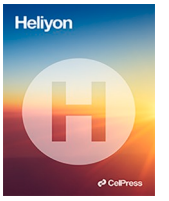
Heliyon 11 (2025) e41120

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon



Research article

Prob-sparse self-attention extraction of time-aligned dynamic functional connectivity for ASD diagnosis

Hongwu Chen [a], Fan Feng [b], Pengwei Lou [c,d], Ying Li [b], MingLi Zhang [b], Feng Zhao [b,c,d,*]

[a] School Hospital, Shandong Technology and Business University, Yantai, China
[b] School of Computer Science and Technology, Shandong Technology and Business University, Yantai, China
[c] Key Laboratory of Xinjiang Coal Resources Green Mining, Ministry of Education, Xinjiang, China
[d] College of Information Engineering, Xinjiang Institute of Engineering, Xinjiang, China

ARTICLE INFO

Keywords:
Autism spectrum disorder
Prob-sparse attention
Sliding window
Time alignment

ABSTRACT

Dynamic functional connectivity (DFC) has shown promise in the diagnosis of Autism Spectrum Disorder (ASD). However, extracting highly discriminative information from the complex DFC matrix remains a challenging task. In this paper, we propose an ASD classification framework PSA-FCN which is based on time-aligned DFC and Prob-Sparse Self-Attention to address this problem. Specifically, we introduce Prob-Sparse Self-Attention to selectively extract global features, and use self-attention distillation as a transition at each layer to capture local patterns and reduce dimensionality. Additionally, we construct a time-aligned DFC matrix to mitigate the time sensitivity of DFC and extend the dataset, thereby alleviating model overfitting. Our model is evaluated on fMRI data from the ABIDE NYU site, and the experimental results demonstrate that the model outperforms other methods in the paper with a classification accuracy of 81.8 %. Additionally, our research findings reveal significant variability in the DFC connections of brain regions of ASD patients, including Cuneus (CUN), Lingual gyrus (LING), Superior occipital gyrus (SOG), Posterior cingulate gyrus (PCG), and Precuneus (PCUN), which is consistent with prior research. In summary, our proposed PSA framework shows potential in ASD diagnosis as well as automatic discovery of critical ASD-related biomarkers.

1. Introduction

Autism Spectrum Disorder (ASD) is a complex neurodevelopmental condition characterized by persistent challenges in social interaction, communication, and repetitive behaviors [1,2]. Individuals with ASD face a range of difficulties that have far-reaching impacts on their personal lives, families, and society as a whole. However, diagnosing ASD is challenging, as it requires careful observation and evaluation by specialized medical professionals. The challenges in diagnosing ASD can lead to delays in patients accessing necessary interventions and treatment. Therefore, early and accurate detection of ASD is of paramount importance.

In recent years, significant advancements in computer science and neuroimaging technology have greatly pushed forward the development of innovative diagnostic methods for ASD [3]. Among these methods, static functional connectivity (SFC) based on functional magnetic resonance imaging (fMRI) has proved to be an effective diagnostic tool for a variety of neurological, cognitive, and

* Corresponding author. School of Computer Science and Technology, Shandong Technology and Business University, Yantai, China.
E-mail address: zhaofeng1016@126.com (F. Zhao).

https://doi.org/10.1016/j.heliyon.2024.e41120
Received 15 June 2024; Received in revised form 7 December 2024; Accepted 9 December 2024
Available online 10 December 2024

psychiatric disorders including ASD, Alzheimer's disease, and others [4,5]. SFC analysis holds the assumption that the connectivity patterns within the brain remain relatively stable throughout the entire scanning process [6], thus involving the analysis of averaged data across the entire scan. However, increasing evidence suggests that the functional connectivity of brain revealed by rs-fMRI data is dynamic instead of static [7]. Consequently, SFC methods cannot fully capture the brain's changes at varying time scales [8] and may miss valuable information for diagnosis.

With the recognition that brain FC is highly dynamic, researchers have paid more attention to dynamic functional connectivity (DFC) analysis as a promising avenue for gaining deeper insights into neurological conditions like ASD. For example, Li et al. used sliding window to construct DFC and used the k-means algorithm to perform a cluster analysis on an individual consisting of 154 windows * 4005 features, and found that DFC variability was higher between the posterior cingulate gyrus (PCC) and middle temporal pole (TPOmid) in ASD subjects [9]. Zhao et al. constructed a higher-order D-FCN net to explore higher-level and more complex interactions between multiple pairs of brain regions and used a central-moment approach to extract the temporal invariance of the data [10]. Their experimental results reveal that the brain regions selected for modeling were related to emotional expression, language comprehension, and motor coordination. Some deep learning methods have also been applied to DFC classification. For example, Fan et al. combined convolutional neural networks (CNN) and long short-term memory (LSTM) networks to extract the spatiotemporal features of DFC [11].

Despite the promising results of the DFC analysis, the existing methods still have some limitations: 1) Most DFC methods lack self-attention mechanisms. Existing DFC construction methods often suffer from high dimensionality and the introduction of redundant connections, leading to significant interference in feature extraction. self-attention mechanism can selectively capture global relationships within the data, allowing the model to focus more on disease-relevant connections, thus contributing to a more nuanced understanding of the dynamic interactions of the brain over time. 2) DFC data typically possess high dimensions, and conventional dimension reduction methods used may fail to learn the intricate temporal patterns in specific time windows. Ultimately, this hinders the ability to provide a detailed description of brain activity. 3). The continuous time scanning nature of fMRI requires strict chronological arrangement of subnetworks generated by sliding windows. This makes DFC highly sensitive to the temporal sequence of these subnetworks. Consequently, establishing temporal correspondences among FC subnetworks from the same time window across different subjects becomes challenging, thereby impeding research and comparisons of DFC at a population level [10].

To address the aforementioned challenges, we propose an ASD classification model based on time-aligned DFC and Prob-Sparse self-attention named PSA-FCN. *Firstly,* we introduce Prob-Sparse attention for feature extraction. It is an enhanced version of self-attention proposed by Wang et al. [12] aimed at reducing the computational complexity in traditional self-attention models. Through the 'Top-K' selection strategy, computations and updates are focused solely on the positions with the highest attention scores. This approach preserves the extraction of global information while eliminating some of the redundant connections in DFC data, enabling selective extraction of spatiotemporal features. *Afterwards,* we use self-attention distillation as a transition layer between the feature extraction modules. The included convolutional layers enhance the learning process of local features, while the pooling layers effectively reduce the dimensionality of DFC data, resulting in a pyramid-like structure within the model. *Additionally*, in order to minimize the time sensitivity of the data, we design an innovative cyclic DFC matrix with head and tail connected. This structural design liberates the DFC sequence from the restriction of commencing at a single time point. Moreover, it augments the training dataset and mitigates the overfitting problem which is typically caused by limited samples. All the source code of PSA-FCN is available at https://github.com/fan2077/PSA-FCN-InformerDFC.

The contributions of this paper can be summarized as follows: 1). We introduce the Prob-Sparse self-attention mechanism to address issues related to DFC. The enhanced self-attention is able to selectively extract important features, allowing the model to capture the connectivity of critical brain regions with high computational efficiency. 2). The utilization of self-attention distillation enables the model to learn global relationships while further capturing local features. It can also reduce data dimensions and enhance model generalization. 3). We take a new perspective on high-dimensional functional connectivity matrix sequences, creatively constructing cyclic DFC matrices, which reduce the time sensitivity of DFC matrices, and alleviate the overfitting problem by data augmentation.

Table 1
The demographic details of the subjects.

Characteristic	ASD	NC	p-values
Gender (M/F)	36/9	36/11	0.6923[a]
Age (mean ± SD)	11.1 ± 2.3	11.0 ± 2.3	0.773[b]
FIQ (mean ± SD)	106.8 ± 17.4	113.3 ± 14.1	0.0510[b]
ADI-R (mean ± SD)	32.2 ± 14.3[c]	–	–
ADOS (mean ± SD)	13.7 ± 5.0	–	–
FD (mean ± SD)	0.14 ± 0.05	0.15 ± 0.07	0.36[b]

M, male; F, female; FIQ, Full Intelligence Quotient; ADI-R, Autism Diagnostic Interview-Revised; ADOS, autism diagnostic observation schedule.

[a] The p-value was obtained by $\chi^2 - test$.

[b] The p-value was obtained by two-sample two-tailed *t*-test.

[c] Two patients do not have the ADI-R score.

2. Materials and preprocessing

The data used in this study is sourced from the Autism Brain Imaging Data Exchange (ABIDE) database, which comprises 17 imaging sites worldwide [13,14]. To address the data heterogeneity arising from different MRI acquisition protocols (such as variations in MRI equipment, scan times, and basic pulse sequence parameters) and uncontrolled preprocessing procedures across sites, we chose to evaluate our model using data from the NYU site. This dataset offers the largest sample size for testing the feasibility of our proposed method. It includes rs-fMRI scans from 45 ASD patients and 47 healthy controls, aged 7 to 15, with no excessive head movement in any direction (displacement less than 1.5 mm or angular rotation less than 1.5°). There were no significant differences between the two groups in terms of gender, age, or full-scale IQ (FIQ) (see Table 1).

The data acquisition and preprocessing follow a standard pipeline, including head movement correction, normalization, denoising, and other processes. Specifically, firstly, the initial 10 rs-fMRI volumes were discarded to achieve magnetization stabilization. Spatial smoothing was then applied using a Gaussian kernel with a full width at half maximum (FWHM) of 6 mm, additional processing steps involved band-pass filtering (0.005–0.1 Hz) and signal detrending to alleviate physiological noise, measurement inaccuracies, and potential scanner magnetic field drifts. Next, nuisance signals including ventricular and global signals were regressed out, Subsequent to this, all rs-fMRI volumes underwent normalization to the Montreal Neurological Institute (MNI) space, adhering to a resolution of 3 $\times$ 3 $\times$ 3 mm3. To mitigate the effects of head motion, six head motion signals were regressed prior to computing functional connectivity [15–18]. The brain was segmented into 116 regions of interest (ROIs) defined in the Automated Anatomical Labeling (AAL) atlas. At the final step, the average Rs-fMRI time series for each of the 116 brain ROIs were computed, resulting in a data matrix $X \in R^{170\times116}$, wherein 170 represents the total number of temporal image volumes and 116 denotes the total number of brain ROIs.

3. Method

The overall structure of PSA-FCN is illustrated in Fig. 1. The model takes the DFC matrix as input, and its feature extraction component (encoder) consists of stacked attention blocks. After each round of attention computation, we introduce a self-attention distillation module, in which the local information of the data is captured by convolutional layers and the dimensionality of the matrix is reduced by half with a pooling layer. This "feature extraction-distillation" process can be repeated multiple times, creating a pyramid-like architecture for the model. The processed data is then fed into the final classification layer, where it is mapped to the classification head through a linear projection layer. In this section, we will first introduce the construction of the DFC matrix and how time alignment is applied, and then elaborate the Prob-Sparse self-attention introduced in the model and the composition of the encoder.

3.1. Construction of DFC matrix

We construct the DFC matrix in a similar manner as Ji et al. [19], as shown in Fig. 2. The fMRI time series can be represented by a matrix of size N^*L, where N is the number of brain regions and L represents the number of time points in the scan (Fig. 2(a)). Next, the time series is locally sampled by using a sliding window (with width W, and step S). For each window, we calculate the Pearson correlation, resulting in i individual matrices of size N^*N (Fig. 2(b)), where i is determined by the sliding window formula: $i = [(L - W)/S] + 1$.

Since the correlation matrices are symmetric square matrices, we extract half of each matrix along the diagonal and flatten it row by row into one-dimensional vectors. Each vector has a length of $l = N^*(N-1)/2$ (excluding the diagonal elements). Subsequently, these vectors are stacked in the order generated by the sliding window, resulting in a matrix with a length of l and a width of i. In this matrix, each row represents the functional connectivity changes of different brain regions at the same time point, while each column represents the temporal changes of the same brain region. This matrix, which combines the spatiotemporal features of brain regions, is defined as a DFC matrix.

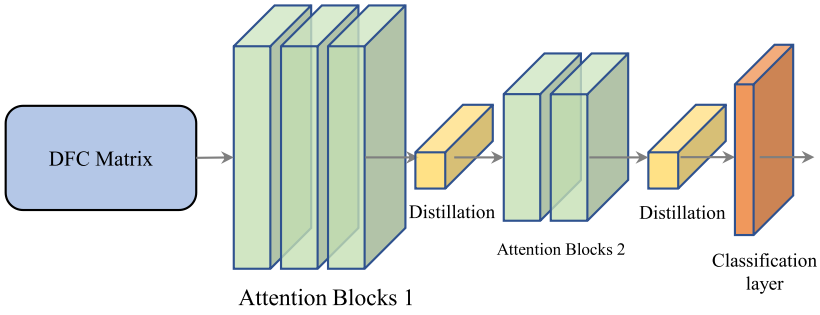


Fig. 1. The overall structure of the model. The model takes a DFC matrix as input, where the green module represents the feature extraction section, composed of stacked Prob-attention encoders. The yellow module signifies the self-attention distillation layer, which, after distillation, reduces the data's dimension by half, resulting in a pyramid-shaped model structure.

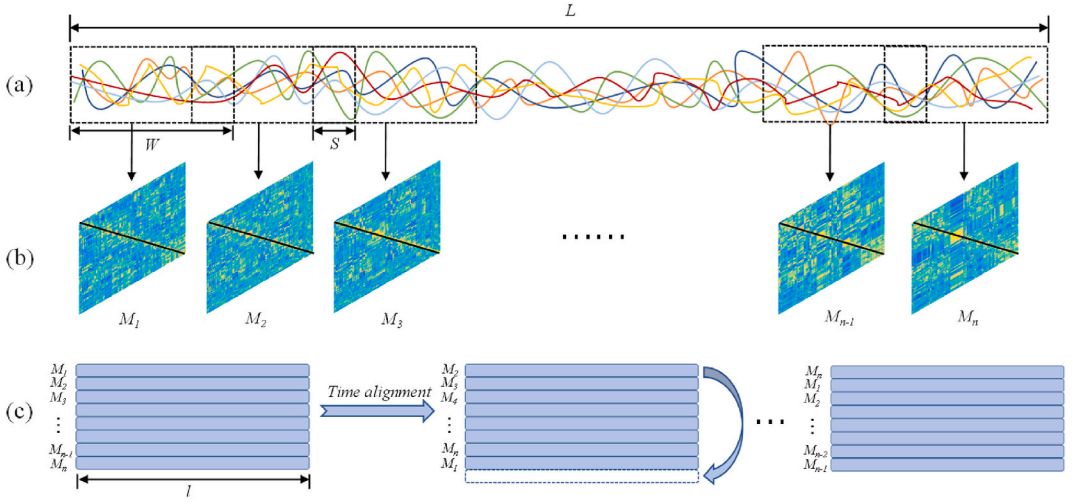


Fig. 2. Construction process of DFC matrix. Where Fig. 2(a) represents the process of sliding window cutting of the original time series, Fig. 2(b) represents the functional connectivity matrix obtained by doing Pearson's correlation for each sliding window and halving it along the diagonal, and Fig. 2(c) demonstrates the constructed DFC matrix, as well as the process of time alignment.

3.2. Time alignment

After constructing the DFC matrix, we introduce a "time alignment" method to address its sensitivity to time order. Specifically, motivated by the concept of cyclic linked list, we connect the first and last rows of the DFC matrix, which allows the matrix to commence from any selected time point and effectively reorder the functional connectivity models at different time points. This method facilitates a more thorough exploration of dynamic changes in functional connectivity and the discovery of potential consistent patterns. This process is illustrated in Fig. 2(c).

In practice, we repeatedly shift one or more rows of the DFC matrix to the end, creating a new data matrix. Our method has two significant advantages: 1). The newly generated matrix not only maintains overall temporal continuity but also extensively explores different starting points in the DFC sequence, enabling a deeper investigation into the dynamic changes in functional connectivity patterns at various time points. 2). The new matrix significantly expands the training dataset, which can effectively avoid overfitting in deep learning with limited training samples. This ensures that the model exhibits robustness and good performance even when dealing with small datasets.

3.3. Prob-sparse self-attention

During the DFC construction process, flattening all pairs of brain regions results in a huge dimensionality of the DFC matrix, which poses a big challenge for feature extraction and model training. Therefore, we introduce Prob-Sparse self-attention into our model. To better understand Prob-Sparse self-attention, we first revisit the Norm-based self-attention mechanism in Ref. [20]. The traditional formula for scaled dot product self-attention is defined as (1):

$$A(Q,K,V) = \text{Softmax}\left(QK^T \Big/ \sqrt{d}\right)V. \tag{1}$$

Where QKV is the query and key-value pair generated from the data, and d denotes the dimension of input. To further elaborate, let q_i, k_i, v_i represent the $i-th$ row in Q, K, V respectively. For the $i-th$ query, its attention score with the $j-th$ key can be defined as a probability p (2):

$$p(k_j|q_i) = \frac{e^{q_i k_j^T} \Big/ \sqrt{d}}{\sum_{l=1}^{L} e^{q_i k_l^T} \Big/ \sqrt{d}}. \tag{2}$$

Accordingly, we can derive the self-attention of q_i on the key K, expressed as $A(q_i,K,V) = \sum_j p(k_j|q_i)v_j$. It is defined as the combination of the attention of the $i-th$ query on all keys with the values V. However, not all the self-attentions generated by queries are equally important. Many studies have revealed that self-attention probability distributions have inherent sparsity. For example, Wang et al. qualitatively evaluated the regularized self-attention patterns and found that self-attention score follows a long-tail distribution, where a few dot products contribute to the main attention while others contribute to negligible attention [12]. Therefore, we aim to select self-attention probability distributions $p(k_j|q_i)$ that are far from the uniform distribution $q(k_j|q_i) = 1/L$ for the query q_i. To this end, we use the KL divergence to quantify the difference between two distributions, and the magnitude of its value reflects the

importance of the query, expressed as (3):

$$KL(q\|p) = ln\sum_{j=1}^{L} e^{\frac{q_i k_j^T}{\sqrt{d}}} - \frac{1}{L}\sum_{j=1}^{L}\frac{q_i k_j^T}{\sqrt{d}} - ln\, L. \tag{3}$$

By removing the constant term, we define the sparsity measure of the $i-th$ query as (4):

$$M(q_i, K) = ln\sum_{j=1}^{L_K} e^{\frac{q_i k_j^T}{\sqrt{d}}} - \frac{1}{L}\sum_{j=1}^{L_K}\frac{q_i k_j^T}{\sqrt{d}}. \tag{4}$$

Queries with larger values of $M(q_i, K)$ are considered to have a greater influence when computing attention, allowing them to capture important contextual information. By constructing sparse queries, the model can achieve a balance between information representation and computational efficiency. However, there is a need for iteration through all query & key to compute correlations. To further reduce computation, Prob-Sparse self-attention adopts an efficient sampling strategy, i.e., randomly selecting a subset of Keys to compute correlations with each q. The top-u queries are selected for subsequent computations. The formula is as follows (5):

$$\overline{M}(q_i, \widetilde{K}) = \max_j\left\{\frac{q_i k_j^T}{\sqrt{d}}\right\} - \frac{1}{\widetilde{L}}\sum_{j=1}^{\widetilde{L}}\frac{q_i k_j^T}{\sqrt{d}}. \tag{5}$$

Where $\widetilde{K}$ represents the randomly sampled Keys, $\widetilde{L}$ represents the number of selected Keys, which is calculated as $\widetilde{L} = L\ln^L$. After computing the dot product between each query and the sampled Keys, only the query corresponding to the Top-u $\overline{M}(q_i, \widetilde{K})$ are used for subsequent self-attention computations. The other dot products can be omitted due to the filled zeros.

3.4. Encoder

The Prob-attention block structure in our model is depicted in Fig. 3(a). We leverage the model from our previous work, but replace the multi-head attention mechanism with Prob-attention. Within the encoder, the Add & Norm layer, which includes residual connections and layer normalization, helps to mitigate the vanishing gradient problem. Additionally, dropout regularization is applied to reduce the risk of overfitting, and the feedforward layer handles the nonlinear transformation of the output. It is worth noting that we also explore the encoder design problem raised in Ref. [12], but our experimental results demonstrate that the encoder we designed is better suited for the current classification task.

3.5. Self-attention distilling

As the encoder layers maintain the input matrix size, dealing with such a large matrix remains a challenge in constructing subsequent deep neural networks. To address this problem, additional hidden layers, called "self-attention distillation", are integrated into our architecture, depicted in Fig. 3(b). This particular module includes a one-dimensional convolutional (Conv1D) layer, which can exploit local data features to complement the global information captured by the self-attention mechanism. Batch Normalization after Conv1D layer is used to stabilize and accelerate the training process. Subsequently, Exponential Linear Units (ELUs) serve as activation

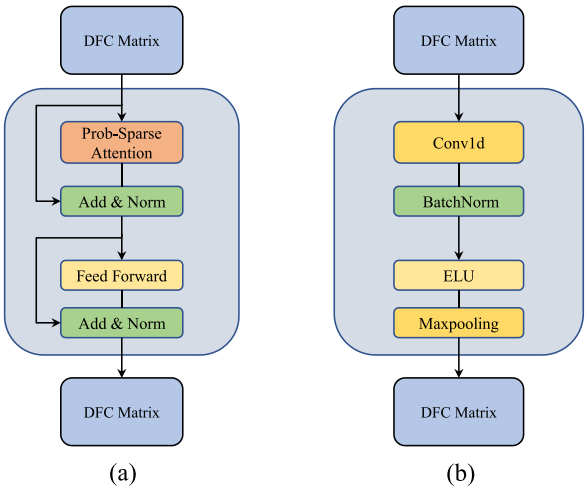


Fig. 3. The overall structure of the model encoder. Fig. 3(a) presents the Prob-attention module in the encoder and Fig. 3(b) presents the self-attention distillation module.

functions to introduce nonlinearities and capture complex relationships in the data. Finally, the maximum pooling layer can effectively reduce matrix dimensionality by aggregating local features, while maintaining essential information for classification.

4. Experiment

We conducted three experiments: 1). The influence of sliding window hyperparameters. In this experiment, we test various combinations of window width and stride to determine the optimal window configuration's impact on the model's performance. 2). Ablation study. This series of experiments systematically involved removing certain components from the model architecture to analyze their individual contributions to overall performance. 3). ASD diagnosis using different methods. This experiment involved comparing the proposed model with existing benchmark models or other excellent methods on the task of ASD vs. NC classification. All experiments were conducted on a workstation equipped with Nvidia Geforce RTX 2060 GPU.

4.1. The influence of sliding window hyperparameters

When constructing the DFC matrix, the sliding window hyperparameters can significantly influence the accuracy and stability of the training results. Specifically, the window width and step size are two crucial factors that determine the temporal resolution of the DFC matrix. The width of sliding window determines the number of time segments included in each window, while the step size defines the time interval between adjacent windows. The selection of these two hyperparameters directly affects the level of detail in the connectivity patterns and the ability to capture dynamic changes in the DFC matrix.

Balancing the trade-off between temporal resolution and computational efficiency is a common challenge in choosing these parameters. Smaller window widths and steps can provide higher temporal resolution and better capture short-term dynamic connectivity patterns. However, smaller window size also increases the computational complexity and storage requirements. On the contrary, larger window widths and steps can reduce computation but may lead to an overly smoothed DFC matrix that fails to capture instantaneous functional connectivity changes.

This section aims to systematically explore the impact of sliding window hyperparameters on the DFC matrix and find the optimal values for these parameters. By employing a specific hyperparameter tuning method, we conduct comprehensive experimental evaluations of different combinations of window widths and steps.

Specifically, we select window widths W = [30, 35, 40, 45, 50, 55, 60] and segment them in increments of step size S = [2–6]. The results are presented in Fig. 4. It is evident the classification accuracy of the models is worst when the step size is 2 regardless of window widths. As the step size increases, the model exhibits improved accuracy, and reaches the highest accuracy when the step size is set to 5 or 6.

For further validation, we conduct an additional experiment. The average cosine similarity between consecutive matrices resulting from segmentation is computed and visualized as a heatmap, as shown in Fig. 5. It can be seen that when the step size is set as 1 or 2, the similarity between matrices remains relatively high, ranging from approximately 94 %–97 %. However, an obvious decline in similarity occurs at a step size of 3, with a slight increase observed at a step size of 6. This explains the slight decrease in accuracy observed at a step size of 6. Importantly, the similarity does not continuously decrease with step size increments but stabilizes at around 90 %. Even when the step size equals the window width, the similarity remains at approximately 89 %.

Therefore, we can draw the following conclusions: 1). When the step size is small, rapid convergence is achieved due to significant overlap between adjacent windows, but this leads to lower accuracy on the validation set and the presence of significant overfitting. 2). Data similarity does not exhibit a continuous decline with an increase in step size; it stabilizes at approximately 90 % even when the step size equals the window width. Based on these findings, we select a segmentation mode with a step size of 5 for sliding window analysis. This can help capture the key features while avoiding overfitting.

4.2. Ablation study

We conduct an ablation study for PSA-FCN. In these experiments, the setup is the same as the above ones, and the dataset is

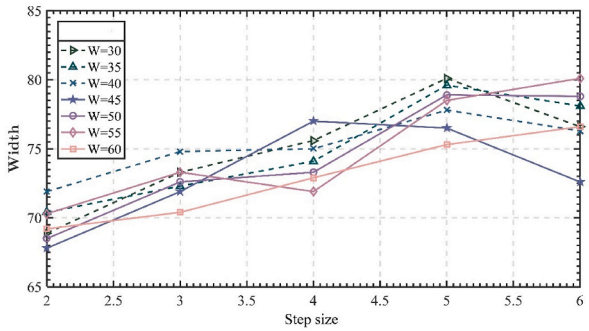


Fig. 4. Hyperparameter study of sliding window.

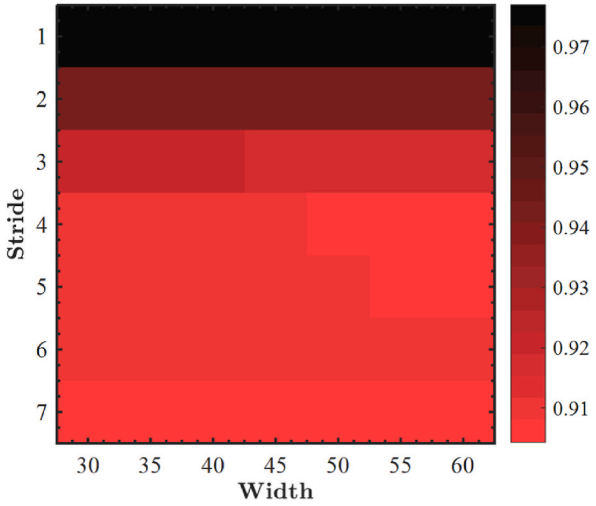


Fig. 5. Heatmap of similarity of neighboring matrices.

segmented using a sliding window with W = 30 and S = 5. Table 2 presents the results of all the ablation experiments, where "√"represents structures or methods used in the model, and "/" indicates unused.

A) The effect of Prob-Sparse attention

In this section, we compare the models that have two different attention mechanisms: Prob-Sparse self-attention blocks and global self-attention blocks (Full-attention). Our goal is to investigate their impact on classification accuracy and computational complexity.

From Table 2, models with Prob-Sparse attention show superior classification accuracy to those using Full-attention. This validates the hypothesis that Prob-attention can facilitate more effective feature representation learning.

Furthermore, we compute the Floating Point Operations Per Second (FLOPs) for both approaches to analyze their computational complexity. From the result, the Prob-attention model exhibit lower FLOPs. Specifically, without self-attention distillation, the Full-attention model requires 108M FLOPs, whereas the Prob-attention model only needs 97M FLOPs. This verifies the assumption that Prob-attention can modestly reduce the model's complexity.

In summary, we find that Prob-Sparse attention outperforms Full-attention in classification accuracy while also offering a slight reduction in computational complexity. Consequently, we opt for the Prob-Sparse self-attention mechanism for our subsequent models.

B) The effect of time alignment

To validate the effectiveness of the time-alignment strategy, we conduct comparative experiments on three different data processing strategies: time alignment, time disorder, and raw data. We evaluate their impacts on model performance.

Time alignment involves generating multiple functional connectivity matrices with slight temporal offsets. This approach increases the sample diversity of the training data while maintaining time connectivity, thus reducing the sensitivity of the data to time order. For comparison, we introduce a time disorder method which reorders the time-series data and cause temporal discontinuity in the DFC matrix. Lastly, raw data represents using the original data without any processing.

Table 3 shows that the time-aligned model outperforms the other approaches in accuracy, validating its effectiveness in ASD classification. Additionally, through comparing with the time disorder method, it reveals that the accuracy improvement is also dependent on preservation of temporal connectivity besides augmented training data.

C) The effect of self-attention distillation

Table 2
Ablation experiments of the models.

Prob-attention	Full-attention	Time alignment	SA Distilling	ACC(%)	FLOPs
/	✓	✓	✓	71.1	96M
/	✓	✓	/	/	108M
✓	/	/	✓	72.3	72M
✓	/	✓	/	73.4	97M
✓	/	✓	✓	79.8	72M

Table 3
Comparison test of time alignment.

Data	ACC(%)
Time alignment	79.8
Time disorder	64.4
Raw data	72.9

Table 2 presents the results without distillation. From the results, it can be seen that when the input matrix size is [6670,24], Prob-attention with self-attention distillation achieve an accuracy of 79.8 %, surpassing the other models without distillation. Moreover, the FLOPs for this model is 72M, significantly smaller than the non-distilled models. This confirms that self-attention distillation can effectively reduce the computation resources.

4.3. Comparison of ASD diagnosis using different methods

Table 4 lists the experimental results of our model and several competing methods on the same dataset. It includes various machine learning and deep learning approaches, with inputs categorized into SFC and DFC.

For static FC, three classic machine learning methods, i.e., Random Forest [21], Logistic Regression [22], and Support Vector Machine [23], are used for classification on SFC. C-FCN is a conventional FC network based on Pearson correlation. SA-GCN integrates low-order SFC network into a graph convolutional network structure with self-attention pooling [24]. MHSA-FCN is a network that utilizes a stacked Transformer encoder for SFC classification.

For DFC classification (see the result below the horizontal line), SVM directly uses DFC matrices as input, while Lo-D-FCN generates DFC matrices by encoding non-stationary interactions between different brain regions using sliding windows, and subsequently employs SVM for classification. $(2D)^2$PCA introduces a novel feature extraction framework grounded in two-directional two-dimensional principal component analysis [25]. CM-FCN extracts central distance features from Lo-D-FCN to mitigate its temporal sensitivity [10]. JLGCN presents a multi-level dynamic brain network joint learning architecture based on GCN [26]. Our proposed method, PSA-FCN, is a DFC network with Prob-parse self-attention as its core component.

All experiments are evaluated using 10x10-fold cross-validation. Specifically, the data is divided into 10 equal-sized subsets, with one subset used as the test set and the remaining 9 subsets as the training set. To further mitigate any potential bias in fold selection, the entire 10-fold cross-validation procedure is repeated 10 times, with different random splits of the samples each time. Four widely used metrics are adopted for the performance assessment, including classification accuracy (ACC), sensitivity or true positive rate (TPR), specificity or true negative rate (TNR), and F1 score. In our experiments, ASD is treated as the positive class while NC as the negative one.

From the result in Table 4, we have the following observations. 1). Among the three traditional machine learning methods, SVM achieves higher accuracy on SFC and effectively extracts features for classification, especially in situations with limited data. Consequently, a significant portion of our subsequent comparative work (C-FCN, Lo-D-FCN, CM-FCN) focuses on features selection based on different machine learning methods, followed by SVM models trained with different types of FCN. However, directly feeding the features from DFC matrix into SVM may lead to overfitting, limiting its direct usage in DFC problems. 2). Deep learning methods consistently outperform conventional machine learning methods for both static and dynamic functional connectivity. This shows the robustness of deep learning methods in handling high-dimensional features in complex tasks. Note that deep learning models achieve higher specificity, which highlights their superior ability to identify negative-class samples. This suggests a potential reduction in false ASD diagnoses, which can avoid unnecessary concerns, reevaluations, and treatments. 3). DFC methods show better performance compared with the methods above the horizontal line. It can be observed that the accuracy of $(2D)^2$PCA, which uses traditional machine learning techniques, is higher than that of all other SFC methods except for MHSA-FCN. MHSA-FCN achieves the highest accuracy among SFC methods at 79.4 %. However, this accuracy is still lower than JLGCN's 80.3 % and our PSA-FCN's 81.8 %. Notably, our method strikes a good balance between high specificity and high sensitivity, ensuring accurate diagnosis and effective

Table 4
Comparison test of multiple models.

Method	ACC(%)	Recall(%)	Specificity(%)	F1 score(%)
RF	67.9 ± 0.03	73.5 ± 0.02	71.0 ± 0.03	67.5 ± 0.02
LR	71.4 ± 0.08	72.5 ± 0.04	72.5 ± 0.05	71.0 ± 0.04
SVM	73.9 ± 0.09	74.5 ± 0.06	74.5 ± 0.06	74.0 ± 0.01
C-FCN	74.2 ± 0.03	71.4 ± 0.02	76.9 ± 0.07	72.4 ± 0.07
SA-GCN	75.0 ± 0.05	70.2 ± 0.04	77.4 ± 0.02	72.1 ± 0.09
MHSA-FCN	79.4 ± 0.06	77.0 ± 0.02	82.5 ± 0.01	80.6 ± 0.05
Lo-D-FCN	74.9 ± 0.02	66.1 ± 0.04	79.2 ± 0.06	69.4 ± 0.06
$(2D)^2$PCA	76.1 ± 0.01	73.3 ± 0.03	78.7 ± 0.01	75.0 ± 0.05
CM-FCN	76.1 ± 0.05	75.4 ± 0.01	74.9 ± 0.05	74.4 ± 0.08
JLGCN	80.3 ± 0.01	75.6 ± 0.06	83.5 ± 0.02	79.4 ± 0.01
PSA-FCN	**81.8 ± 0.08**	**81.2 ± 0.02**	**83.8 ± 0.07**	**82.5 ± 0.03**

patient screening.

Additionally, to assess whether the model has learned features that generalize across different sites, we use the PSA-FCN to evaluate its performance across various sites and compared it with the SVM, RF, CNN, and MHSA-FCN methods [27]. Given the differences in MRI scanning protocols mentioned earlier, we select sites with scan times of 170 units or more, truncating data that exceeded 170 units and discarding those below. Table 5 presents the results of the 10-fold cross-validation for the selected sites. As shown in the table, our proposed PSA-FCN method consistently outperforms multiple methods across most sites, achieving the highest average accuracy of 78.1 %. This performance advantage demonstrates the robustness of the PSA-FCN model, validating its ability to perform well on diverse datasets from different sources and the potential for ASD classification on MRI datasets from different sources.

5. Discussion

In this paper, we present an ASD classification model based on time-aligned DFC matrices and Prob-Sparse self-attention. The novelty of our models lies in the utilization of time alignment for the construction of the cyclic DFC matrix, the introduction of the Prob-attention mechanism to extract for spatiotemporal characteristics, and distilling and downscale the self-attention through a convolution module. The time-aligned DFC matrix reduces time sensitivity, while the expanded dataset avoids overfitting. The Prob-attention mechanism introduces probabilistic sparsity, which enables the model to focus on significant positions/channels during computation and optimize the processing of high-dimensional features like DFC. Furthermore, the self-attention distillation aids the model in learning local features of the data. Especially, the convolution module enables the model to capture both local patterns and global relationships. The presence of pooling layers also reduces the data dimensionality and enhances the model's generalization capabilities. Based on experimental results, we will provide a more detailed exploration of PSA-FCN.

5.1. Effects of other hyperparameters

We also explore several other hyperparameters that potentially impact the model. Firstly, the number of stacked encoder blocks, as shown in Table 6, influences the model's accuracy. The highest accuracy of 81.8 % is achieved when the number of layers is set to 3. As this number increases, the model's FLOPs get a slight increase. However, due to the increasing number of distillation layers and multiple dimension reductions, the overall model parameters decrease progressively.

Secondly, the scaling factor in Prob-attention also plays a role in the model. First, it determines the number of samples during attention key-value computation. It selects a subset of "Keys" for computation with all "Queries" to reduce computation. Secondly, it controls the range of attention scores. By multiplying with the factor, it expands or narrows attention scores, thus impact the attention distribution, i.e., more concentrated or dispersed. Through adjusting the factor, we can make the model's attention mechanism adapt to different tasks. However, in our experiments we find that the factor changes have limited impact on the classification accuracy. As shown Table 7, The highest accuracy of 81.8 % is achieved when factor is set to 5.

5.2. Computational complexity in practical

The computational complexity of the proposed model may raise concerns about its practicality for real-time applications. To address this, we have conducted a detailed analysis of the training environment and the training/inference times of the model. All models in this study were trained on a workstation equipped with an NVIDIA 2060 GPU. With a dataset of 1034 samples and a batch size of 2, each training epoch took an average of 4 min, achieving an iteration speed of 4.07 iterations per second (it/s). During the inference phase, with a dataset of 92 samples, the inference time was recorded at 46 s, with an iteration speed of 1.99 it/s.

When compared to traditional machine learning methods such as Random Forest, which can achieve thousands of iterations per second on the same dataset, it is evident that the PSA-FCN model requires a longer training time. This increased duration is an inherent characteristic of complex deep learning models.

However, it is important to highlight that the computational complexity of PSA-FCN has been significantly optimized relative to other deep learning models employed in this study. For example, a full-attention FCN requires 10 min and 56 s per epoch with an

Table 5
Performance in different sites datasets.

Site	Size	PSA-FCN(%)	SVM(%)	RF(%)	MHSA-FCN(%)	CNN(%)
CALTECH	37	71.4	70.2	48.5	64.6	54.0
LEUVEN	61	76.9	63.0	63.2	70.4	65.3
Olin	25	68.5	68.7	64.3	76.0	58.5
Pitt	44	79.9	70.8	67.7	65.8	69.9
SBL	26	85.6	53.0	60.8	89.3	56.0
SDSU	33	74.1	72.6	63.1	75.7	75.7
Stanford	36	78.6	61.5	51.2	77.5	48.8
USM	61	77.5	76.8	67.5	76.9	66.3
UM	112	83.4	79.9	70.1	74.9	77.4
Yale	48	84.7	71.0	66.2	75.1	69.7
Avg		78.1	68.7	62.6	74.6	64.2

Table 6
Impact of the encoder layers.

Layers	ACC	FLOPs	Total Parameters
2	74.5	58M	171482
3	81.8	72M	98042
4	78.6	79M	64586
5	79.5	82M	51146

Table 7
Impact of scaling factor.

Factor	ACC
2	80.2
3	79.8
4	81.5
5	81.8

iteration speed of 1.53 it/s on the same dataset. In contrast, a Prob-Sparse self-attention model without self-attention distillation takes 8 min and 33 s per epoch, with an iteration speed of 2.01 it/s. This clearly demonstrates that the use of Prob-Sparse self-attention and self-attention distillation in PSA-FCN offers substantial improvements in computational efficiency, making it a more feasible option for practical applications compared to other models.

5.3. Discriminative features for ASD diagnosis

To validate the effectiveness of the Prob-attention mechanism, we extract biomarkers that play a decisive role in the classification process. Specifically, A pre-trained model of PSA (without the self-attention distillation) is used to process the test data, followed by extracting two sets of attention weight matrices from correctly classified data. The difference between these matrices yielded the differential attention weight matrix. Since there is no self-attention distillation, the data dimensions remain unchanged during forward propagation, implying a direct connection between the attention matrix and the original brain regions, where the magnitude of weights reflects the degree of attention to specific brain regions. After that, we conduct a statistical analysis on the brain region pairs corresponding to the top 20 % largest values in the differential matrix. These pairs are ranked by occurrence frequencies, identifying the top 10 brain regions as the primary classification biomarkers. The results are shown in Fig. 6.

From this analysis, several inferences can be drawn: 1). The model's selection of brain areas, including Cuneus (CUN), Lingual gyrus (LING), Superior occipital gyrus (SOG), Posterior cingulate gyrus (PCG), and Precuneus (PCUN), aligns with clinical observations. Among these, CUN plays a key role in visual information processing, and LING is responsible for visual and language processing, which have been revealed in previous studies [28,29]. SOG is involved in the high-level processing of visual stimuli, and individuals with ASD may show difficulties in high-level visual tasks [30]. PCG is the brain region that monitors sensory and stereotactic localization and memory roles, and the presence of abnormal findings in the PCG has been highlighted in many studies [31,32]. And PCUN is associated with many high-level cognitive functions, such as situational memory, self-related information processing, and aspects of consciousness. Patients with ASD also differ significantly from normal patients in these cognition capabilities [28,30,33]. 2). The abnormal brain regions identified by the model are not confined to specific hemispheres or areas but are distributed uniformly across the entire brain structure. This suggests that individuals with ASD exhibit widespread abnormal brain activity. Furthermore, this finding supports the notion that Prob-Sparse self-attention can effectively extract critical features for classification by taking a broad, comprehensive view of brain activity. The selected brain regions thus offer a clinically meaningful and interpretable insight into ASD-related neural patterns.

5.4. Limitations

Despite the promising classification performance, our model still has several limitations. First, the proposed model still lacks some interpretability. Specifically, while we can analyze the selection of importance of brain regions by the Prob-Sparse self-attention mechanism alone, the dimensionality of the data changes during forward propagation in the PSA due to the presence of the self-attention distillation layer. This means that the attentional weights of subsequent layers only focus some higher-level features of the data, rather than a direct correspondence to the original data.

Second, the sliding window method, as a fundamental component in our model, lacks a gold standard for optimal parameter selection [34,35]. Determining the optimal window parameters is very challenging due to the time-consuming grid search process. Additionally, the sliding window method requires relatively high computational cost, especially for long sequential data.

6. Conclusion

In this paper, we propose an ASD classification model based on time-aligned DFC and Prob-Sparse self-attention. Based on the

ROIS	Probability of occurrence (%)
CUN.R	11.32
CUN.L	9.89
SOG.R	9.18
SOG.L	8.53
LING.R	6.39
LING.L	4.39
PCG.L	4.25
PCG.R	3.95
PCUN.L	3.67
PCUN.R	2.98

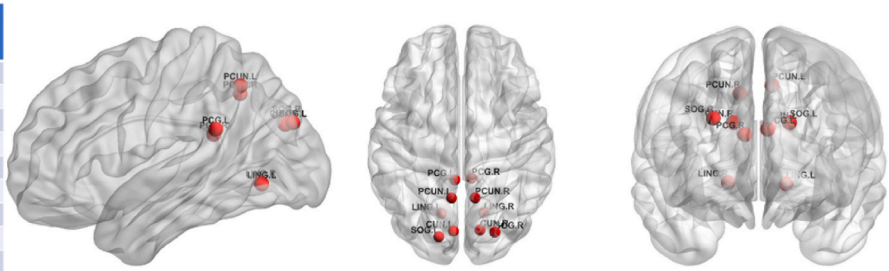


Fig. 6. The top 10 ROIs via ROI selection.

experimental results we derived the following empirical findings: 1). Prob-Sparse self-attention performs well in DFC analysis and has the potential to identify classified biomarkers. 2). Self-attention distillation significantly reduces the model's computational complexity and improves analytical efficiency. 3). Time alignment can effectively help reduce the time sensitivity of data while increasing the classification accuracy.

In future studies, we will explore multi-site data fusion or using generative methods to expand the dataset to address the overfitting problem that may arise due to limited sample size. In addition, we will explore the application of our method to other neurological and psychiatric disorders (e.g., depression, Alzheimer's, etc.) to assess its validity in a variety of contexts.

CRediT authorship contribution statement

Hongwu Chen: Writing – original draft, Validation, Methodology, Investigation, Formal analysis, Conceptualization. **Fan Feng:** Writing – original draft, Validation, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Pengwei Lou:** Writing – review & editing. **Ying Li:** Writing – review & editing. **MingLi Zhang:** Writing – review & editing, Conceptualization. **Feng Zhao:** Writing – review & editing, Methodology, Conceptualization.

Data availability statement

The code and datasets mentioned in this article are available for download at https://github.com/fan2077/PSA-FCN-InformerDFC.

Ethics approval

Review and/or approval by an ethics committee was not needed for this study because it does not involve any animal or human experimental research.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (62176140, 61972235, 61976124, 61873117, 82001775, 61976125), the Central Guidance on Local Science and Technology Development Fund of Shandong Province (YDZX2022093), Technology Innovation Program of Yantai City (2023XDRH001) and Shandong Technology and Business University Wealth Management Interdisciplinary Research Project: Research on the Innovation Path of Art Business and Management from the perspective of Big data Ecology(ID:2023TSXM007).

References

[1] C. Lord, M. Elsabbagh, G. Baird, J. Veenstra-Vanderweele, Autism spectrum disorder, Lancet 392 (10146) (2018) 508–520, https://doi.org/10.1016/s0140-6736(18)31129-2.
[2] Y. Saqr, E. Braun, K. Porter, D. Barnette, C. Hanks, Addressing medical needs of adolescents and adults with autism spectrum disorders in a primary care setting, Autism 22 (1) (2018) 51–61, https://doi.org/10.1177/1362361317709970.
[3] V.D. Calhoun, J. Sui, Multimodal fusion of brain imaging data: a key to finding the missing link(s) in complex mental illness, Biol. Psychiatr.: Cognitive Neuroscience and Neuroimaging 1 (3) (2016) 230–244, https://doi.org/10.1016/j.bpsc.2015.12.005.
[4] J. Cai, A. Liu, T. Mi, S. Garg, W. Trappe, M.J. McKeown, Z.J. Wang, Dynamic graph theoretical analysis of functional connectivity in Parkinson's disease: the importance of fiedler value, IEEE Journal of Biomedical and Health Informatics 23 (4) (2019) 1720–1729, https://doi.org/10.1109/jbhi.2018.2875456.
[5] Z. Cui, G. Gong, The effect of machine learning regression algorithms and sample size on individualized behavioral prediction with functional connectivity features, Neuroimage (2018) 622–637, https://doi.org/10.1016/j.neuroimage.2018.06.001.

[6] R.M. Hutchison, T. Womelsdorf, E.A. Allen, P.A. Bandettini, V.D. Calhoun, M. Corbetta, C. Chang, Dynamic functional connectivity: promise, issues, and interpretations, Neuroimage (2013) 360–378, https://doi.org/10.1016/j.neuroimage.2013.05.079.

[7] M.G. Preti, T.A. Bolton, D. Van De Ville, The dynamic functional connectome: state-of-the-art and perspectives, Neuroimage 160 (2017) 41–54, https://doi.org/10.1016/j.neuroimage.2016.12.061.

[8] R. Ciric, J.S. Nomi, L.Q. Uddin, et al., Contextual connectivity: a framework for understanding the intrinsic dynamic architecture of large-scale functional brain networks, Sci. Rep. 7 (1) (2017) 6537.

[9] Y. Li, Y. Zhu, B.A. Nguchu, Y. Wang, H. Wang, B. Qiu, X. Wang, Dynamic functional connectivity reveals abnormal variability and hyper-connected pattern in autism spectrum disorder, Autism Res. (2020) 230–243, https://doi.org/10.1002/aur.2212.

[10] F. Zhao, Z. Chen, I. Rekik, S.-W. Lee, D. Shen, Diagnosis of autism spectrum disorder using central-moment features from low- and high-order dynamic resting-state functional connectivity networks, Front. Neurosci. 14 (2020), https://doi.org/10.3389/fnins.2020.00258.

[11] L. Fan, J. Su, J. Qin, D. Hu, H. Shen, A deep network model on dynamic functional connectivity with applications to gender classification and intelligence prediction, Front. Neurosci. (2020), https://doi.org/10.3389/fnins.2020.00881.

[12] X. Wang, S. Sun, L. Xie, L. Ma, Efficient Conformer with Prob-Sparse Attention Mechanism for End-To-End Speech Recognition, Interspeech, 2021, https://doi.org/10.21437/interspeech.2021-415, 2021.

[13] M. Assaf, K. Jagannathan, V.D. Calhoun, L. Miller, M.C. Stevens, R. Sahl, G.D. Pearlson, Abnormal functional connectivity of default mode sub-networks in autism spectrum disorder patients, Neuroimage 53 (1) (2010) 247–256, https://doi.org/10.1016/j.neuroimage.2010.05.067.

[14] A. Di Martino, C.G. Yan, Q. Li, E. Denio, F.X. Castellanos, K. Alaerts, et al., The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism, Mol. Psychiatr. 19 (2014) 659–667, https://doi.org/10.1038/mp.2013.78.

[15] D.L. Murdaugh, S.V. Shinkareva, H.R. Deshpande, J. Wang, M.R. Pennick, R.K. Kana, Differential deactivation during mentalizing and classification of autism based on default mode network connectivity, PLoS One 7 (11) (2012) e50064, https://doi.org/10.1371/journal.pone.0050064.

[16] T.D. Satterthwaite, M.A. Elliott, R.T. Gerraty, K. Ruparel, J. Loughead, M.E. Calkins, D.H. Wolf, An improved framework for confound regression and filtering for control of motion artifact in the preprocessing of resting-state functional connectivity data, Neuroimage (2013) 240–256, https://doi.org/10.1016/j.neuroimage.2012.08.052.

[17] S.D. Washington, E.M. Gordon, J. Brar, S. Warburton, A.T. Sawyer, A. Wolfe, J.W. VanMeter, Dysmaturation of the default mode network in autism, Hum. Brain Mapp. 35 (4) (2014) 1284–1296, https://doi.org/10.1002/hbm.22252.

[18] R.C. Leung, E.W. Pang, D. Cassel, J.A. Brian, M.L. Smith, M.J. Taylor, Early neural activation during facial affect processing in adolescents with Autism Spectrum Disorder, Neuroimage: Clinical 7 (2015) 203–212, https://doi.org/10.1016/j.nicl.2014.11.009.

[19] J. Ji, Z. Chen, C. Yang, Convolutional neural network with sparse strategies to classify dynamic functional connectivity, IEEE Journal of Biomedical and Health Informatics (2022) 1219–1228, https://doi.org/10.1109/jbhi.2021.3100559.

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, N. Gomez Aidan, I. Polosukhin, Attention is all you need, Neural Inf. Process. Syst. (2017) 6000–6010.

[21] G. Biau, E. Scornet, A random forest guided tour, Test 25 (2016) 197–227.

[22] S. Menard. Applied Logistic Regression Analysis. Technometrics, 1996, pp. 192-192, https://doi.org/10.1080/00401706.1996.10484485.

[23] W.S. Noble, What is a support vector machine, Nat. Biotechnol. (2006) 1565–1567, https://doi.org/10.1038/nbt1206-1565.

[24] F. Zhao, N. Li, H. Pan, X. Chen, Y. Li, H. Zhang, N. Mao, D. Cheng, Multi-view feature enhancement based on self-attention mechanism graph convolutional network for autism spectrum disorder diagnosis, Front. Hum. Neurosci. 16 (2022), https://doi.org/10.3389/fnhum.2022.918969.

[25] F. Zhao, K. Lv, S. Ye, X. Chen, H. Chen, S. Fan, Y. Ren, Integration of temporal & spatial properties of dynamic functional connectivity based on two-directional two-dimensional principal component analysis for disease analysis, PeerJ 12 (2024) e17078.

[26] N. Li, J. Xiao, N. Mao, D. Cheng, X. Chen, F. Zhao, Z. Shi, Joint learning of multi-level dynamic brain networks for autism spectrum disorder diagnosis, Comput. Biol. Med. 171 (2024) 108054.

[27] F. Zhao, F. Feng, S. Ye, Y. Mao, X. Chen, Y. Li, M. Zhang, Multi-head self-attention mechanism-based global feature learning model for ASD diagnosis, Biomed. Signal Process Control 91 (2024) 106090.

[28] K.A.R. Doyle-Thomas, W. Lee, N.E.V. Foster, A. Tryfon, T. Ouimet, K.L. Hyde, Atypical functional brain connectivity during rest in autism spectrum disorders, Ann. Neurol. 77 (5) (2015) 866–876, https://doi.org/10.1002/ana.24391.

[29] R.J. Jao Keehn, S.S. Sanchez, C.R. Stewart, W. Zhao, E.L. Grenesko-Stevens, B. Keehn, R. Müller, Impaired downregulation of visual cortex during auditory processing is associated with autism symptomatology in children and adolescents with autism spectrum disorder, Autism Res. 10 (1) (2017) 130–143, https://doi.org/10.1002/aur.1636.

[30] Q. Wang, H.-Y. Li, Y.-D. Li, Y.-T. Lv, H.-B. Ma, A.-F. Xiang, D.-Q. Liu, Resting-state abnormalities in functional connectivity of the default mode network in autism spectrum disorder: a meta-analysis, Brain Imaging and Behavior 15 (5) (2021) 2583–2592, https://doi.org/10.1007/s11682-021-00460-5.

[31] A.L. Oblak, D.L. Rosene, T.L. Kemper, M.L. Bauman, G.J. Blatt, Altered posterior cingulate cortical cytoarchitecture, but normal density of neurons and interneurons in the posterior cingulate cortex and fusiform gyrus in autism, Autism Res. 4 (3) (2011) 200–211, https://doi.org/10.1002/aur.188.

[32] J. Wang, K. Fu, L. Chen, X. Duan, X. Guo, H. Chen, H. Chen, Increased gray matter volume and resting-state functional connectivity in somatosensory cortex and their relationship with autistic symptoms in young boys with autism spectrum disorder, Front. Physiol. 8 (2017), https://doi.org/10.3389/fphys.2017.00588.

[33] M. Assaf, K. Jagannathan, V.D. Calhoun, L. Miller, M.C. Stevens, R. Sahl, G.D. Pearlson, Abnormal functional connectivity of default mode sub-networks in autism spectrum disorder patients, Neuroimage 53 (1) (2010) 247–256, https://doi.org/10.1016/j.neuroimage.2010.05.067.

[34] L.E. Mash, A.C. Linke, L.A. Olson, I. Fishman, T.T. Liu, R. Müller, Transient states of network connectivity are atypical in autism: a dynamic functional connectivity study, Hum. Brain Mapp. (2019) 2377–2389, https://doi.org/10.1002/hbm.24529.

[35] Y. Li, J. Liu, Z. Tang, B. Lei, Deep spatial-temporal feature fusion from adaptive dynamic functional connectivity for MCI identification, IEEE Trans. Med. Imag. (2020) 2818–2830, https://doi.org/10.1109/tmi.2020.2976825.