+k}|c_t)/p(x_{t+k}) + \sum_{x_j \in X_{\text{neg}}} p(x_j|c_t)/p(x_j)} \right] \\
&\geq \mathbb{E}_X \left[\log \frac{p(x_{t+k})}{p(x_{t+k}|c_t)} N \right] \qquad (26) \\
&= -I(x_{t+k}; c_t) + \log N. \qquad (27)
\end{aligned}$$

Therefore, we can conclude that while minimizing the loss function $\mathcal{L}_N$, we are also constantly approximating the mutual information of raw data distribution $p(x_{t+k})$ and contextual semantic distribution $p(c_t)$. It turns out that InfoNCE is indeed a well-established loss function designed for self-supervised contrastive learning.

C IMPLEMENTATION DETAILS OF *MBRAIN*

The non-linear encoder g_{enc} used in *MBrain* is composed of three 1-D convolution layers and a one-layer LSTM model [19] is used as the autoregressive model g_{ar}. The model is optimized using Adam optimizer [21] with a learning rate of 2e-4 and weight decay of 1e-6 for the self-supervised learning stage. And for the downstream training stage, the downstream model is optimized with a learning rate of 5e-4 and weight decay of 1e-6 while the SSL model is fine-tuned with a low learning rate of 1e-6. For the hyperparameters of *MBrain*, we set $\theta_1 = 0.5$ and $\theta_2 = 0.5$. We set the maximum value of k_1 in instantaneous time shift task as 8. As Figure 5 shows, we set $K_2 = 7$ so as to take into account the step with the most significant correlation in delayed time shift task. Lastly, we build our model using PyTorch 1.8 [30] and train it on a workstation with 4 NVIDIA GeForce RTX 3090.

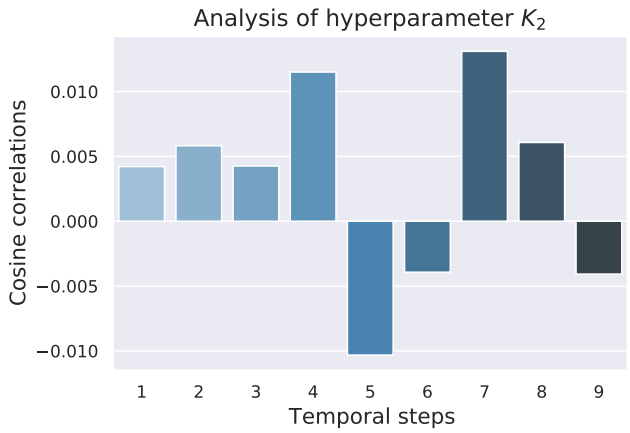


Figure 5: The data observation of how to choose hyperparameter K_2. We first average the correlations between each channel and all other channels in each time step. Then we average those of all channels in the same time step.

For the downstream task, we first utilize an LSTM model [19] to encode the segment representations of each channel in chronological order independently. One-layer self-attention [40] is then adopted to all channels within the same time step. Finally, a two-layer MLP classifier is used to predict whether seizure is occurring in the time segments. All baselines share the same downstream model in our experiments.

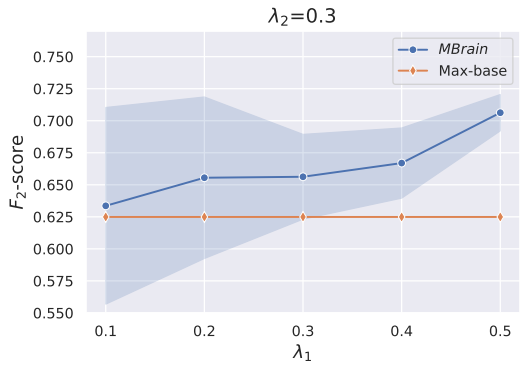
D DETAILS OF ABLATION STUDY

Replace the method to aggregate channel information. We design two strategies to combine multi-channel CNN or MLP into CPC respectively to learn representations for each channel.

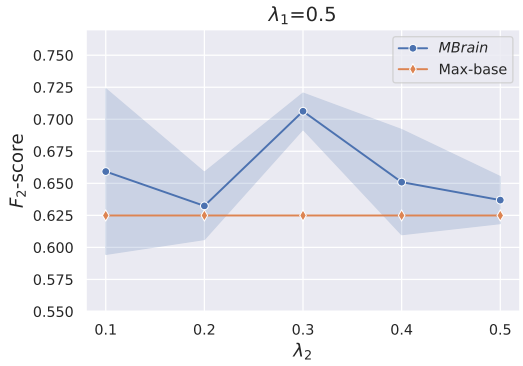
- Directly use 1-Dimension CNN to encode the whole time series data and the number of channels during the process is $\mathbf{C} \rightarrow 256 \rightarrow 256 \rightarrow \mathbf{C}\times 256$, and split the output into $\mathbf{C}$ representations, each of which is a 256-dimensional representation. Then an LSTM is implemented to it. Then we execute the self-supervised task and the downstream task of CPC based on the representations for each channel as *MBrain* does, this variant is denoted as CPC-Conv.
- We use the contextual representations of all n channels as input to an MLP in a fixed order, but we set the representation of the target channel to 0 tensor when we aggregate them. By using

the output of MLP as the aggregated representation of other channels, we perform subsequent experiments following exactly the same steps as *MBrain*. We name this variant as CPC-MLP.

E HYPERPARAMETER ANALYSIS



(a) Weights search for $\mathcal{L}_2$.



(b) Weights search for $\mathcal{L}_3$.

Figure 6: Sensitivity analysis on loss weights.

Sensitivity analysis on loss weights. Our loss function is defined as: $\mathcal{L} = (1-\lambda_1-\lambda_2)\mathcal{L}_1+\lambda_1\mathcal{L}_2+\lambda_2\mathcal{L}_3$, where $\mathcal{L}_1$, $\mathcal{L}_2$ and $\mathcal{L}_3$ are the loss of instantaneous time shift prediction task, delayed time shift prediction task and replace discriminative task respectively, and λ_1 and λ_2 are hyperparameters to balance the three pre-training tasks. We search both of the weights of λ_1 and λ_2 in the set {0.1, 0.2, 0.3, 0.4, 0.5} and report the tuning results with F_2-score for seizure detection task on subject-A from SEEG dataset. In 6(a) and 6(b), we can see that λ_1 = 0.5 and λ_2 = 0.3 lead to the optimal performance. In addition, *MBrain* consistently performs better than the best baseline.

Sensitivity analysis on replace ratio. We perform sensitivity analysis on replace ratio r% from replace discriminative task. We search the replace ratio from 5% to 95% and report the tuning results with F_2-score for seizure detection task on subject-A from SEEG dataset. As Figure 7 shows, when the replace ratio is set as 45%, *MBrain* has the best performance of 71.06±3.41. While *MBrain* gets the smallest standard deviation and the second best performance of 70.63±1.41 when the replace ratio is set as 15%.

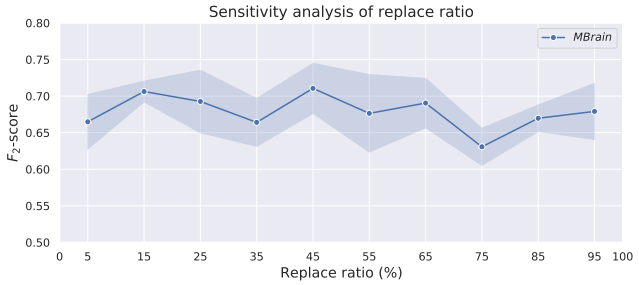


Figure 7: Sensitivity analysis on replace ratio r%.

F EMOTION RECOGNITION TASK

To measure the performance of our model on different downstream tasks, we use the SJTU Emotion EEG Dataset (SEED) [15] to test the model's performance in the emotion recognition task. In SEED, fifteen Chinese film clips (positive, neutral and negative emotions) were chosen from the pool of materials as stimuli used in the experiments. The duration of each film clip is approximately 4 minutes. We divide each EEG segment into 24-second segments without overlapping. For experimental efficiency, we downsample the segments to half the original frequency for each 24-second EEG segment. We randomly split the SEED dataset by subjects into train set, valid set and test set at a ratio of 3:1:1. We sample 3500 and 2000 EEG clips from the training patients for SSL and downstream task. We then sample 500 clips as validation set. Finally, we use all the data from the testing patients to evaluate models.

Table 5: The performance of models on SEED dataset.

Models	Acc.	AUROC
MiniRocket	49.80±0.60	75.28±0.17
CPC	48.23±4.36	73.48±1.51
SimCLR	44.84±5.82	63.05±6.52
T-Loss	47.90±3.99	68.56±5.96
TST	35.13±0.34	53.49±1.26
GTS	39.85±0.34	60.18±1.30
TS-TCC	40.10±5.50	66.38±3.39
TS2Vec	48.75±2.74	71.60±2.16
MBrain	**52.44±1.21**	**75.52±1.27**

Table 5 shows the results of *MBrain* and all baseline models on the emotion recognition task on SEED dataset. Since this is a 3-class classification task with balanced samples for each class, we only report the two metrics of Accuracy (Acc.) and AUROC. As can be seen from the table, *MBrain* improves the Accuracy by 5.30% and the AUROC by 0.32% on SEED dataset compared to the best result of baseline methods, demonstrating that *MBrain* still has a good performance on other downstream tasks of brain signals. It is worth noting that, like the TUSZ dataset, the SEED dataset only has coarse labels for each EEG clip, but our proposed *MBrain* aims to learn fine-grained representations. Therefore, the performance improvement on the SEED dataset is not as obvious as that on the SEEG dataset, but *MBrain* still outperforms all baseline models.