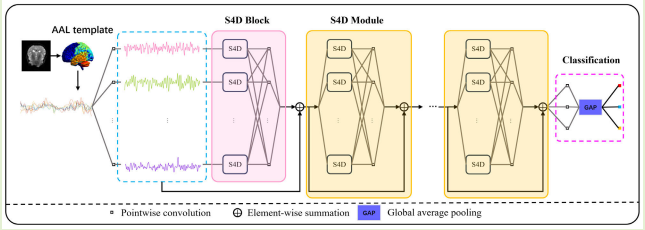
A Diagonal-Structured-State-Space-Sequence-Model-Based Deep Learning Framework for Effective Diagnosis of Mild Cognitive Impairment

Tangwei Cao, Xin Liu, Zuyu Du, Jiankui Zhou, Jie Zheng, and Lin Xu, *Senior Member, IEEE*

***Abstract*—Early diagnosis of mild cognitive impairment (MCI) may effectively prevent its development to Alzheimer's disease. Function connectivity (FC) of the brain networks is a widely used biomarker for MCI detection. However, FC estimated by predefined metrics may be unable to fully characterize the brain signals. This article aims to develop a deep learning framework directly applied to the brain signals for improved MCI diagnosis. A resting-state functional magnetic resonance imaging (rs-fMRI) dataset containing normal controls (NCs), early MCI (EMCI), and late MCI (LMCI) was used to develop and evaluate our model. Blood-oxygenation-level-dependent (BOLD) signals were measured by the fMRI. A 1-D pointwise convolution was employed to freely capture the spatial features, and a diagonal structured state space sequence (S4D) model was designed to extract the temporal features, particularly the long-term dependence of the BOLD signals. The proposed model was evaluated on three classification tasks, i.e., NC versus EMCI, EMCI versus LMCI, and NC versus EMCI versus LMCI, with repeated tenfold cross validation. Accuracy, sensitivity, specificity, and area under the receiver operating characteristic curve were calculated as performance metrics. For the two binary classification tasks, our model achieves the best performance in all metrics among seven state-of-the-art (SOTA) methods. For the three-category classification, despite slightly lower sensitivity, our model produces an overall superior performance than other methods. Our results indicate that long-term dependence of the BOLD signals may contribute significantly to MCI detection, providing useful information for automated diagnosis of MCI.**



***Index Terms*— Functional magnetic resonance imaging (fMRI), mild cognitive impairment (MCI), spatial filter, structured state-space sequence (S4) model.**

I. Introduction

MILD cognitive impairment (MCI) is a prodromal stage of Alzheimer's disease (AD), i.e., the most prevalent dementia form in the elderly population characterized by progressive perceptive and cognitive deficits. AD is an irreversible neurodegenerative disease and lacks effective clinical treatment [1]. It is reported that a considerable amount of MCI patients, e.g., 10%–15% per year, may convert to AD [2]. Early diagnosis and appropriate treatment and intervention of MCI can effectively prevent such conversion [3]. However, diagnosing MCI in an early stage can be a challenging task due mainly to the uncertain etiology, relatively mild symptoms, and subtle differences in the cognitive function between MCI patients and healthy elderly people [4]. Automated MCI diagnosis tools may, therefore, be demanded in clinical practice.

With recent advances in artificial intelligence, digital biomarkers have been extensively employed for automatic detection of many diseases [5], [6], [7]. Several studies have struggled to develop automated MCI diagnosis using advanced brain imaging techniques, such as electroencephalography (EEG) [8], functional near-infrared spectroscopy (fNIRS) [9], and magnetic resonance imaging (MRI) [10]. In particular, the blood-oxygenation-level-dependent (BOLD) signals measured by the resting-state functional

Manuscript received 18 October 2023; revised 2 April 2024; accepted 8 April 2024. Date of publication 16 April 2024; date of current version 15 May 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 62171284 and in part by a start-up funding from the ShanghaiTech University. The associate editor coordinating the review of this article and approving it for publication was Dr. Marko Vauhkonen. *(Corresponding author: Lin Xu.)*

Tangwei Cao, Xin Liu, Zuyu Du, Jiankui Zhou, and Jie Zheng are with the School of Information Science and Technology, ShanghaiTech University, Shanghai 201210, China.

Lin Xu is with the School of Information Science and Technology, ShanghaiTech University, Shanghai 201210, China, and also with the Shanghai Engineering Research Center of Energy Efficient and Custom AI IC, Shanghai 201210, China (e-mail: xulin1@shanghaitech.edu.cn).

Digital Object Identifier 10.1109/JSEN.2024.3387103

MRI (rs-fMRI) can capture neuronal activities by measuring the time-varying blood oxygenation level within structurally segregated and functionally specialized brain regions. It is, therefore, widely employed as a noninvasive measure for MCI diagnosis.

Based on the BOLD signals, brain functional connectivity (FC) has been extensively employed for MCI identification [11], [12], [13], [14], [15], as MCI is reported to associate with disconnection syndrome within the brain network [15]. By assuming stationary interactions between different brain regions throughout the whole recording period, static FC was first investigated by Qiao et al. [13], Yu et al. [16], and Zhang et al. [17]. Besides, taking the discover that the FC is dynamic rather than stationary into consideration [12], recent studies have proposed dynamic FC for MCI diagnosis by dividing the BOLD signals into consecutive segments using a sliding window [12], [15], [18].

Several measures, such as Pearson's correlation [19], partial correlation [20], coherence [21], and phase-locked value [22], have been proposed to estimate the interactions between different brain regions in order to establish stationary or dynamic FC. Based on the established FC, traditional machine learning approaches consisting of manual feature extraction and classification [16], [17], [18], [23], [24] or deep learning approach extracting the spatial, temporal, and/or spatiotemporal representation automatically have been investigated for effective MCI diagnosis [25], [26], [27], [28], [29], [30]. These methods can extract information from predefined FC and have achieved promising performance in MCI diagnosis.

However, the effectiveness of FC-based methods may be determined by the metrics used in the estimation of the interactions between different regions. In general, a complex FC incorporating multiple metrics has a stronger representation ability for MCI diagnosis than a simple FC using only one metric [23]. Furthermore, a predefined FC is usually insufficient to fully characterize the BOLD signals and, thus, may result in loss of valuable information for following analysis, limiting its performance for MCI diagnosis.

To overcome these limitations, some studies have proposed to use deep neural networks for the estimation of the dynamic FC of the brain [31]. Yet, a more straightforward idea is to extract spatial, temporal, and/or spatial–temporal representation directly from the original BOLD signals rather than the FC. In fact, many studies have employed 1-D convolutional neural networks (1-D CNNs) [32] and long short-term memory (LSTM) networks [33] to learn features directly from the BOLD signals for the diagnosis of autism. One limitation of the 1-D CNN is that the output features at each layer are constrained by the receptive field of the kernel size, resulting in limited (local) temporal representation.

With the assumption that the observed BOLD signals are generated by a sequence of underlying mental processes, recent studies have proposed deep state-space models (SSMs) to learn spatial–temporal features from the BOLD signals [34]. Moreover, a structured state-space sequence (S4) model can be obtained by specifying the state-transition matrix in the SSM as a high-order polynomial projection operator (HiPPO) [35]. The HiPPO matrix allows the S4 model to capture long-term temporal dependencies in the BOLD signals and, therefore, has achieved promising performance in related tasks, such as the diagnosis of autism spectrum disorders and depression disorders [34]. However, training the S4 model, in particular, the state-transition matrix, efficiently requires several computational tricks. Parameterizing the state-transition matrix as a diagonal plus low-rank (DPLR) matrix was proposed for faster computation [36]. Moreover, a single diagonal matrix is further proposed as an approximation of the state-transition matrix for improved computation, which is referred to as S4D [36], [37].

Despite the promising performance in the diagnosis of autism and depression, the aforementioned methods directly applied to the BOLD signals have seldom been employed for the diagnosis of MCI. The aim of this article is to develop a deep learning model directly applied to the BOLD signals for effective MCI diagnosis. To this end, a pointwise spatial convolution is employed to learn spatial representation directly from the BOLD signals, permitting the interpretability of extracted spatial features. In addition, an S4D model is employed to capture the temporal, particularly long-term, representation. The effectiveness of the proposed model on MCI diagnosis has been evaluated on the Alzheimer's Disease Neuroimaging Initiative (ADNI) database.

II. Materials and Methods

A. Materials and Preprocessing

The adopted public dataset, ADNI-2, consists of structural and functional MRI of NC and patients with early MCI (EMCI), late MCI (LMCI), and AD. Three classification tasks were tested in this article, including two binary classification tasks, i.e., NC versus EMCI and EMCI versus LMCI, and one classification among NC, EMCI, and LMCI. The adopted patient size was 172, 188, and 161 for NC, EMCI, and LMCI, respectively.

For each subject, $T1$-weighted structural and functional images were acquired at the rest state (subjects had their eyes open). The fMRI parameters were TE $=$ 30 ms, TR $=$ 3000 ms, voxel size $= 3 \times 3 \times 3.3$ mm^3, number of slices $=$ 48, flip angle $=$ 80 degrees, and number of volumes $=$ 140. The first three volumes were excluded from further analysis in order to avoid magnetization disequilibrium.

A standard pipeline was employed to preprocess the structural and functional images. First, the structural images were anterior commissure (AC)—posterior commissure (PC) corrected and resampled to $256 \times 256 \times 256$ with a resolution of $1 \times 1 \times 1$ mm^3. The N3 algorithm [38] was then used for intensity-inhomogeneity correction. The corrected structural images were finally skull-stripped and registered to the Montreal Neurological Institute (MNI) space.

The functional images were first corrected for slice timing by interpolation. Head motion was then corrected by rigid body transformation with mutual information as the cost function. The corrected fMRI was rigidly registered to the corresponding $T1$ MR images and aligned to the MNI space. The automated anatomical labeling (AAL) template [39] was then employed to demarcate the functional images into 116 brain regions of interest (ROIs), resulting in 116-channel BOLD signals. Spatial filtering using a Gaussian kernel with a 4-mm

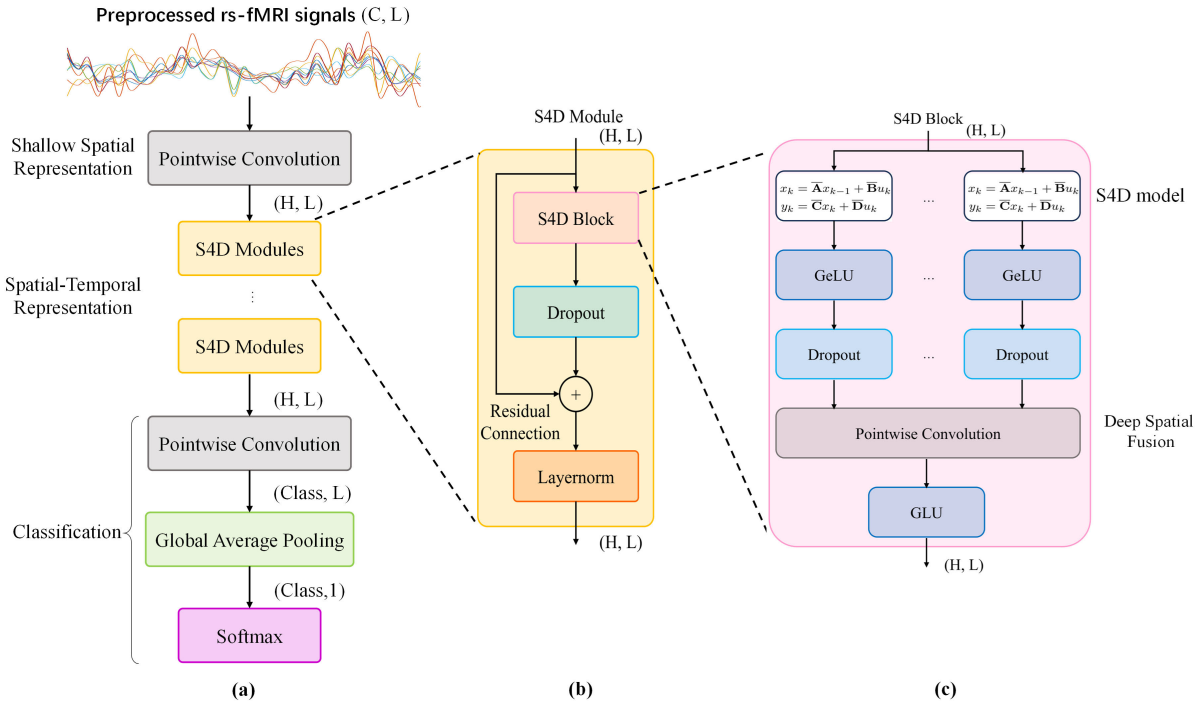


Fig. 1. Scheme of the proposed deep learning frame work. (a) Entire framework, (b) S4D module, and (c) S4D block.

full-width at half-maximum (FWHM) and temporally filtering between 0.01 and 0.1 Hz were applied to the obtained BOLD signals.

B. Proposed Model

The idea of this article was to develop a deep learning model that extracts spatial and temporal features directly from the BOLD signals for effective MCI diagnosis. The scheme of the proposed model is shown in Fig. 1, which consists of three main parts. First, a pointwise convolution layer was applied for a direct and shallow spatial representation of the preprocessed multichannel BOLD signals in order to mix information from different channels (ROIs). Then, an S4D module was employed to learn the long-term dependency and the discriminant temporal information from the spatially mixed time sequences. Finally, a classification part consisting of a pointwise convolution layer and a global average pooling (GAP) was utilized, responding for category-feature extraction and disease prediction.

1) Shallow Spatial Representation: Pointwise convolution is a very simple structure (1 × 1 convolution) and has been widely employed to mix information across multiple channels as a spatial filter [40], [41], [42], [43]. It is, therefore, applied directly to the multichannel BOLD signals for a shallow spatial representation in this article, as shown in Fig. 1(a). In general, the pointwise spatial convolution can be formulated as follows:

$$\mathbf{Y} = [\mathbf{W}_1 \quad \mathbf{W}_2 \quad \cdots \quad \mathbf{W}_H]^T [\mathbf{X}_1 \quad \mathbf{X}_2 \quad \cdots \quad \mathbf{X}_L] \quad (1)$$

where $\mathbf{X}_k$ ($k = 1{:}L$, L the number of time samples) is a C-channel column vector, e.g., the BOLD signal at the time instance k, $\mathbf{W}_i$ ($i = 1{:}H$, H the number of spatial filters) is a C-dimensional spatial filter (column vector), and $\mathbf{Y}$ is the output with the dimensions of $H \times L$.

In this article, the pointwise convolution layer was employed for effective learning of the interaction patterns among multiple brain regions. Each feature map generated by one pointwise filter contains information from all channels (ROIs), while H pointwise convolution filters could extract H different interaction patterns according to the weights of each filter. To alleviate the overfitting issue, the pointwise convolution layer was followed by a dropout layer in this article, which randomly dropped units (along with their connections) from the neural network during training to prevent excessive coadapting [44].

Different from graphic-neural-network (GNN)-based methods that generally take FC as the input [28], [29], the pointwise convolution layer frees the calculation of FC and is easier to implement for learning spatial patterns or interaction patterns. Moreover, by analyzing the weights of pointwise convolution filters, it is possible to show the significant brain regions and discriminative interaction patterns.

2) Temporal Representation and Deep Spatial Fusion: The core of the proposed method was to integrate a diagonal structured SSM (S4D) model into the deep learning framework, enabling its ability to learn temporal features with long-term dependencies. The working principles of the S4D model are briefly summarized hereafter.

An SSM represents a linear, time-invariant, and finite-dimensional dynamical system as a set of input, output, and state variables in the first-order differential equations [45]. Its continuous form is defined as follows:

$$\begin{aligned} x'(t) &= \mathbf{A}x(t) + \mathbf{B}u(t) \\ y(t) &= \mathbf{C}x(t) + \mathbf{D}u(t) \end{aligned} \quad (2)$$

where $u(t)$ is a 1-D input signal; $x(t)$ is an N-dimensional latent state; $y(t)$ is a 1-D output signal; and $\mathbf{A} \in \mathcal{C}^{N\times N}$,

$\mathbf{B} \in \mathcal{C}^{N\times 1}$, $\mathbf{C} \in \mathcal{C}^{1\times N}$, and $\mathbf{D} \in \mathcal{C}^{1\times 1}$ are the state-space matrices.

With a bilinear method, the discrete form SSM is defined as follows [46]:

$$\begin{aligned} x_k &= \overline{\mathbf{A}}x_{k-1} + \overline{\mathbf{B}}u_k \\ y_k &= \overline{\mathbf{C}}x_k + \overline{\mathbf{D}}u_k \end{aligned} \tag{3}$$

where $\overline{\mathbf{A}} = (\mathbf{I}-\Delta/2\cdot\mathbf{A})^{-1}(\mathbf{I}+\Delta/2\cdot\mathbf{A})$, $\overline{\mathbf{B}} = (\mathbf{I}-\Delta/2\cdot\mathbf{A})^{-1}\Delta\mathbf{B}$, $\overline{\mathbf{C}} = \mathbf{C}$, $\overline{\mathbf{D}} = \mathbf{D}$, and Δ is the step size utilized to convert the continuous sequence to a discrete one. Omitting $\overline{\mathbf{D}}$ and setting the initial state $x_0 = \mathbf{0}$, (3) can be unrolled as follows [37]:

$$\begin{aligned} y_k &= \overline{\mathbf{C}}\overline{\mathbf{A}}^k\overline{\mathbf{B}}u_0 + \overline{\mathbf{C}}\overline{\mathbf{A}}^{k-1}\overline{\mathbf{B}}u_1 + \cdots + \overline{\mathbf{C}}\overline{\mathbf{A}}\overline{\mathbf{B}}u_{k-1} + \overline{\mathbf{C}}\overline{\mathbf{B}}u_k \\ y &= \overline{\mathbf{K}} * u, \quad \overline{\mathbf{K}} = \left(\overline{\mathbf{C}}\overline{\mathbf{B}}, \overline{\mathbf{C}}\overline{\mathbf{A}}\overline{\mathbf{B}}, \ldots, \overline{\mathbf{C}}\overline{\mathbf{A}}^{L-1}\overline{\mathbf{B}}\right) \end{aligned} \tag{4}$$

where L is the length of the input sequence u_k. The SSM model can, therefore, be interpreted as a sequence-to-sequence convolution layer, in which $\overline{\mathbf{K}}$ is the convolution kernel or filter with learnable parameters $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, and $\mathbf{D}$.

The basic SSM performs actually very poorly in practice due most probably to possible gradient scaling exponentially in the sequence length, i.e., the vanishing gradients problem [47]. The HiPPO theory of continuous-time memorization [35] is then incorporated into the SSM model [37]. Specifying the structure of the state matrices A and B as HiPPO matrices can theoretically and empirically enable the state $x(t)$ to memorize the history of the input $u(t)$ [35], [48]. Several structures have been proposed for the state matrices [35], [49], among which the HiPPO-FouT structure introduced in [49] is employed in this article, as given by

$$\boldsymbol{A}_{nk} = \begin{cases} -2, & n = k = 1 \\ -2\sqrt{2}, & n = 1, k \text{ odd} \\ -2\sqrt{2}, & k = 1, n \text{ odd} \\ -4, & n, k \text{ odd} \\ 2\pi k, & n - k = 1, k \text{ odd} \\ -2\pi n, & k - n = 1, n \text{ odd} \\ 0, & \text{otherwise} \end{cases}$$

$$\boldsymbol{B}_n = \begin{cases} 2, & n = 1 \\ 2\sqrt{2}, & n \text{ odd} \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

Training $\overline{\mathbf{K}}$ efficiently requires several computational tricks. Studies have proposed to parameterize A as a DPLR matrix for faster computation of the convolution kernel using the Cauchy-kernel algorithm [36], [37]. Recent studies have proposed a single diagonal matrix to approximate the state-transition matrix $\mathbf{A}$ for improved computational efficiency, resulting in a model referred to as S4D, which is the key component of the deep learning model proposed in this article, as shown in Fig. 1(c).

The S4D model performs sequence-to-sequence mapping of 1-D input signal. In order to process the H signals generated by the spatial convolution layer, a number of S4D models, e.g., H, were employed with each handling one channel. The feature maps generated by each S4D were activated by a Gaussian error linear unit (GeLU), which is reported to generate better performance than traditional activations, such as ReLU and ELU [50]. The GeLU is generally calculated as follows:

$$\text{GeLU}(x) = 0.5x\left(1 + \tanh\left[\sqrt{2/\pi}\left(x + 0.044715x^3\right)\right]\right). \tag{6}$$

A dropout layer was used after the GeLU in order to alleviate the overfitting issue.

Then, a pointwise convolution layer was employed to fuse multichannel features, i.e., mapping an $H \times L$ feature map to a $2H \times L$ feature map. The output of the pointwise convolution was divided into two feature maps with dimensions of $H \times L$ for each, denoted as a and b, by a gated linear unit (GLU). The GLU outputted the $(H \times L)$ feature map, calculated as $a \otimes \sigma(b)$, where $\otimes$ represented elementwise product and σ represented sigmoid activation, which performed a gating mechanism that helped gradients flow through the layers and allowed for faster convergence [51]. The H S4D models, GeLUs, dropout, pointwise convolution, and the GLU composed a basic S4D block, as shown in Fig. 1(c).

The S4D block was the key component of the S4D module designed for spatial–temporal representation of the BOLD signals [Fig. 1(b)]. The output feature maps of the S4D block first passed through another dropout layer and were then added to the input feature maps of the S4D block, building a residual connection that enabled improved optimization through mitigating and vanishing gradients [52]. A normalization layer was employed to normalize the residual feature maps on every single batch along the input dimension of the channel, leading to smoother gradients, faster convergence, and better generalization [53]. In this article, a cascade of S4D modules was used to build a deep network for effective capture of the long-term dependency of the BOLD signals.

3) Classification: A classification part was designed to predict the score of each label. It consisted of another pointwise convolution layer, a dropout layer, and a GAP layer. The pointwise convolution layer was designed to extract category features from the input feature maps and generate one feature map for each corresponding category using 1×1 convolution. The GAP transformed the category features into category confidence or prediction scores by averaging the features of each category. Finally, the classification result was generated by feeding the GAP-resulted vector directly into the softmax layer. Compared to fully connected layers, GAP could avoid overfitting caused by the full connection and was more robust to time translations of the BOLD signals, as it summed out the temporal information [54].

C. Methods for Comparison

In order to evaluate its effectiveness, the proposed method was compared with the following state-of-the-art (SOTA) methods.

1) Strength- and Similarity-Guided Group Sparse Representation (SSGSR): The GSR is jointly guided by low-order FC and high-order FC. The resulted weights are then used as features in SVM for classification [23].

2) Attention-Diffusion-Bilinear Neural Network: The strength of node interactions is learned by an attention-diffusion-bilinear neural network (ADB-NN) framework through coupling FC and structural connectivity (SC) and is further aggregated for diagnosis [25].

3) Spatiotemporal Graph Convolutional Network: A spatiotemporal graph convolutional network (ST-GCN) is trained on short subsequences of the BOLD time series to model the nonstationary nature of FC, and the importance of graph edges within ST-GCN is helpful for prediction [26].

4) Dynamic Effective Connectivity With a Virtual Adversarial Training Convolutional Neural Network: The dynamic effective connectivity is built with a group-constrained Kalman filter algorithm. Then, the local features are extracted by a virtual adversarial training convolutional neural network (VAT-CNN) and inputted into a high-order connectivity weight-guided graph attention network for feature aggregation [27].

5) Sparse Representation With Latent Temporal Dependency (SRiLT): The brain FC is constructed by introducing temporal dependency and sequential information with a latent variable. The coefficients of the constructed brain FC are used as features for MCI classification [24].

6) Spatiotemporal Aggregated Attention Network: The spatial and dynamic relationships among different ROIs are simultaneously extracted by the spatiotemporal aggregated attention network (STAA) module with a self-attention mechanism and dynamic GCN. The extracted features are aggregated and further fed into a classifier [28].

The methods in [23], [24], [25], [26], and [27] are originally evaluated independently under quite different sample sizes. These methods are later reproduced in [28] based on the same dataset as that used in this article. The results reproduced in [28] are, therefore, considered for simple and fair comparisons. However, the patient size in [28] (154 NC, 168 EMCI, and 120 LMCI) is slightly smaller than that of this article (172 NC, 188 EMCI, and 161 LMCI) due to the specific exclusion criterion adopted in [28].

D. Experiment Setting and Implementation Details

Based on the adopted ADNI-2 dataset, three classification tasks were tested in our study in order to evaluate the effectiveness of the proposed model, including two binary classification tasks, i.e., NC versus EMCI and EMCI versus LMCI, and a three-category classification task among NC, EMCI, and LMCI. To fully assess the generalization performance of the model, a tenfold cross-validation strategy was used. The model was trained on nine groups and applied to the remaining group to evaluate the classification performance, rounding until all groups underwent classification. Accuracy (ACC), sensitivity (SEN), and area under the receiver operating characteristic curve (AUC) were considered as performance metrics. The cross-validation procedure was repeated for ten random subdivisions, and the average and standard deviation of the adopted metrics were considered.

To train the S4D model embedded in the proposed deep learning framework, the state matrices **A**, **B**, **C**, and **D** needed to be first initialized. In this article, the S4D-Lin initialization [36] was implemented, where the S4-FouT HiPPO matrix **A** in (5) was approximated by a diagonal matrix, with the diagonal elements $a_{nn} = jn\pi - 0.5$. Diagonal approximation of **A** permitted fast computation of the SSM model while keeping the ability to capture long-term dependencies. Besides, for model simplicity, all elements in **B** was set to 1, as it were reported that a trainable **B** provides minor improvement [36]. Finally, **C** and **D** were randomly initialized.

Then, the AdamW optimizer and a cosine annealing scheduler were used for weight decay regularization as suggested by previous work [36]. The AdamW optimizer modifies Adam to recover the original formulation of weight decay regularization and, therefore, is a powerful tool for improving the accuracy and speed of deep learning models [55]. The cosine annealing scheduler allows the learning rate to decay in the shape of a cosine function during training and, therefore, is helpful for convergence [56].

Following prior works [34], [36], the training parameters were set as follows: dropout rate $= 0.5$, epoch $= 200$, batch size $= 16$, number of S4D module or layer $= 4$, learning rate $= 1 \times 10^{-4}$ and weight decay $= 0$ for S4D, and learning rate $= 5 \times 10^{-4}$ and weight decay $= 1 \times 10^{-4}$ for other layers. Moreover, the number of spatial filters, i.e., H in (1), was set to 256, and the same for the number of state variables (N). The proposed model as well as the SOTA methods for comparison were implemented with PyTorch on a GPU (NVIDIA RTX 3090 24 GB).

III. Results

A. Performances of Our Model and the SOTA Methods

The classification performance of the proposed deep learning framework for the three classification tasks is reported in Table I, together with the SOTA methods. It is clear that, for the two binary classification tasks, the proposed model performs the best in all metrics as compared to the SOTA methods. In particular, it achieves an ACC, AUC, and SEN of 87.4%, 95%, and 86.4%, respectively, for the classification between NC and EMCI, and 85.0%, 93%, and 89.0% for the classification between EMCI and LMCI.

For the three-category classification task among NC, EMCI, and LMCI, the proposed model yields the highest ACC and AUC among all the methods, i.e., 77.9% and 92%, respectively. The sensitivity of the proposed model for distinguishing NC among all the labels, i.e., SEN_N, is relatively low as compared to the SOTA methods, i.e., ranking the fourth among the seven methods. Apart from SEN_N, the proposed model produces slightly lower sensitivity for distinguishing EMCI and LMCI, i.e., $\text{SEN}_E = 80.1\%$ and $\text{SEN}_L = 76.1\%$, respectively, both ranking the second among all the methods. Therefore, taking all the metrics into consideration, the proposed model produces the best overall performance for the three-category classification task.

B. Ablation Experiments

To illustrate the effectiveness of the proposed deep learning framework for spatial–temporal representation, two ablation experiments are performed on the same dataset. The first one was to assess the contribution of the shallow spatial presentation in Fig. 1(a) by simply removing this layer. The second one was to examine the effectiveness of the S4D model for temporal representation by replacing the S4D blocks in Fig. 1(c) with popular recurrent neural networks (RNNs), which also perform spatial–temporal representation through

TABLE I
PERFORMANCE OF OUR MODEL AND THE SOTA METHODS

Method	NC vs. EMCI			EMCI vs. LMCI			NC vs. EMCI vs. LMCI				
	ACC (%)	AUC	SEN (%)	ACC (%)	AUC	SEN (%)	ACC (%)	AUC	SEN_N (%)	SEN_E (%)	SEN_L (%)
SSGSR	80.1±1.7*	0.82±0.04*	82.5±1.4*	75.4±1.9*	0.79±0.09*	66.1±3.5*	71.5±2.3*	0.79±0.10*	77.9±0.3	67.3±2.6*	69.2±2.7*
ADB-NN	81.4±2.0*	0.84±0.07*	80.4±1.9*	82.6±2.3*	0.84±0.06*	74.2±2.4*	73.5±1.7*	0.83±0.03*	76.6±1.0	70.8±2.1*	73.3±1.4*
ST-GCN	82.3±1.1*	0.86±0.05*	83.3±1.7*	80.6±1.5*	0.86±0.03*	71.7±2.1*	75.6±0.7*	0.84±0.04*	**82.5**±0.6*	72.0±0.8*	71.7±1.2*
VAT-CNN	83.2±1.6*	0.83±0.05*	84.5±1.3*	80.9±2.7*	0.81±0.08*	67.5±3.3*	72.2±2.5*	0.81±0.07*	79.2±2.0 *	69.1±2.9*	67.5±3.0*
SRiLT	80.1±4.2*	0.85±0.06*	81.6±2.3*	84.4±2.4	0.88±0.03*	75.0±2.9*	76.0±1.2*	0.85±0.05*	74.0±2.1*	77.4±1.5*	**76.7**±1.3
STAA	78.3±0.5*	0.84±0.04*	79.2±0.7*	77.8±1.0*	0.81±0.04*	67.5±0.6*	72.4±1.3*	0.83±0.04*	65.6±1.2*	**82.7**±1.1*	66.7±0.9*
Proposed	**87.4**±0.7	**0.95**±0.01	**86.4**±1.8	**85.0**±1.0	**0.93**±0.01	**89.0**±1.7	**77.9**±1.1	**0.92**±0.01	77.2±2.1	80.1±2.5	76.1±2.2

Data are expressed as mean ± standard deviation. * indicates significant difference ($p < 0.05$) between each method and the proposed method.

TABLE II
RESULTS OF THE ABLATION EXPERIMENTS

Method	NC vs. EMCI			EMCI vs. LMCI			NC vs. EMCI vs. LMCI				
	ACC (%)	AUC	SEN (%)	ACC (%)	AUC	SEN (%)	ACC (%)	AUC	SEN_N (%)	SEN_E (%)	SEN_L (%)
GRU	82.1±1.3*	0.89±0.01*	83.1±2.0*	82.0±1.6*	0.88±0.01*	80.7±1.3*	72.5±1.3*	0.88±0.01*	73.6±2.6*	72.5±2.0*	71.3±3.6*
LSTM	81.3±1.4*	0.89±0.02*	83.8±2.2*	80.8±2.1*	0.88±0.01*	80.3±3.2*	71.8±2.0*	0.87±0.01*	72.0±3.5*	72.7±1.4*	70.5±3.4*
BGRU	82.8±1.0*	0.89±0.01*	84.1±1.4*	82.5±1.4*	0.90±0.01*	82.8±2.1*	74.4±1.0*	0.88±0.01*	77.2±3.0	73.6±2.1*	72.6±2.7*
BLSTM	81.8±0.9*	0.90±0.01*	83.2±1.5*	81.6±1.1*	0.89±0.01*	82.0±2.4*	73.2±0.9*	0.88±0.01*	72.7±2.8*	72.5±2.5*	74.4±1.9
Proposed	**87.4**±0.7	**0.95**±0.01	**86.4**±1.8	**85.0**±1.0	**0.93**±0.01	**89.0**±1.7	**77.9**±1.1	**0.92**±0.01	**77.2**±2.1	**80.1**±2.5	**76.1**±2.2
Proposed-S[1]	86.2±1.1*	0.94±0.01*	**86.5**±2.09	83.5±1.2*	0.92±0.01*	87.4±2.4	76.3±0.9*	0.91±0.01*	76.6±4.1	77.3±2.1*	74.8±3.5

[1] The proposed method without the shallow spatial representation in Fig. 1a. Data are expressed as mean ± standard deviation. * indicates significant difference ($p < 0.05$) between each method and the proposed method.

sequence-to-sequence mapping and involve an intermediate hidden state, similar as the S4D block. The RNNs adopted in this article for ablation experiments include GRU [57], LSTM [58], bidirectional GRU (BGRU), and bidirectional LSTM (BLSTM) [58].

The results of the ablation experiments are reported in Table II. While removing the shallow spatial representation, the performance of the proposed method decreases in all metrics for the three classification tasks except for one, i.e., the sensitivity of the classification between NC and EMCI. These results demonstrate the contribution of shallow spatial pointwise convolution to the proposed deep learning framework. Besides, among all the models for temporal representation, the S4D method yields the highest values in all metrics, indicating the superiority of the S4D method for temporal representation of the BOLD signals.

IV. DISCUSSION

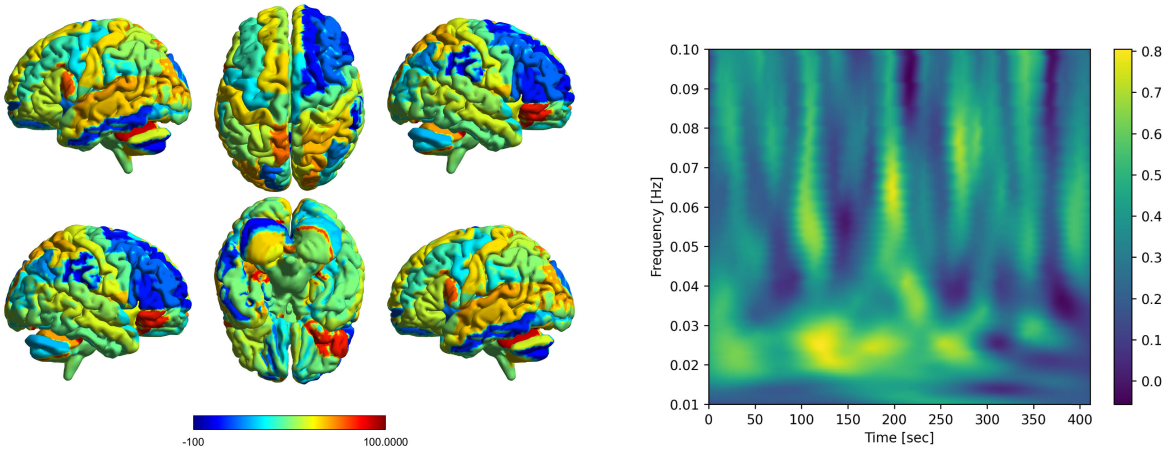
A. General Discussion

This article proposes an effective deep learning framework for improved MCI diagnosis. A key feature of the proposed framework is the employment of a diagonal-structured SSM for temporal representation of the BOLD signals. In fact, SSM has been widely used to model a time series as well as the fMRI and achieved promising results in the diagnosis of autism spectrum disorders and depression disorders [34]. In our study, we integrate the SSM in our deep learning framework for the first time for MCI detection, which lies in the main novelty of this article. Besides, a simple 1-D pointwise convolution is adopted for spatial representation, enabling fully free spatial characterization of the BOLD signals as well as the interpretability of the learned spatial features.

The observed superior results in Tables I and II demonstrate the effectiveness of the proposed deep learning framework. In particular, all metrics in Table II show better performance of S4D than other temporal representation methods, such as GRU [57], LSTM [58], BGRU, and BLSTM [58]. These results may suggest that long-term dependence of the fMRI may contribute significantly to MCI detection, similar to the diagnosis of autism spectrum disorders and depression disorders [34].

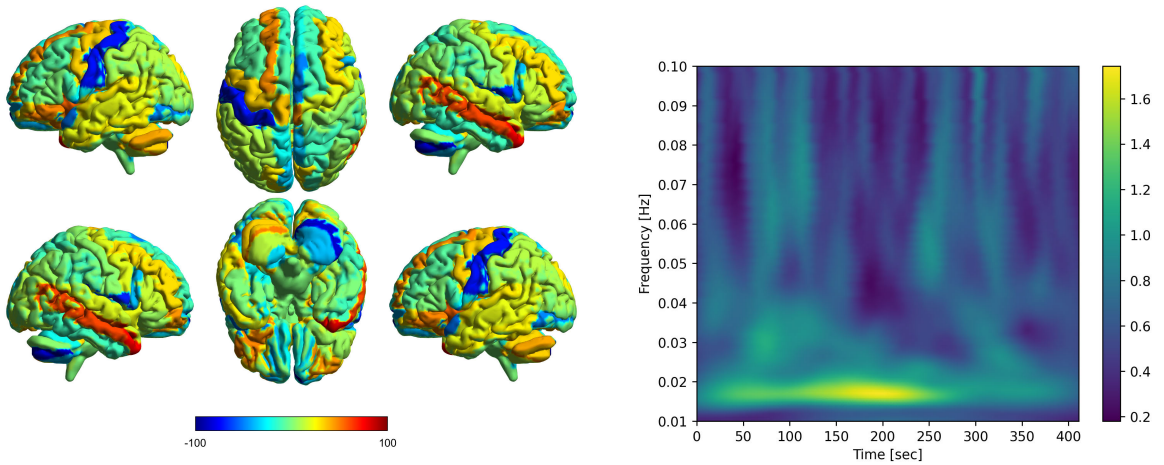
The ability of the S4D to capture long-term dependency depends on the structuration of the state-transition matrix to an HiPPO. Among several available HiPPO forms, a general framework referred to as HiPPO-FouT is adopted in this article, which performs a projection onto truncated Fourier functions [49]. For computational efficiency, the HiPPO matrix is usually decomposed into a DPLR matrix [37]. In this article, only a diagonal matrix is used as an approximation of the HiPPO matrix. The superior results for all the three classification tasks confirm our choice of HiPPO-FouT as well as the diagonal approximation.

Note that the sensitivities of the three-category classification are slightly lower than some of the SOTA methods. This is most probably due to the fact that, in this article, the model parameters are picked from previous works [34], [36], which are mainly optimized for accuracy rather than sensitivity. Besides, those studies focus only on binary classification [34], [36]. In order to improve the sensitivity of three-category classification, one strategy may be the use of nested cross validation, in which the model parameters can be optimized based on the sensitivity obtained on the validation set. Another possibility is to train three binary classifiers, with each distinguishing between one label and others. Then, the final



(a) (b)

Fig. 2. Visualization of the spatial filter for the classification between EMCI and NC. (a) Heat map of the spatial filter. (b) Difference in the averaged WT of the resulted features ($WT_{EMCI} - WT_{NC}$).



(a) (b)

Fig. 3. Visualization of the spatial filter for the classification between EMCI and LMCI. (a) Heat map of the spatial filter. (b) Difference in the averaged WT of the resulted features ($WT_{EMCI} - WT_{LMCI}$).

classification result of a test sample can be obtained using the voting of the three binary classifiers. In fact, our preliminary results show very good sensitivity for the binary classification between NC and MCI (EMCI + LMCI), i.e., ACC = 85.2%, AUC = 0.90, and SEN = 92.1%.

B. Feature Characterization

The shallow pointwise spatial convolution enables the visualization of the spatial filters as well as the generated feature maps. Specifically, the most significant spatial filter determined by ablation experiment, e.g., removal of which will result in the largest drop in the performance, is visualized. In addition, Morlet wavelet transform (WT) is applied to the feature map outputted from the determined spatial filter. The average WT over all subjects in each group, i.e., NC, EMCI, and LMCI, is calculated, and the difference in the average WT between two groups is illustrated.

Fig. 2 shows the heat map of the most significant spatial filter (a) and the difference in the average WT of the outputs of the spatial filter for the classification between NC and EMCI (b). As shown in Fig. 2(a), large weights of the spatial filter can be observed in several brain regions, such as regions, such as vermis, cerebellum, and the right frontal gyrus. Fig. 2(b) shows a clear power increase in the frequency band around 0.02–0.03 and 0.05–0.08 Hz in the outputs of the spatial filter for EMCI patients as compared to NC.

For the classification between LMCI and EMCI, large weights of the spatial filter are observed in right-middle frontal orbital gyrus, left postcentral gyrus, right-middle temporal pole, and right crus II of cerebellum, as shown in Fig. 3(a). Besides, a clearly higher power is observed in the frequency band around 0.01–0.02 Hz in the resulted feature maps of the EMCI patients as compared to that of the LMCI patients, as illustrated in Fig. 3(b).

For the three-category classification task among LMCI, EMCI, and NC, large weights of the spatial filter are observed in left supplementary motor area, left olfactory cortex, left cerebellum 7b, right insula, left superior occipital gyrus, left-middle temporal gyrus, left putamen, and vermis 6, as shown in Fig. 4(a). Moreover, the resulted feature maps of each

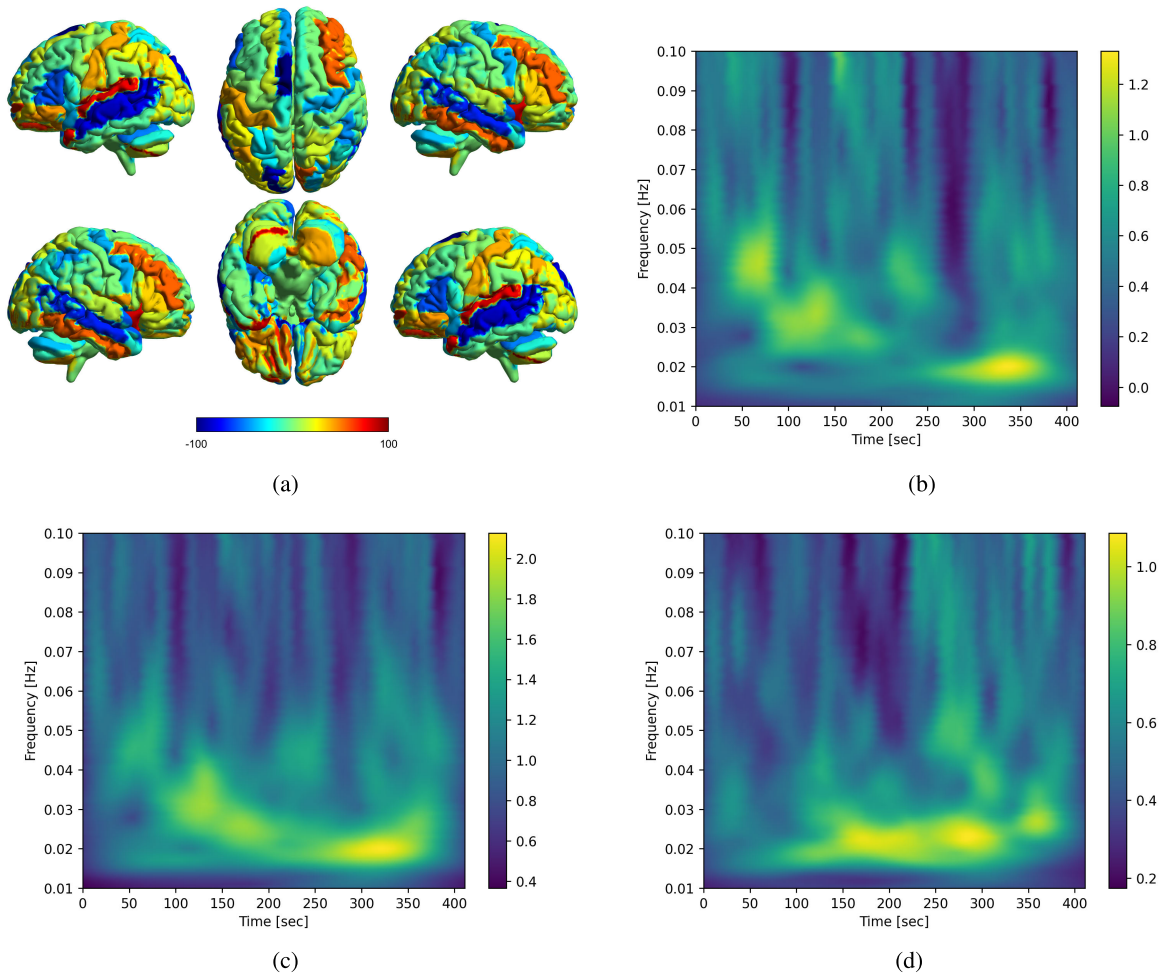


Fig. 4. Visualization of the spatial filter for the three-category classification. (a) Heat map of the spatial filter. (b) Difference in the averaged WT of the resulted features between NC and LMCI. (c) Difference in the averaged WT of the resulted features between NC and LMCI. (d) Difference in the averaged WT of the resulted features between EMCI and LMCI.

two classes show clear power difference in certain frequency bands, as illustrated in Fig. 4(b)–(d).

It is noted that the spatial patterns learned by our model are different from the traditional FC spatial patterns. This is most probably due to the fact that an FC extracts the temporal correlations between pairs of ROIs, while our pointwise spatial convolution performs a free linear combination of all brain regions.

C. Most Discriminative ROIs

To determine the most discriminative ROIs, we scaled the absolute weights between 0 and 1 for each spatial filter and ranked their importance according to the average scaled weights over the ten repeated tenfold cross validation. Consequently, the most ten discriminative ROIs for the three different classification tasks are reported in Table III. It is clear that a number of the ROIs identified from the three-category classification task overlap with that from the two binary classification tasks, including TPOmid.R, HIP.L, ITG.L, OLF.R, CAU.R, Cerebellum 6R, ROL.R, and Cerebellum 4&5L, confirming the importance of these ROIs for MCI diagnosis.

Note that many of the determined ROIs are located in occipital, frontal, and temporal regions, suggesting those regions to play an important role in MCI diagnosis, as reported in previous studies [59], [60]. Other ROIs, such as INS.L, DCG.R, OLF.R, ROL.R, left hippocampus, and CAU.R, are also important for tracking the progression of AD/MCI, in line with previous studies [29], [61], [62]. Besides, a few discriminative ROIs are identified in the cerebellum and vermis, indicating that the cerebellum and vermis also contain important pathology information for MCI diagnosis, in line with previous studies [31], [63]. These results suggest the

TABLE III
TEN MOST DISCRIMINATIVE ROIS FOR THE THREE CLASSIFICATION TASKS

NC vs. EMCI	EMCI vs. LMCI	NC vs. EMCI vs. LMCI
100 (Cerebellum 6R)	37 (HIP.L)	88 (TPOmid.R)
84 (TPOsup.R)	89 (ITG.L)	37 (HIP.L)
97 (Cerebellum 4&5L)	29 (INS.L)	89 (ITG.L)
111 (Vermis 4&5)	17 (ROL.L)	22 (OLF.R)
88 (TPOmid.R)	26 (ORBsupmed.R)	72 (CAU.R)
89 (ITG.L)	91 (Cerebellum Crus1L)	100 (Cerebellum 6R)
72 (CAU.R)	18 (ROL.R)	56 (FFG.R)
34 (DCG.R)	14 (IFGtriang.R)	32 (ACG.R)
22 (OLF.R)	88 (TPOmid.R)	18 (ROL.R)
87 (TPOmid.L)	97 (Cerebellum 4&5L)	97 (Cerebellum 4&5L)

spatial filters proposed in our model to reliably extract significant activity from pathologically related brain regions.

D. Limitations and Future Work

Worthily to note that, despite the effectiveness of the proposed pointwise spatial convolution, there are many other methods for spatial presentation. However, these methods are not investigated in this article due to the fact that the main focus of this article is the employment of the S4D model for temporal, particularly long-term, representation. Combining the S4D model and different spatial representation methods as spatial–temporal representation of the BOLD signals for MCI diagnosis may be an interesting direction for future studies.

Besides, it is noticed that, in this article, the BOLD signals consist of 116 channels (ROIs) derived from the AAL template. Using other templates with more subdivided ROIs can lead to an increase in the spatial resolution and may, therefore, have an impact on the performance of MCI diagnosis. Investigation of new templates with increased spatial resolution and its influence on our proposed deep learning framework for MCI detection may also be interesting directions in future studies.

Moreover, a forward information flow is considered in our S4D model. However, as a sequence-to-sequence mapping method, it is also possible to implement the S4D model with backward information flow. A bidirectional S4D may extract more significant temporal representation as similar bidirectional sequence to sequence methods, e.g., bidirectional GRU and bidirectional LSTM (BLSTM) [58]. Future studies may explore the effect of bidirectional S4D on temporal representation of the BOLD signals and, thus, on the diagnosis of MCI.

V. Conclusion

In this work, a novel deep learning neural network is proposed for effective MCI diagnosis, in which a simple pointwise convolution is employed for spatial representation and an S4D model for temporal representation. The evaluation results on the ADNI dataset with 521 samples show superior performance of our methods for EMCI and LMCI diagnosis, outperforming a number of SOTA methods. Moreover, visualization of the spatial filters and the corresponding outputs show the proposed method to be able to extract physiologically interpretable features. Future studies may focus on bidirectional S4D, a combination of S4D and different spatial representation methods, as well as improving the spatial resolution of the BOLD signals.

References

[1] K. Blennow, M. J. de Leon, and H. Zetterberg, "Alzheimer's disease," *Lancet*, vol. 368, no. 9533, pp. 387–403, 2006.

[2] R. C. Petersen et al., "Current concepts in mild cognitive impairment," *Arch. Neurol.*, vol. 58, no. 12, pp. 1985–1992, 2001.

[3] A. M. McGrattan et al., "Risk of conversion from mild cognitive impairment to dementia in low-and middle-income countries: A systematic review and meta-analysis," *Alzheimer's Dementia: Transl. Res. Clin. Interventions*, vol. 8, no. 1, 2022, Art. no. e12267.

[4] S. Saunders, C. W. Ritchie, T. C. Russ, G. Muniz-Terrera, and R. Milne, "Assessing and disclosing test results for 'mild cognitive impairment': The perspective of old age psychiatrists in Scotland," *BMC Geriatrics*, vol. 22, no. 1, pp. 1–12, Dec. 2022.

[5] Z. Chen et al., "A CRISPR/Cas12a-empowered surface plasmon resonance platform for rapid and specific diagnosis of the Omicron variant of SARS-CoV-2," *Nat. Sci. Rev.*, vol. 9, no. 8, 2022, Art. no. nwac104.

[6] F. Zheng et al., "A highly sensitive CRISPR-empowered surface plasmon resonance sensor for diagnosis of inherited diseases with femtomolar-level real-time quantification," *Adv. Sci.*, vol. 9, no. 14, May 2022, Art. no. 2105231.

[7] T. Xue et al., "Ultrasensitive detection of miRNA with an antimonene-based surface plasmon resonance sensor," *Nature Commun.*, vol. 10, no. 1, p. 28, Jan. 2019.

[8] S.-S. Poil, W. de Haan, W. M. van der Flier, H. D. Mansvelder, P. Scheltens, and K. Linkenkaer-Hansen, "Integrative EEG biomarkers predict progression to Alzheimer's disease at the MCI stage," *Frontiers Aging Neurosci.*, vol. 5, p. 58, Oct. 2013.

[9] D. Yang, K.-S. Hong, S.-H. Yoo, and C.-S. Kim, "Evaluation of neural degeneration biomarkers in the prefrontal cortex for early identification of patients with mild cognitive impairment: An fNIRS study," *Frontiers Hum. Neurosci.*, vol. 13, p. 317, Sep. 2019.

[10] S. H. Hojjati, A. Ebrahimzadeh, and A. Babajani-Feremi, "Identification of the early stage of Alzheimer's disease using structural MRI and resting-state fMRI," *Front. Neurol.*, vol. 10, p. 904, Aug. 2019.

[11] M. Rubinov and O. Sporns, "Complex network measures of brain connectivity: Uses and interpretations," *NeuroImage*, vol. 52, no. 3, pp. 1059–1069, Sep. 2010.

[12] R. M. Hutchison et al., "Dynamic functional connectivity: Promise, issues, and interpretations," *Neuroimage*, vol. 80, pp. 360–378, Oct. 2013.

[13] L. Qiao, H. Zhang, M. Kim, S. Teng, L. Zhang, and D. Shen, "Estimating functional brain networks by incorporating a modularity prior," *NeuroImage*, vol. 141, pp. 399–407, Nov. 2016.

[14] T. Cao, R. Lin, Y. Zheng, D. Shen, and L. Xu, "A novel approach analysing the dynamic brain functional connectivity for improved MCI detection," *IEEE Trans. Biomed. Eng.*, vol. 71, no. 1, pp. 207–216, Jan. 2024.

[15] X. Chen, H. Zhang, L. Zhang, C. Shen, S. Lee, and D. Shen, "Extraction of dynamic functional connectivity from brain grey matter and white matter for MCI classification," *Hum. Brain Mapping*, vol. 38, no. 10, pp. 5019–5034, Oct. 2017.

[16] R. Yu, H. Zhang, L. An, X. Chen, Z. Wei, and D. Shen, "Connectivity strength-weighted sparse group representation-based brain network construction for M CI classification," *Hum. Brain Mapping*, vol. 38, no. 5, pp. 2370–2383, May 2017.

[17] Y. Zhang et al., "Strength and similarity guided group-level brain functional network construction for MCI diagnosis," *Pattern Recognit.*, vol. 88, pp. 421–430, Apr. 2019.

[18] C.-Y. Wee, S. Yang, P.-T. Yap, D. Shen, and A. D. N. Initiative, "Sparse temporally dynamic resting-state functional connectivity networks for early MCI identification," *Brain Imag. Behav.*, vol. 10, pp. 342–356, 2016.

[19] W. Li, Z. Wang, L. Zhang, L. Qiao, and D. Shen, "Remodeling Pearson's correlation for functional brain network estimation and autism spectrum disorder identification," *Frontiers Neuroinform.*, vol. 11, p. 55, Aug. 2017.

[20] Y. Wang, J. Kang, P. B. Kemmer, and Y. Guo, "An efficient and reliable statistical method for estimating functional connectivity in large scale brain networks using partial correlation," *Frontiers Neurosci.*, vol. 10, p. 123, Mar. 2016.

[21] S. M. Bowyer, "Coherence a measure of the brain networks: Past and present," *Neuropsychiatric Electrophysiol.*, vol. 2, pp. 1–12, Jan. 2016.

[22] B. T. Schmidt, A. S. Ghuman, and T. J. Huppert, "Whole brain functional connectivity using phase locking measures of resting state magnetoencephalography," *Frontiers Neurosci.*, vol. 8, p. 141, Jun. 2014.

[23] Z. Zhou et al., "A toolbox for brain network construction and classification (BrainNetClass)," *Hum. Brain Mapping*, vol. 41, no. 10, pp. 2808–2826, Jul. 2020.

[24] Y. Xue, Y. Zhang, L. Zhang, S.-W. Lee, L. Qiao, and D. Shen, "Learning brain functional networks with latent temporal dependency for MCI identification," *IEEE Trans. Biomed. Eng.*, vol. 69, no. 2, pp. 590–601, Feb. 2022.

[25] J. Huang, L. Zhou, L. Wang, and D. Zhang, "Attention-diffusion-bilinear neural network for brain network analysis," *IEEE Trans. Med. Imag.*, vol. 39, no. 7, pp. 2541–2552, Jul. 2020.

[26] S. Gadgil, Q. Zhao, A. Pfefferbaum, E. V. Sullivan, E. Adeli, and K. M. Pohl, "Spatio-temporal graph convolution for resting-state fMRI analysis," in *Proc. 23rd Int. Conf. Med. Image Comput. Comput. Assist. Intervent. (MICCAI)*. Lima, Peru: Springer, 2020, pp. 528–538.

[27] Y. Li, J. Liu, Y. Jiang, Y. Liu, and B. Lei, "Virtual adversarial training-based deep feature aggregation network from dynamic effective connectivity for MCI identification," *IEEE Trans. Med. Imag.*, vol. 41, no. 1, pp. 237–251, Jan. 2022.

[28] W. Cui et al., "Personalized functional connectivity based spatio-temporal aggregated attention network for MCI identification," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 31, pp. 2257–2267, 2023.

[29] X. Xing et al., "DS-GCNs: Connectome classification using dynamic spectral graph convolution networks with assistant task training," *Cerebral Cortex*, vol. 31, no. 2, pp. 1259–1269, Jan. 2021.

[30] T.-E. Kam, H. Zhang, Z. Jiao, and D. Shen, "Deep learning of static and dynamic brain functional networks for early MCI detection," *IEEE Trans. Med. Imag.*, vol. 39, no. 2, pp. 478–487, Feb. 2020.

[31] M. Wang, C. Lian, D. Yao, D. Zhang, M. Liu, and D. Shen, "Spatial-temporal dependency modeling and network hub detection for functional MRI analysis via convolutional-recurrent network," *IEEE Trans. Biomed. Eng.*, vol. 67, no. 8, pp. 2241–2252, Aug. 2020.

[32] A. El Gazzar, L. Cerliani, G. van Wingen, and R. M. Thomas, "Simple 1-D convolutional networks for resting-state fMRI based classification in autism," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2019, pp. 1–6.

[33] N. C. Dvornek, P. Ventola, K. A. Pelphrey, and J. S. Duncan, "Identifying autism from resting-state fMRI using long short-term memory networks," in *Proc. Mach. Learn. Med. Imag., 8th Int. Workshop, MLMI, Held Conjunct (MICCAI)*. Quebec City, QC, Canada: Springer, 2017, pp. 362–370.

[34] A. El-Gazzar, R. M. Thomas, and G. Van Wingen, "fMRI-S4: Learning short-and long-range dynamic fMRI dependencies using 1D convolutions and state space models," in *Proc. Int. Workshop Mach. Learn. Clin. Neuroimag.* Springer, 2022, pp. 158–168.

[35] A. Gu, T. Dao, S. Ermon, A. Rudra, and C. Ré, "HiPPO: Recurrent memory with optimal polynomial projections," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 1474–1487.

[36] A. Gu, K. Goel, A. Gupta, and C. Ré, "On the parameterization and initialization of diagonal state space models," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 35971–35983.

[37] A. Gu, K. Goel, and C. Ré, "Efficiently modeling long sequences with structured state spaces," 2021, *arXiv:2111.00396*.

[38] J. G. Sled, A. P. Zijdenbos, and A. C. Evans, "A nonparametric method for automatic correction of intensity nonuniformity in MRI data," *IEEE Trans. Med. Imag.*, vol. 17, pp. 87–97, 1998.

[39] N. Tzourio-Mazoyer et al., "Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain," *NeuroImage*, vol. 15, no. 1, pp. 273–289, 2002.

[40] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[41] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826.

[42] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 4700–4708.

[43] V. Lawhern, A. Solon, N. Waytowich, S. M. Gordon, C. Hung, and B. J. Lance, "EEGNet: A compact convolutional neural network for EEG-based brain–computer interfaces," *J. Neural Eng.*, vol. 15, no. 5, 2018, Art. no. 056013.

[44] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.

[45] R. L. Williams et al., *Linear State-Space Control Systems*. Hoboken, NJ, USA: Wiley, 2007.

[46] A. Tustin, "A method of analysing the behaviour of linear systems in terms of time series," *J. Inst. Elect. Eng.-IIA: Autom. Regulators Servo Mech.*, vol. 94, no. 1, pp. 130–142, 1947.

[47] R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 1310–1318.

[48] A. Gu et al., "Combining recurrent, convolutional, and continuous-time models with linear state space layers," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 572–585.

[49] A. Gu, I. Johnson, A. Timalsina, A. Rudra, and C. Ré, "How to train your HiPPO: State space models with generalized orthogonal basis projections," 2022, *arXiv:2206.12037*.

[50] D. Hendrycks and K. Gimpel, "Gaussian error linear units (GELUs)," 2016, *arXiv:1606.08415*.

[51] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, "Language modeling with gated convolutional networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 933–941.

[52] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *Proc. 14th Eur. Conf.* Amsterdam, The Netherlands: Springer, 2016, pp. 630–645.

[53] J. Lei Ba, J. Ryan Kiros, and G. E. Hinton, "Layer normalization," 2016, *arXiv:1607.06450*.

[54] M. Lin, Q. Chen, and S. Yan, "Network in network," 2013, *arXiv:1312.4400*.

[55] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2017, *arXiv:1711.05101*.

[56] I. Loshchilov and F. Hutter, "SGDR: Stochastic gradient descent with warm restarts," 2016, *arXiv:1608.03983*.

[57] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," 2014, *arXiv:1412.3555*.

[58] R. C. Staudemeyer and E. Rothstein Morris, "Understanding LSTM—A tutorial into long short-term memory recurrent neural networks," 2019, *arXiv:1909.09586*.

[59] L. Sun et al., "Mining brain region connectivity for Alzheimer's disease study via sparse inverse covariance estimation," in *Proc. 15th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2009, pp. 1335–1344.

[60] T. Zhang et al., "Predicting MCI to AD conversation using integrated sMRI and rs-fMRI: Machine learning and graph theory approach," *Frontiers Aging Neurosci.*, vol. 13, Jul. 2021, Art. no. 688926.

[61] A. Khazaee, A. Ebrahimzadeh, and A. Babajani-Feremi, "Identifying patients with Alzheimer's disease using resting-state fMRI and graph theory," *Clin. Neurophysiol.*, vol. 126, no. 11, pp. 2132–2141, Nov. 2015.

[62] H.-I. Suk, S.-W. Lee, and D. Shen, "Latent feature representation with stacked auto-encoder for AD/MCI diagnosis," *Brain Struct. Function*, vol. 220, pp. 841–859, Dec. 2015.

[63] Z. Wang et al., "Spatial patterns of intrinsic brain activity in mild cognitive impairment and Alzheimer's disease: A resting-state functional MRI study," *Hum. Brain Mapping*, vol. 32, no. 10, pp. 1720–1740, Oct. 2011.