



This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

# BrainWaveNet: Wavelet-based Transformer for Autism Spectrum Disorder Diagnosis

Ah-Yeong Jeong<sup>1\*</sup>[0009-0008-6696-1652], Da-Woon Heo<sup>2\*</sup>[0000-0001-9281-8325], Eunsong Kang<sup>3</sup>[0009-0007-3010-5144], and Heung-Il Suk<sup>2†</sup>[0000-0001-7019-8962]

<sup>1</sup> Dept. of Biomedical Engineering, Korea University, Seoul, Republic of Korea

<sup>2</sup> Dept. of Artificial Intelligence, Korea University, Seoul, Republic of Korea

<sup>3</sup> Dept. of Brain and Cognitive Engineering, Korea University, Seoul, Republic of Korea

{wjddkdud6469, daheo, eunsong1210, hisuk}@korea.ac.kr

**Abstract.** The diagnosis of Autism Spectrum Disorder (ASD) using resting-state functional Magnetic Resonance Imaging (rs-fMRI) is commonly analyzed through functional connectivity (FC) between Regions of Interest (ROIs) in the time domain. However, the time domain has limitations in capturing global information. To overcome this problem, we propose a wavelet-based Transformer, BrainWaveNet, that leverages the frequency domain and learns spatial-temporal information for rs-fMRI brain diagnosis. Specifically, BrainWaveNet learns inter-relations between two different frequency-based features (real and imaginary parts) by cross-attention mechanisms, which allows for a deeper exploration of ASD. In our experiments using the ABIDE dataset, we validated the superiority of BrainWaveNet by comparing it with competing deep learning methods. Furthermore, we analyzed significant regions of ASD for neurological interpretation.

**Keywords:** Resting-state fMRI · Autism Spectrum Disorder · Continuous Wavelet Transform · Transformer

## 1 Introduction

Autism Spectrum Disorder (ASD) is a neurodevelopmental disorder exhibiting deficiencies in social interaction, stereotypic behaviors, and hyper- or hypo-reactivity to sensory input, where early diagnosis is pivotal for reducing symptom severity through early intervention [17,8]. Resting-state functional magnetic resonance imaging (rs-fMRI) measures spontaneous blood oxygen level-dependent (BOLD) signals reflecting active brain regions, facilitating neurological condition diagnoses. Leveraging deep-learning methods enables analysis of rs-fMRI data, including BOLD signals, functional connectivity (FC), and the amplitude of low-frequency fluctuations (ALFF) of BOLD signals. Especially, frequency-based approaches (*i.e.*, fast Fourier transform; FFT, short-time Fourier transform; STFT)

\* Equally contributed.

† Corresponding author.

for fMRI-based diagnosis have gained prominence due to the cost and time efficiency [11]. However, FFT has intrinsic limitations, including sensitivity to low signal-to-noise ratios and difficulty in capturing rapidly changing non-stationary signals [9]. In addition, its complexity increases in multivariate data, leading to challenges in accurate predictions [23]. STFT enables the observation of frequency components over time, but it is limited by a trade-off between frequency and temporal resolution by selecting the window sizes [19].

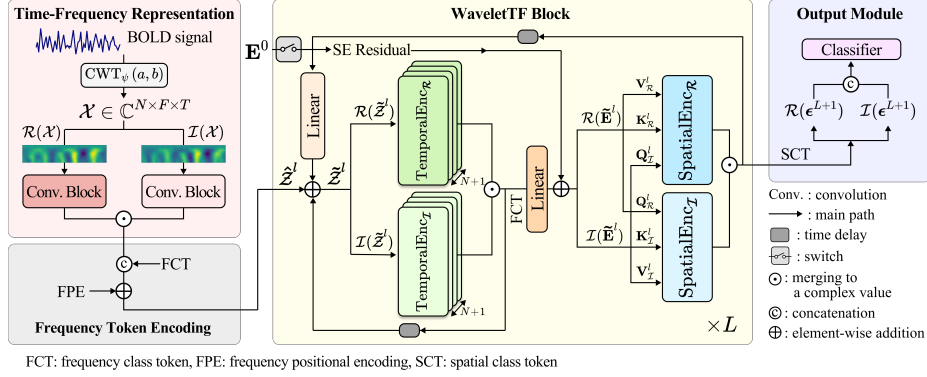
To address the limitations mentioned above and enhance ASD diagnosis, we employ the continuous wavelet transform (CWT) to process rs-fMRI BOLD signals into complex values, consisting of both real and imaginary parts. The CWT captures both frequency and temporal information, using scalable wavelets, providing a multi-resolution analysis that allows for capturing various scales and frequencies of signals [3]. This makes it more suitable for non-stationary signals and better at capturing transient features in biological signals than FFT-based methods [1]. The significance of each—real and imaginary parts—in fMRI studies is notable: most existing studies that investigate fMRI in the frequency domain, such as ALFF, use only the magnitude or real component. However, as [20] demonstrated, analyzing the real and imaginary parts of the spectral features can provide a more comprehensive understanding of the fMRI signal. This motivates us to use both the real and imaginary parts derived from CWT. By leveraging the frequency-wise temporal information from CWT and employing a Transformer model for both the temporal (*i.e.*, frequency-based) and spatial (*i.e.*, regions of interest; ROIs) axes, our proposed BrainWaveNet learns the complex dynamics of signals. It captures detailed spatial-temporal representations and regional relationships between ROIs, aiming to deepen our understanding of the intricate relationships between temporal dynamics and spatial structures among ROIs.

The main contributions of our work are as follows: (1) We introduce a novel breakthrough in the fMRI domain by effectively leveraging the multi-scale spectral features of raw data, utilizing both the real and imaginary parts of the complex-valued features obtained from the CWT. (2) We propose BrainWaveNet<sup>1</sup>, a CWT-based Transformer that effectively captures global temporal-to-spatial patterns by applying the Transformer model to both temporal and spatial axes. (3) Utilizing self-attention and cross-attention mechanisms, our model captures not only the intra-correlations within the real and imaginary parts separately but also explores their inter-relationships. (4) Our proposed model demonstrates notable efficacy in our comparative analysis, indicating its potential as a valuable tool alongside other baseline models.

## 2 Method

The BrainWaveNet architecture consists of a time-frequency representation, frequency token encoding, WaveletTF block, and output module, as illustrated in

<sup>1</sup> Our code is available at <https://github.com/ku-milab/BrainWaveNet>.



**Fig. 1.** An overview of BrainWaveNet: time-frequency representation, frequency token encoding, WaveletTF block, and output module.

Fig. 1. The input BOLD signal passes through the time-frequency representation to conduct the CWT and then passes through the residual convolutional block to extract frequency-based representation. The frequency token encoding attaches the frequency class token ( $FCT$ ) and adds the positional information. The WaveletTF block consists of a temporal Transformer encoder (TemporalEnc) and a spatial Transformer encoder (SpatialEnc) hierarchically inspired by [18]. The TemporalEnc learns the temporal embedding of each ROI at a local-level. Subsequently, only a specific token, the  $FCT$ , is forwarded to the SpatialEnc, and it learns the spatial embedding of relationships between ROIs at a global-level. Lastly, the output module projects the values from the final WaveletTF block for the classification task. We elaborate on the details of each module in the following subsections.

## 2.1 Time-Frequency Representation

Coefficients are computed by convolving the original signal with the selected mother wavelet  $\psi(t)$  over time and across various frequency scales using the CWT. The CWT of a signal  $x(t)$  can be expressed as:

$$\text{CWT}_{\psi}(a, b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} x(t) \psi^* \left( \frac{t-b}{a} \right) dt, \quad (1)$$

where  $*$  denotes complex conjugation,  $a$  and  $b$  are the scale and translation factor. For the mother wavelet, the Morlet wavelet is preferred due to its Gaussian shape in the frequency domain, which minimizes ripple effects that could be misinterpreted as oscillations [3]. Additionally, the Morlet wavelet has an optimal ratio between the Fourier period and wavelet scale, facilitating interpretation in the frequency domain [21]. Therefore, we use the complex Morlet wavelet to leverage its characteristics and utilize the real and imaginary parts of the spectral features. The complex Morlet wavelet with center frequency  $f_c$  and bandwidth

$f_b$  can be expressed as a product of a Gaussian function  $g(t)$  and sinusoidal functions as follows [16]:

$$\psi(t) = \frac{1}{\sqrt{\pi f_b}} e^{-\frac{t^2}{f_b}} e^{j2\pi f_c t} = g(t)[\cos(2\pi f_c t) + j \sin(2\pi f_c t)]. \quad (2)$$

This implies that the real and imaginary parts of the results obtained through CWT are coefficients obtained by scaling a cosine and sine function with Gaussian distribution respectively.

The input data  $\mathcal{X} \in \mathbb{C}^{N \times F \times T}$  is obtained by applying CWT to the raw data, where  $N$  is the number of ROIs,  $F$  is the number of frequency bins, and  $T$  is the number of time-steps.  $\mathcal{X}$  is a complex value containing time-frequency information, and we used both real (*i.e.*,  $\mathcal{R}(\mathcal{X}) \in \mathbb{R}^{N \times F \times T}$ ) and imaginary (*i.e.*,  $\mathcal{I}(\mathcal{X}) \in \mathbb{R}^{N \times F \times T}$ ) coefficients of the complex value. The real and imaginary parts of  $\mathcal{X}$  are passed into a residual convolutional block independently and their respective outputs are merged into a complex value  $\mathcal{X}' \in \mathbb{C}^{N \times F \times T}$  for time-frequency representation.

## 2.2 Frequency Token Encoding

**Frequency Class Token** We introduce a learnable vector  $FCT$  (*i.e.*,  $\mathbf{c}_n \in \mathbb{C}^{1 \times T}$ ) for each ROI to facilitate the extraction of spectral features, leveraging the conventional classification token [5]. It is concatenated with the time-frequency representation  $\mathbf{X}'$  as follows:

$$\mathbf{Z}_n = [\mathbf{c}_n; \mathbf{X}'_n] \in \mathbb{C}^{(F+1) \times T}, \quad (3)$$

where  $\mathbf{X}'_n \in \mathbb{C}^{F \times T}$  represents the  $n$ -th ROI in  $\mathcal{X}'$ , and  $[\mathbf{c}_n; \mathbf{X}'_n]$  denotes the concatenation of  $\mathbf{c}_n$  and  $\mathbf{X}'_n$  along the row dimension, resulting in  $\mathcal{Z} = [\mathbf{Z}_n]_{n=1}^N$ . The  $FCT$  not only serves to extract spectral information but also facilitates communication between Transformers in the WaveletTF block.

**Frequency Positional Encoding** To learn the frequency sequences of different bands for each ROI, we apply zero-initialized learnable frequency positional encoding (*i.e.*,  $FPE \in \mathbb{C}^{(F+1) \times T}$ ). We add the  $FPE$  to  $\mathcal{Z}$ :

$$\hat{\mathbf{Z}}_n = \mathbf{Z}_n + FPE = [\hat{\mathbf{c}}_n; \hat{\mathbf{X}}_n] \in \mathbb{C}^{(F+1) \times T}, \quad (4)$$

where  $\hat{\mathbf{c}}_n$  and  $\hat{\mathbf{X}}_n$  denote the  $FCT$  and  $\mathbf{X}'$  after applying  $FPE$ , resulting in  $\hat{\mathcal{Z}} = [\hat{\mathbf{Z}}_n]_{n=1}^N$ .

## 2.3 WaveletTF Block

Our proposed WaveletTF Block comprises two phases: (i) the TemporalEnc, a local-level representation learning phase, performs self-attention on the data containing time-frequency information along with Temporal Embedding (TE),

and (ii) the SpatialEnc, a global-level representation learning phase, performs cross-attention on the compact time-frequency information from TE outputs with Spatial Embedding (SE).

Concretely, the output of the frequency token encoding module  $\hat{\mathcal{Z}}$  (*i.e.*, TE) serves as an input for TemporalEnc, which comprises a stack of WaveletTF blocks. Before  $\hat{\mathcal{Z}}$  passes through the TemporalEnc, the learnable spatial embedding  $\mathbf{E}^l = [\mathbf{e}^l; \mathbf{e}_1^l; \mathbf{e}_2^l; \dots; \mathbf{e}_N^l] \in \mathbb{C}^{(N+1) \times D}$  (*i.e.*, SE), where  $l = \{1, \dots, L\}$  denotes an index of a WaveletTF block, which includes a spatial class token (*SCT*)  $\mathbf{e}^l \in \mathbb{C}^{1 \times D}$ , is linearly projected and added to the *FCT* in  $\hat{\mathcal{Z}}$  (*i.e.*,  $\hat{\mathbf{c}}_n$ ). At this stage,  $\hat{\mathcal{Z}}$  is zero-padded to align its dimensions with  $\mathbf{E}^l$ , resulting in the *FCT* being  $\hat{\mathbf{C}} = [\mathbf{0}; \hat{\mathbf{c}}_1; \dots; \hat{\mathbf{c}}_N] \in \mathbb{C}^{(N+1) \times T}$ , where  $\mathbf{0} \in \mathbb{R}^{1 \times T}$ . This operation in the  $l$ -th block for *FCT* can be described as follows:

$$\tilde{\mathbf{C}}^l = \hat{\mathbf{C}}^l + \text{Linear}(\mathbf{E}^l) \in \mathbb{C}^{(N+1) \times T}. \quad (5)$$

$\tilde{\mathcal{Z}}^l$  with updated *FCT* (*i.e.*,  $\tilde{\mathbf{C}}^l$ ) is fed into the TemporalEnc. This process distributes information along the ROIs axis, enabling the learning of spatial and global features.

**Temporal Transformer Encoder** To independently learning the temporal dependencies of the real and imaginary parts of the input through self-attention,  $\tilde{\mathcal{Z}}^l$  is divided into the real  $\mathcal{R}(\tilde{\mathcal{Z}}^l)$  and imaginary parts  $\mathcal{I}(\tilde{\mathcal{Z}}^l)$  and separately passed through the TemporalEnc. The process for the real part in TemporalEnc is as follows:

$$\mathcal{R}(\hat{\mathcal{Z}}^{l+1}) = \text{TemporalEnc}_{\mathcal{R}}(\mathcal{R}(\tilde{\mathcal{Z}}^l)) \in \mathbb{R}^{(N+1) \times (F+1) \times T}. \quad (6)$$

The outputs of TemporalEnc from each real and imaginary part are merged into a complex value (*i.e.*,  $\hat{\mathcal{Z}}^{l+1}$ ). Then, the *FCT* vector (*i.e.*,  $\hat{\mathbf{C}}^{l+1}$ ) from the TemporalEnc output is linearly projected and added to the corresponding residual SE vector (*i.e.*,  $\mathbf{E}^l$ ):

$$\tilde{\mathbf{E}}^l = \mathbf{E}^l + \text{Linear}(\hat{\mathbf{C}}^{l+1}) \in \mathbb{C}^{(N+1) \times D}. \quad (7)$$

**Spatial Transformer Encoder** To learn the spatial dynamics of the data through cross-attention between learned real and imaginary parts of coefficients, we propose SpatialEnc. SpatialEnc consists of two multi-head attention blocks for real and imaginary parts. Specifically, the cross-attention mechanism is applied in both configurations: 1) with the imaginary part as the query and the real part as the key and value, and 2) vice versa. By taking  $\tilde{\mathbf{E}}^l$  as an input, the cross-attention equation for the real part in SpatialEnc is as follows, with a similar equation for the imaginary part by changing their roles of query ( $\mathbf{Q}$ ), key ( $\mathbf{K}$ ), and value ( $\mathbf{V}$ ):

$$\text{Attention}_{\mathcal{R}}(\mathbf{Q}_{\mathcal{I}}^l, \mathbf{K}_{\mathcal{R}}^l, \mathbf{V}_{\mathcal{R}}^l) = \text{Softmax} \left( \frac{\mathbf{Q}_{\mathcal{I}}^l (\mathbf{K}_{\mathcal{R}}^l)^T}{\sqrt{d_k}} \right) \mathbf{V}_{\mathcal{R}}^l, \quad (8)$$

**Table 1.** Classification Results (mean  $\pm$  standard deviation)

Method	AUC	ACC (%)	SEN (%)	SPC (%)
BrainNetCNN [12]	0.6907 $\pm$ 0.01	64.36 $\pm$ 1.19	53.40 $\pm$ 3.79	<b>75.32<math>\pm</math>2.31</b>
BrainNetTF [10]	0.7141 $\pm$ 0.02	66.80 $\pm$ 1.78	61.06 $\pm$ 4.96	72.55 $\pm$ 8.08
STAGIN (GARO) [13]	0.5886 $\pm$ 0.07	57.23 $\pm$ 5.69	57.02 $\pm$ 11.62	57.45 $\pm$ 7.41
STAGIN (SERO) [13]	0.5839 $\pm$ 0.05	57.88 $\pm$ 1.89	51.92 $\pm$ 2.18	61.28 $\pm$ 4.02
BolT [2]	0.6989 $\pm$ 0.02	62.66 $\pm$ 3.54	55.32 $\pm$ 5.53	70.00 $\pm$ 3.31
P w/o Imag	0.6411 $\pm$ 0.06	60.43 $\pm$ 5.17	59.36 $\pm$ 15.39	61.49 $\pm$ 14.02
P w/o Real	0.6371 $\pm$ 0.04	58.83 $\pm$ 4.23	43.62 $\pm$ 17.14	74.04 $\pm$ 9.26
P w/o Cross-attn	0.6668 $\pm$ 0.02	62.45 $\pm$ 2.24	59.79 $\pm$ 4.28	65.11 $\pm$ 5.29
BrainWaveNet(Ours)	<b>0.7388<math>\pm</math>0.02</b>	<b>67.55<math>\pm</math>2.03</b>	<b>66.49<math>\pm</math>9.17</b>	68.35 $\pm$ 9.80

where  $\mathbf{Q}_{\mathcal{I}}^l$  denotes the queries of  $\mathcal{I}(\tilde{\mathbf{E}}^l) \in \mathbb{R}^{(N+1) \times D}$ , and  $\mathbf{K}_{\mathcal{R}}^l$  and  $\mathbf{V}_{\mathcal{R}}^l$  denote the keys and values of  $\mathcal{R}(\tilde{\mathbf{E}}^l) \in \mathbb{R}^{(N+1) \times D}$  respectively. The process for the real part in SpatialEnc can be expressed as follows:

$$\mathcal{R}(\mathbf{E}^{l+1}) = \text{SpatialEnc}_{\mathcal{R}}(\mathcal{I}(\tilde{\mathbf{E}}^l), \mathcal{R}(\tilde{\mathbf{E}}^l), \mathcal{R}(\tilde{\mathbf{E}}^l)) \in \mathbb{R}^{(N+1) \times D}. \quad (9)$$

The outputs of SpatialEnc from each real and imaginary part are then merged into a complex value  $\mathbf{E}^{l+1} \in \mathbb{C}^{(N+1) \times D}$ . This module not only captures the relationship between real and imaginary parts but also learns the spatial dependencies of tokens containing time-frequency information for each ROI.

## 2.4 Output Module

As mentioned earlier we utilize the *SCT* (*i.e.*,  $\epsilon^l \in \mathbb{C}^{1 \times D}$ ), for classification tasks. The final output  $\mathbf{E}^{L+1}$  from a stack of WaveletTF Blocks, can be described as follows:

$$\mathbf{E}^{L+1} = \{\epsilon^{L+1}, \mathbf{e}_1^{L+1}, \mathbf{e}_2^{L+1}, \dots, \mathbf{e}_N^{L+1}\}. \quad (10)$$

The complex value  $\epsilon^{L+1}$  is divided into its imaginary and real parts, each with a dimension of  $\mathbb{R}^{1 \times D}$ . These parts are then concatenated, resulting in a dimension of  $\mathbb{R}^{1 \times 2D}$ . This is then passed through a fully-connected layer, and the probability result is obtained through a softmax layer. We utilize the cross-entropy loss for training.

## 3 Experiments and Analysis

### 3.1 Dataset and Experimental Settings

**Dataset** We conducted our experiments on the publicly available Autism Brain Imaging Data Exchange (ABIDE)-I [6] dataset, the rs-fMRI data collected from 17 international sites. Based on the given quality control scores, 897 subjects (414 with ASD and 483 with TC) from 1,112 subjects were selected. The dataset, pre-processed by the Configurable Pipeline for the Analysis of Connectomes (CPAC)

tool, underwent band-pass filtering (0.01 – 0.1Hz) without global signal regression. The brain was parcellated using the Craddock 200 atlas [4]. Given the time variability across sites, with volumes between 78 and 316, we standardized the time size by cropping sequences to 78 volumes without overlapping. This uniform approach ensures consistency across participants, crucial for maintaining comparability in analyses. CWT was performed using the PyWavelets library [15], evenly dividing the frequency range of 0.01 – 0.1 Hz into five subbands with corresponding scale factors. Additionally, we adjusted the CWT process to account for the repetition time (TR) at each site by converting the TR to the corresponding sampling frequency. This site-specific sampling frequency was then used to scale the frequency components, ensuring accurate time-frequency representation of the fMRI signals.

**Implementation Details** In the Frequency Token Encoding, the  $FCT$  and  $FPE$  were initialized with zeros. The spatial embedding matrix  $\mathbf{E}^l$ , excluding the zero-initialized  $SCT$ , was randomly initialized with a dimension of  $D = 128$ . We set the number of WaveletTF blocks to  $L = 2$ . TemporalEnc had 3 layers with a model dimension of  $d_t = 16$ , 4 heads, and a feed-forward dimension of  $h_t = 32$ . SpatialEnc had 1 layer with a model dimension of  $d_s = 64$ , 8 heads, and a feed-forward dimension of  $h_s = 128$ . We used GELU for activation and applied a dropout rate of 0.3 to the Transformer in WaveletTF blocks. We split the dataset into 7 : 2 : 1 for training, validation, and test sets, and performed 5-fold cross-validation with a batch size of 32. Adam optimizer with an initial learning rate of  $3 \times 10^{-5}$  and weight decay of  $10^{-4}$  was used. Models were trained for 20 epochs, and the highest area under the curve (AUC) on the validation set was used for model selection. To assess performance, AUC, accuracy, sensitivity, and specificity were employed.

### 3.2 Performance Comparison

We compared our model with four baseline models as shown in Table 1.: (i) BrainNetCNN [12], a convolution-based deep neural network; (ii) BrainNetTF [10], a Transformer-based model; (iii) STAGIN [13], a model that combines graph and Transformer-based approaches for dynamic graph representation learning, featuring spatio-temporal attention. It included two readout methods, graph-attention readout (GARO) and squeeze-excitation readout (SERO); and (iv) BoIT [2], a Transformer-based model, which employed fused window multi-head self-attention. The overall data setting was the same as the proposed method.

### 3.3 Ablation studies

We ablated our BrainWaveNet in terms of complex-valued CWT and cross-attention in SpatialEnc. To investigate the impact of utilizing both real and imaginary parts through complex-valued CWT, we compared the performance of the model without the imaginary part (P w/o Imag) and the model without



## 4 Conclusion

In this paper, we propose BrainWaveNet, a novel frequency-based model that captures information of brain activation levels and the temporal dynamics between regions from fMRI data through the CWT. This work represents a breakthrough application of the CWT in the domain of fMRI analysis. In our experiments with the ABIDE dataset, BrainWaveNet demonstrated superior performance in classifying ASD when compared to existing methods. Moreover, we investigated regions related to ASD for neuroscientific interpretation. Future research will extend the scope of the analysis to encompass diverse time points and atlases.

**Acknowledgments.** This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2019-II190079, Artificial Intelligence Graduate School Program (Korea University)) and the National Research Foundation of Korea (NRF) grant funded by the Korea government (No. 2022R1A4A1033856).

**Disclosure of Interests.** The authors have no competing interests to declare that are relevant to the content of this article.

## References

1. Adeli, H., Zhou, Z., Dadmehr, N.: Analysis of EEG records in an epileptic patient using wavelet transform. *Journal of Neuroscience Methods* **123**(1), 69–87 (2003)
2. Bedel, H.A., Sivgin, I., Dalmaz, O., Dar, S.U., Çukur, T.: BolT: Fused window Transformers for fMRI time series analysis. *Medical Image Analysis* **88**, 102841 (2023)
3. Cohen, M.X.: A better way to define and describe Morlet wavelets for time-frequency analysis. *NeuroImage* **199**, 81–86 (2019)
4. Craddock, R.C., James, G.A., Holtzheimer III, P.E., Hu, X.P., Mayberg, H.S.: A whole brain fMRI atlas generated via spatially constrained spectral clustering. *Human Brain Mapping* **33**(8), 1914–1928 (2012)
5. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional Transformers for language understanding. In: *North American Chapter of the Association for Computational Linguistics* (2019)
6. Di Martino, A., Yan, C.G., Li, Q., Denio, E., Castellanos, F.X., Alaerts, K., Anderson, J.S., Assaf, M., Bookheimer, S.Y., Dapretto, M., et al.: The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. *Molecular Psychiatry* **19**(6), 659–667 (2014)
7. Doyle-Thomas, K.A., Lee, W., Foster, N.E., Tryfon, A., Ouimet, T., Hyde, K.L., Evans, A.C., Lewis, J., Zwaigenbaum, L., Anagnostou, E., et al.: Atypical functional brain connectivity during rest in autism spectrum disorders. *Annals of Neurology* **77**(5), 866–876 (2015)
8. Gabbay-Dizdar, N., Ilan, M., Meiri, G., Faroy, M., Michaelovski, A., Flusser, H., Menashe, I., Koller, J., Zachor, D.A., Dinstein, I.: Early diagnosis of autism in the community is associated with marked improvement in social symptoms within 1–2 years. *Autism* **26**(6), 1353–1363 (2022)

9. de Jesus Romero-Troncoso, R.: Multirate signal processing to improve FFT-based analysis for detecting faults in induction motors. *IEEE Transactions on Industrial Informatics* **13**(3), 1291–1300 (2016)
10. Kan, X., Dai, W., Cui, H., Zhang, Z., Guo, Y., Yang, C.: Brain network Transformer. *Advances in Neural Information Processing Systems* **35**, 25586–25599 (2022)
11. Karavallil Achuthan, S., Coburn, K.L., Beckerson, M.E., Kana, R.K.: Amplitude of low frequency fluctuations during resting state fMRI in autistic children. *Autism Research* **16**(1), 84–98 (2023)
12. Kawahara, J., Brown, C.J., Miller, S.P., Booth, B.G., Chau, V., Grunau, R.E., Zwicker, J.G., Hamarneh, G.: BrainNetCNN: Convolutional neural networks for brain networks; towards predicting neurodevelopment. *NeuroImage* **146**, 1038–1049 (2017)
13. Kim, B.H., Ye, J.C., Kim, J.J.: Learning dynamic graph representation of brain connectome with spatio-temporal attention. *Advances in Neural Information Processing Systems* **34**, 4314–4327 (2021)
14. van Kooten, I.A., Palmen, S.J., von Cappeln, P., Steinbusch, H.W., Korr, H., Heinsen, H., Hof, P.R., van Engeland, H., Schmitz, C.: Neurons in the fusiform gyrus are fewer and smaller in autism. *Brain* **131**(4), 987–999 (2008)
15. Lee, G., Gommers, R., Waselewski, F., Wohlfahrt, K., O’Leary, A.: PyWavelets: A python package for wavelet analysis. *Journal of Open Source Software* **4**(36), 1237 (2019)
16. Li, S., Ma, S., Wang, S.: Optimal complex Morlet wavelet parameters for quantitative time-frequency analysis of molecular vibration. *Applied Sciences* **13**(4), 2734 (2023)
17. Lord, C., Brugha, T.S., Charman, T., Cusack, J., Dumas, G., Frazier, T., Jones, E.J., Jones, R.M., Pickles, A., State, M.W., et al.: Autism spectrum disorder. *Nature Reviews Disease Primers* **6**(1), 1–23 (2020)
18. Lu, W.T., Wang, J.C., Won, M., Choi, K., Song, X.: SpecTNT: A time-frequency Transformer for music audio. *International Society for Music Information Retrieval Conference* (2021)
19. Mateo, C., Talavera, J.A.: Short-time Fourier transform with the window size fixed in the frequency domain. *Digital Signal Processing* **77**, 13–21 (2018)
20. Rowe, D.B., Logan, B.R.: A complex way to compute fMRI activation. *NeuroImage* **23**(3), 1078–1092 (2004)
21. Torrence, C., Compo, G.P.: A practical guide to wavelet analysis. *Bulletin of the American Meteorological Society* **79**(1), 61–78 (1998)
22. Vissers, M.E., Cohen, M.X., Geurts, H.M.: Brain connectivity and high functioning autism: a promising path of research that needs refined models, methodological convergence, and stronger behavioral links. *Neuroscience & Biobehavioral Reviews* **36**(1), 604–625 (2012)
23. Zhou, T., Ma, Z., Wen, Q., Wang, X., Sun, L., Jin, R.: Fedformer: Frequency enhanced decomposed Transformer for long-term series forecasting. In: *International Conference on Machine Learning*. pp. 27268–27286. PMLR (2022)