A CNN-Transformer Hybrid Architecture for Capturing Spatiotemporal Dependencies in Resting-State fMRI for Autism Diagnosis

Mengzhu Yu
Information Technology Department, Qingdao Vocational and Technical College of Hotel Management, Qingdao, Shandong, 266100, China
E-mail: yumz1030@163.com

Keywords: autism, fMRI data, CNN, transformer, classification

Received: October 30, 2025

Autism Spectrum Disorder (ASD) diagnosis remains challenging because of its heterogeneity and reliance on subjective behavioral assessments. Resting-state functional MRI (fMRI) presents a compelling opportunity avenue for identifying objective biomarkers, but decoding its complex spatiotemporal patterns requires advanced computational models. While Deep Learning (DL) approaches have progressed, many struggle to concurrently capture local neural dynamics and global temporal dependencies. A novel end-to-end CNN-Transformer hybrid framework designed for fMRI-based autism diagnosis is proposed to address this. Our model leverages a two-layer convolutional module (temporal + depth-wise spatial convolution) to extract localized spatiotemporal features, which are then processed by a 4-layer Transformer encoder with a 4-head Multi-Head Self-Attention (MHSA) mechanism to model long-range, global dependencies through a Multi-Head Self-Attention (MHSA) mechanism. Evaluated via a rigorous 10-fold cross-validation on the large multi-site ABIDE-I dataset (N=1,035), the suggested model achieved state-of-the-art performance with an accuracy of 77.85%, a sensitivity of 76.52%, a specificity of 78.90%, and an F1-score of 77.71%. Ablation studies confirmed the critical contribution of each architectural component, demonstrating that the integration of the Transformer encoder and residual connections provided a significant performance boost over the CNN-only baseline. and comparisons with pre-trained CNNs and other leading methods demonstrated superior and statistically significant performance (p<0.05). Despite an observed performance drop in site-specific evaluations, underscoring the challenge of scanner heterogeneity, our results affirm that the synergistic integration of local feature learning and global contextual modeling is a powerful paradigm for neuroimaging-based diagnostic applications.

Povzetek: Predlagan je hibridni CNN-Transformer model za diagnozo ASD iz mirujočih fMRI posnetkov, ki združi učenje lokalnih vzorcev in dolgoročnih odvisnosti ter na podatkovni zbirki ABIDE-I doseže približno 78 % natančnost in boljše rezultate od primerjanih metod.

1 Introduction

It is tough to effectively diagnose neurodevelopmental disorders like autism because such disorders have a vast variety of symptoms, especially in the younger population [1], [2]. In general, psychiatric diagnosis is mainly based on the observation of the behavior of the patient, which is subjective and is done according to the criteria given in manuals like DSM-5 and ICD-10; this kind of diagnosis can be wrong sometimes [3], [4]. In contrast to many physical diseases, for example, HIV and diabetes, which can be verified by objective laboratory measures, mental health disorders do not have consistent biological markers. This makes differential diagnosis difficult because of overlapping symptoms and no definitive tests [5]. ASD is a condition that is often described as a neurodevelopmental disorder with a particular emphasis on the lifelong trajectory. ASD is described by social communication problems with friendship and peer relationships, repetitive behavior difficulties, and restricted interests, usually identifiable in early childhood [6,8]. The case of a child with autism is just one among 100 children worldwide [4,9]. Hence, finding the disorder as early as possible is necessary to intervene in time. Besides, demographic variations in the rates of autism diagnosis are quite considerable, as male-to-female differences in the rates of diagnosis are around four times, implying gender-based disparities [10]. To increase the accuracy of diagnosis, the scientists have employed various means such as structured behavioral observation, study of demographic trends, and brain imaging [11], [12]. The most recent research also deals with behavioral signal processing; the work of Alkahtani et al. exemplifies this [13] and Pandian et al. [14], where the authors have used automated analysis of facial expressions and gaze patterns, respectively, as one of the main behavioral signals in their work. Besides that, Zunino et al. [15] harnessed the power of recurrent neural networks by employing them to analyze video data to identify individuals on the autism spectrum and those without neurodevelopmental conditions, i.e., neurotypical controls.

Quantitative evaluations of neuroimaging data may identify important biomarkers that could greatly enhance the diagnostic capability of neurological and psychiatric disorders [16], [17]. Machine learning methods have become popular over the last few years to examine basic and functional Magnetic Resonance Imaging (MRI and fMRI) data to diagnose diseases, namely, Alzheimer's [18], ADHD [19], and Autism [20]. This article primarily distinguishes between autistic and typically developed individuals using resting-state fMRI. The utilization of fMRI for detecting autism has become a popular topic, mainly due to large-scale public datasets like the Autism Brain Imaging Data Exchange (ABIDE), which collects brain imaging data from different sites worldwide [21]. Many studies utilized ABIDE data to develop and evaluate a number of classification models [22], [23], [24], [25]. Some studies specifically employed several age groups of the data set; for example, Iidaka developed a probabilistic neural network to classify the rs-fMRI data for subjects younger than 20 years of age [26]. Plitt et al. tested two different segments of the ABIDE rs-fMRI data and claimed a classification accuracy of 76.67% [27]. Parisot et al. improved a graph convolutional network model representing individuals as nodes with features extracted from the images and using phenotypic data as edge weights, reaching an accuracy of 70.4% [28]. Sen et al., in a broader study, presented a hybrid algorithm that integrated features from both sMRI and fMRI and hence achieved 64.3% accuracy in 1,111 participants from both groups [29], [30]. Moreover, some researchers have also examined the phenotypic markers besides the imaging data: Parikh et al., for example, implemented demographic variables—like age, sex, handedness, and IQ scores—as features and evaluated their utility in the ABIDE repository using different machine learning methods [30].

During the last couple of years, deep learning (DL) architectures, along with neural network models—like autoencoders, Convolutional Neural Networks (CNNs), and Long Short-Term Memory (LSTM) networks—have been extremely popular as a means of achieving better autism classification [31], [32]. Brown et al. presented a novel DL framework that combined an element-wise layer with data-informed structural priors, thus obtaining a 68.7% accuracy level for a total of 1,013 individuals (474 with autism and 539 healthy controls) [33]. Heinsfeld et al. similarly executed a DL strategy on the same data, but the library contained one thousand thirty-five subjects (five hundred five with ASD and five hundred thirty controls). They reached 70% accuracy and claimed that their performance was better than that of the previous works [34]. Moreover, Qiang et al. moved the research forward by creating a hierarchically structured recurrent variational autoencoder without supervision for detecting autism with ABIDE data, which resulted in a remarkably high accuracy of 82.1% [35]. Additionally, Subah et al. were able to achieve autism classification accuracy of 88% by building a DL model that utilized the AAL and CC200 brain atlases, and they reported that prediction of individuals with autism could be made with high accuracy [36].

These models, however, have significant limitations in that they are not easily parallelized for training and have a limited capacity for modeling long-range temporal dependencies. Besides, a single, fully comprehensive, accurate model suitable for clinical practice has not yet been presented. Meanwhile, the phenomenal success of the Self-Attention (SA)-based Transformer model in domains like computer vision has led to its adoption for fMRI analysis [37]. The main power of the architecture comes from its SA mechanism, which offers a global receptive field that can fetch data from the whole sequence to enhance the model's efficacy. So far, the use of such Transformer-based architectures specifically for autism classification via fMRI data is very limited. Indeed, the most recent approaches have, in fact, gradually supplanted the SA mechanism of the Transformer with their strategy that provides a greater receptive field, and higher potential to include global contextual information, and yields improved performance. A typical limitation associated with these approaches is that they are inclined to underestimate the importance of local feature learning, which is at the core of the fMRI signal decoding process. While these Transformer-based architectures have surpassed the old ones in decoding accuracy, a large room for improvement still exists. To address such shortcomings, a hybrid framework is developed that effectively merges the two most promising technologies, CNNs and Transformers, thus creating complementary effects between them. The suggested network is an end-to-end architecture comprising a convolutional module, followed sequentially by a Transformer encoder, and culminating in a classification layer. To formally guide this investigation, the study is structured around the following research questions (RQs) and corresponding hypotheses (Hs):

RQ1: Can a hybrid CNN-Transformer architecture more accurately classify ASD from rs-fMRI data compared to existing CNN-only or Transformer-only models?

H1: The proposed hybrid model will achieve superior classification performance on the multi-site ABIDE-I dataset by synergistically combining local spatiotemporal feature extraction with global temporal dependency modeling, outperforming state-of-the-art benchmarks.

RQ2: What is the individual contribution of the core architectural components (i.e., the convolutional module, the Transformer encoder, and their residual integration) to the overall model performance?

H2: Ablation studies will confirm that each component is critical, with the Transformer encoder providing a significant performance boost by capturing long-range dependencies, and the residual fusion of local and global features yielding the most robust and accurate model.

RQ3: How does the model's performance generalize across heterogeneous data acquisition sites, and what is the impact of key hyperparameters like the number of Transformer layers and attention heads?

H3: While the model will demonstrate strong overall generalizability, its performance will exhibit site-specific variability due to scanner heterogeneity. Furthermore, an optimal configuration of architectural hyperparameters (e.g., 4 layers, 4 heads) will exist, balancing model complexity with effective learning.

The primary points of the work are essentially the following: (I) A novel robust model structure is built utilizing CNNs to localize features and a Transformer encoder to effectively model long-range dependencies across the entire context of fMRI data; (II) The suggested model is the new state-of-the-art benchmark, performing at the forefront and thus setting the standard for future research. These findings imply the robustness of the model's generalization capacity and indicate its potential as a reference model in subsequent research in fMRI-based decoding; and (III) Several component-based comparisons and ablation studies are performed to account for the component contributions of our framework.

2 Methods

A convolutional transformer network with a MHSA mechanism is presented to classify resting-state fMRI data as an objective nonverbal tool for autism analysis. Fig. 1 illustrates the block diagram of the suggested framework. This type of system allows for directly classifying fMRI time series from raw data, thus eliminating the need for manually engineered features. Three main components make up the model: a fully connected classifier, a Transformer encoder, and a convolutional module. The convolutional module detects local spatiotemporal patterns in the brain signals obtained from fMRI by applying two convolutional layers: a one-dimensional temporal convolution, followed by a depth-wise spatial convolution. The input for the module consists of standard regional time series, and it can encode both the temporal changes in blood-oxygen-level-dependent (BOLD) signals, as well as the functional interactions between different brain areas. The result is a new feature sequence that captures more advanced temporal information. The resultant feature sequence is the input to a Transformer encoder that employs a MHSA mechanism to select and assign weights to the most important features concerning the entire sequence in question. The model culminates in a compact classification module, consisting of a fully connected layer, that generates the final diagnostic outcome. Each component of this model is detailed below.

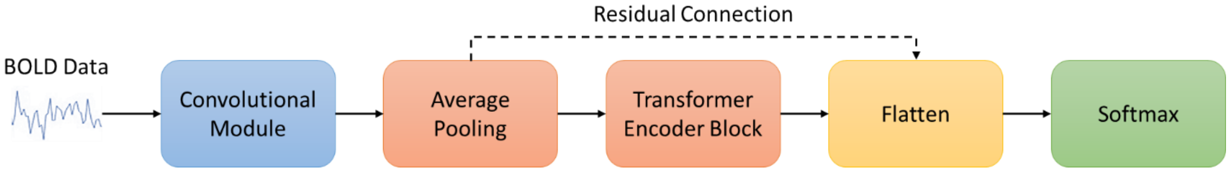

Figure 1: General block diagram of the proposed framework

2.1 ABIDE-I database

In cases where the scan was performed without the member being given a task, it is called resting-state fMRI, a common paradigm in the work of neurological and psychiatric illnesses [38], [39]. This work employed the preprocessed ABIDE-I dataset, a publicly available dataset composed of 1,112 resting-state fMRI scans collected from 505 persons with ASD and 530 typically developing controls at 17 different sites internationally, summarized in Table 1. The dataset contains representative time series extracted from several a priori defined regions of interest (ROIs) informed by multiple brain atlases and processed from various preprocessing pipelines. The current study utilized the data which had been preprocessed utilizing the Configurable Pipeline for the Analysis of Connectomes (C-PAC) [40] and parcellated into 116 coherent functional regions based on the AAL atlas [41] The preprocessing procedure included standard procedures and included the removal of initial volumes to allow for stabilization of the magnetic field, motion realignment, regression of nuisance signals, slice-timing correction, band-pass filtering to decrease low-frequency drift and high-frequency noise, and intensity normalization. One limitation to keep in mind is that the gaining factors, namely, Echo Time (TE), Repetition Time (TR), spatial and temporal resolutions, and whether participants' eyes were opened or closed were not consistent across scanning sites.

Table 1: Details of the ABIDE-I database for every imaging location

Site	Yale	USM	UM	UCLA	Trinity	Stanford	SDSU	SBL	PITT	OLIN	OHSU	NYU	MaxMun	Leuven	KKI	CMU	Caltech
#Of Autism	28	46	66	54	22	19	14	15	29	19	12	75	24	29	20	14	19
#Of Control	28	25	74	44	25	20	22	15	27	15	14	100	28	34	28	13	18
#Of Male	40	71	113	86	47	31	29	30	48	29	26	139	48	55	36	21	29

2.2 Convolutional module

Our framework (Fig. 2) has a convolutional module that identifies and isolates spatiotemporal features of raw data in BOLD time series form. The data are arranged in terms of the batch size, the quantity of ROIs, and the sampling points in terms of time (32, 116, 200). First, a one-dimensional temporal convolution is executed that goes across the time axis for each ROI separately. In this layer, 16 filters with a kernel size of 64 capture local temporal patterns and fluctuations within each region's signal; thus, the input is effectively expanded into a feature-rich representation. The result is immediately standardized with batch normalization and gets to the layer with the ReLU activation function for stable and non-linear processing. This stage changes the input to a different shape of (32, 16, 116, 200), where the second dimension now stands for those learned temporal features.

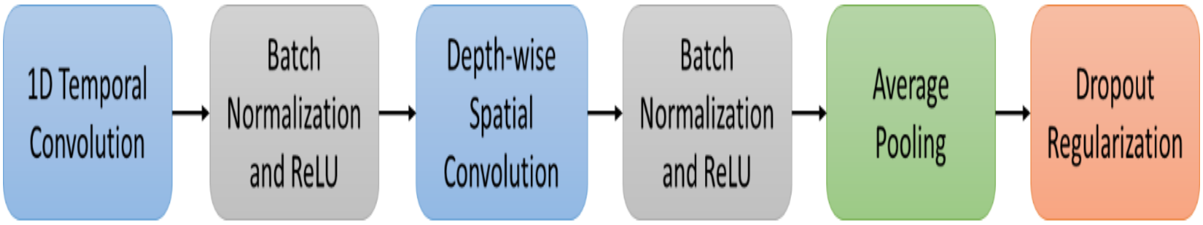
1D Temporal Convolution
Batch Normalization and ReLU
Depth-wise Spatial Convolution
Batch Normalization and ReLU
Average Pooling
Dropout Regularization

Figure 2: Convolutional module of our proposed framework to extract spatiotemporal features from the raw BOLD data

After temporal feature extraction, the depth-wise spatial convolution is executed to represent the interactions of different brain regions. A kernel size of 116, i.e., the number of ROIs, and a depth multiplier of 2 are used in this layer, which indicates that it learns two spatial filters for each of the 16 temporal feature maps. To perform the "depth-wise" operation, the quantity of groups is assigned the value of the quantity of input channels (16); each temporal feature is processed by its set of spatial filters without cross-feature interference. Therefore, the output shape is (32, 32, 1, 200), which combines the spatial relationships across the brain in a compact form. The next sequence is another batch normalization followed by the ReLU activation. To reduce computational complexity for subsequent transformer layers and highlight the most salient features, an average pooling operation with a size and stride of 8 is applied along the temporal dimension, downsizing the sequence length from 200 to 25. A dropout layer with a dropout rate of 0.3 is employed for regularization, and the output is finally reshaped into a sequence of 25 tokens, each with 32 features (32, 25, 32), preparing it for global temporal modeling in the transformer encoder.

2.3 Transformer encoder block

The Transformer encoder block (see Fig. 3) is tasked with modeling the global, long-range temporal dependencies within the feature-embedded sequence produced by the convolutional module. The input to this block is a sequence of 25 tokens, each represented as a 32-dimensional vector, effectively forming a compact, higher-level representation of the brain's activity over time for a batch of 32 subjects. The encoder is composed of a stack of four identical layers, each with 2 main sublayers. The first sublayer is a MHSA mechanism (Fig. 4) that utilizes four attention heads. This enables the model to simultaneously focus on data from multiple distinct representation subspaces; splitting the features (32-dimensional) into four 8-dimensional heads allows the mechanism to attend to differences in temporal dynamics within the entire BOLD signal and thus identify complex, non-local interactions across the whole-time sequence. For an input sequence of token embeddgins X ∈ $\mathbb{R}^{T\times dmodel}$ where T = 25 is the sequence length and d_{model} = 32 is the feature dimension. First, the input X is projected into Queries (Q), Keys (K), and Values (V) using learned weight matrices W^Q, W^K, W^V ∈ $\mathbb{R}^{dmodel\times dk}$, where d_k = d_{model}/h = 8 and h = 4 is the number of attention heads:

$$Q = XW^Q, K = XW^K, V = XW^V \qquad (1)$$

The scaled dot-product attention for each head is then computed as:

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \qquad (2)$$

The scaling factor $\frac{1}{\sqrt{d_k}}$ prevents the softmax function from entering regions with extremely small gradients. The outputs of all h heads are concatenated and projected back to the original dimension d_{model} using a learned weight matrix W^O ∈ $\mathbb{R}^{h.dk\times dmodel}$ to form the final MHSA output:

$$\text{MHSA(Q, K, V)=Concat}(\text{head}_1, \ldots, \text{head}_h)W^O \qquad (3)$$

Where $head_i = Attention(XW_i^Q, XW_i^K, XW_i^V)$. In our framework, this mechanism enables each of the 25 temporal tokens to compute a weighted sum over all other tokens in the sequence. The attention weights $softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)$ explicitly quantify the pairwise influence between timepoints, allowing the model to identify critical, globally-informative moments in the fMRI sequence that are most relevant for the autism classification task.

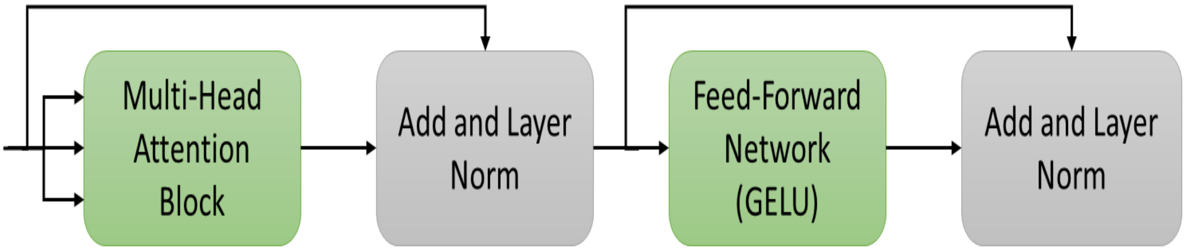

Figure 3: Transformer encoder block of our proposed framework

Subsequent to the computation of the attention weights, the resulting output is integrated with the original input via a residual connection, followed by normalization using LayerNorm. This critical step helps the training of the deep network to be stable. The scaled output is received by the 2nd sub-layer, which is a position-wise Feed-Forward Network (FFN). The FFN contains a linear projection that expands the dimensionality of the model from 32 values to 128 values, and it applies a GELU activation function to make the function nonlinear, before projecting back to the original 32 dimensions. The operation is independent and identical for each token, thus allowing further non-linear feature transformation. The output of the FFN is once more combined with its input through another residual connection and LayerNorm. The structured combination of SA and feed-forward processing, with the help of residuals and normalization, allows each of the 25 tokens to be informed by every other token in the sequence. Thus, the Transformer encoder turns the feature embeddings into a potent representation where each element is globally-informed and understands the entire temporal dynamics, which is necessary for the final autism classification.

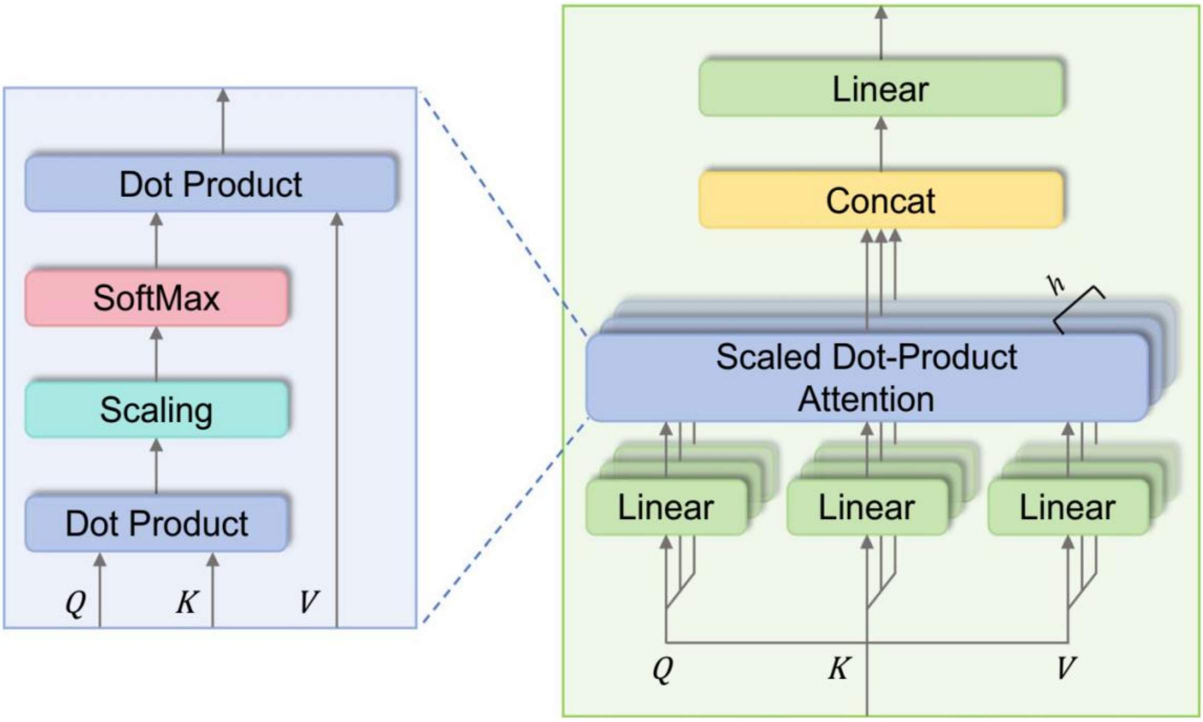

Figure 4: Multi-head attention

2.4 Classification module

At the classification tier of the model, the feature maps obtained from the convolutional module are fused with the output of the Transformer encoder through a residual connection. This additive coupling guarantees that the localized spatiotemporal patterns identified by the convolutional layers and the globally contextualized representations derived by the Transformer are preserved and made available for direct propagation to the subsequent layers. After that, the combined feature maps are transformed from a multi-dimensional tensor into a one-dimensional vector to be compatible with the dense layers. To reduce overfitting and enhance the model's generalization capability on unseen data, a dropout layer with a ratio of 0.5 is employed, whereby 50% of the activations are randomly deactivated during the training process. The resulting feature vector is subsequently input into the final fully connected layer, which comprises 2 output units corresponding to the 2 diagnostic categories in the autism classification task (i.e., autism and healthy). The complete network is trained in an end-to-end manner utilizing the cross-entropy loss function, which quantifies the divergence between the predicted probability distribution and the true diagnostic labels.

3 Results

The training process was performed on an NVIDIA RTX 2080 Ti GPU (11 GB VRAM). The TensorFlow DL library ran on a Windows 11 OS and on an Intel Core i7-12700K processor. The ABIDE-I dataset evaluated the proposed framework's stability, accuracy, and generalizability. The model was trained for a maximum of 150 epochs using the AdamW optimizer with a weight decay of 0.00001 to regularize learning. A cosine annealing schedule was applied to the learning rate, starting at 0.0001 and decaying to a minimum of 0.000001. We employed a batch size of 32 and minimized the categorical cross-entropy loss. To prevent overfitting and ensure model selection, early stopping was triggered if the validation loss failed to improve for 25 consecutive epochs, with the best-performing weights being restored. Furthermore, all random number generators for Python, NumPy, and TensorFlow were fixed with a seed of 42 to guarantee reproducible weight initialization and data shuffling across all experiments. Fig. 5 shows the steps of preprocessing fMRI data.

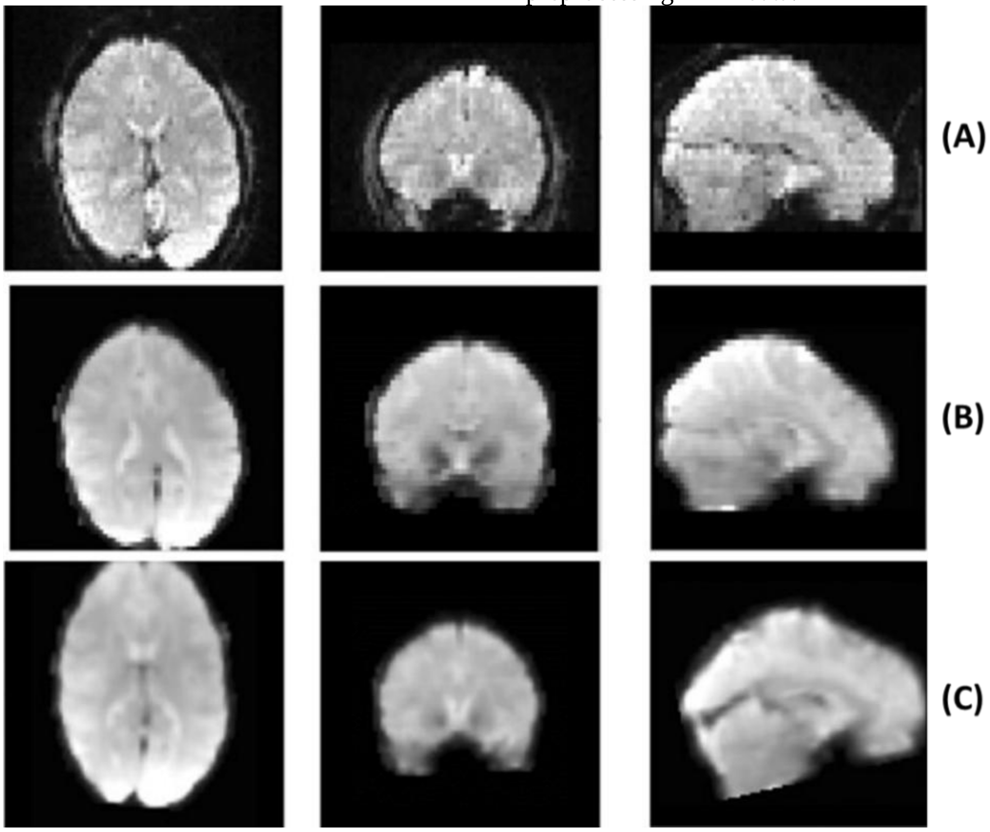

Figure 5: Sample data after preprocessing. (A) Raw resting-state fMRI images; (B) brain extraction, filtering, and slice timing correction; and (C) image registration to standard space

The proposed model architecture, detailed in Table 2, is a hybrid network designed to classify fMRI time-series data. The specific hyperparameters for this architecture were not chosen arbitrarily; as shown in Table 3, they were selected from established ranges in the literature through empirical testing and validation, balancing model complexity with performance and regularization to prevent overfitting. This approach is analogous to optimizing a control system for stability under varying operating conditions. We employed a grid search strategy combined with 5-fold cross-validation on a held-out development set from the multi-site ABIDE-I data. The primary objective was to identify a configuration that maintained high performance across different data sources, thereby inherently building resilience to site-specific scanner variations. Specifically, hyperparameters such as the number of Transformer layers and attention heads were optimized not just for peak accuracy on the aggregate data, but for consistent performance across the folds, which represent different data partitions and, implicitly, different scanner contributions. Similarly, temporal kernel sizes were evaluated for their ability to capture biologically plausible hemodynamic response dynamics across different TR (Repetition Time) parameters present in the multi-site dataset. This process ensured the model was tuned to be less sensitive to the noise introduced by scanner heterogeneity, prioritizing generalizable feature learning over site-specific overfitting.

Table 2: Model architecture: spatiotemporal convolutional network with Transformer encoder

Module	Layer / Operation	Key Parameters	Input Shape	Output Shape	Notes
Input	fMRI Time-Series	-	(32, 116, 200)	(32, 116, 200)	116 ROIs (AAL atlas), 200 timepoints.
Convolutional Module	Temporal Conv1D	F1=16 filters, kernel size= (1, 64), stride=1, padding='same'	(32, 116, 200)	(32, 16, 116, 200)	Learns temporal features. Adds a dimension.
	BatchNorm + ReLU	-	(32, 16, 116, 200)	(32, 16, 116, 200)	Stabilizes and introduces non-linearity.
	Depth-wise Spatial Conv1D	D=2, kernel size= (116, 1), groups=16, padding='valid'	(32, 16, 116, 200)	(32, 32, 1, 200)	D=2 spatial filters per temporal filter. groups=F1 makes it depth-wise.
	BatchNorm + ReLU	-	(32, 32, 1, 200)	(32, 32, 1, 200)	Processes spatial features.
	Average Pooling	pool size= (1, 8), stride= (1, 8)	(32, 32, 1, 200)	(32, 32, 1, 25)	Reduces sequence length from 200 to 25.
	Dropout	p=0.3	(32, 32, 1, 25)	(32, 32, 1, 25)	Regularization during training.
	Reshape for Transformer	-	(32, 32, 1, 25)	(32, 25, 32)	T'_c = 25 tokens, each with d_model = 32 features.
Transformer Encoder (x L layers)	LayerNorm	-	(32, 25, 32)	(32, 25, 32)	Normalizes before attention.
	Multi-Head Attention	heads=4, key_dim=8	(32, 25, 32)	(32, 25, 32)	4 heads, each with dimension 8 (32/4).
	Add + LayerNorm	Residual Connection	(32, 25, 32)	(32, 25, 32)	Adds input to attention output.
	Feed-Forward Network	dim_ff=128, activation='GELU'	(32, 25, 32)	(32, 25, 32)	Expands to 128 dims, then back to 32.
	Add + LayerNorm	Residual Connection	(32, 25, 32)	(32, 25, 32)	Adds input to the FFN output.
Classifier	Flatten	-	(32, 25, 32)	(32, 25 * 32) = (32, 800)	Prepares for a dense layer.
	Dropout	p=0.5	(32, 800)	(32, 800)	Final regularization.
	Fully Connected	units=2	(32, 800)	(32, 2)	Final feature fusion.
	Softmax	-	(32, 2)	(32, 2)	Output: [P(ASD), P(Control)].
Output	Class Probabilities	-	-	(32, 2)	Final prediction.

Table 3: Hyperparameter selection: literature-informed ranges and final model configuration

Parameter	Tested Value/Range (from Literature)	Selected Value & Rationale
Batch Size	16, 32, 64	32
# Temporal Filters	8, 16, 32, 64	16
Temporal Kernel Size	~0.25-1s of data (e.g., 32, 64, 128 for TR=0.5s-2.0s)	64
Depth Multiplier	1, 2, 4	2
Pool Size (Temporal)	4, 8, 16	8
Dropout (Conv)	0.2 - 0.5	0.3
Transformer Dimension	32, 64, 128	32
Number of Heads	2, 4, 8	4
FFN Dimension	128, 256, 512 (Often 4 * d_model)	128 (4 * 32)
Number of Layers	2, 4, 6, 8	4
Dropout (Classifier)	0.4 - 0.7	0.5

The model's evaluation involved a two-stage, 10-fold cross-validation process. Initially, it was tested on the full multi-site dataset (N=1,035) to assess its ability to generalize across diverse fMRI scanning parameters. After that, its performance was evaluated locally at each site based on smaller, site-specific datasets. Standard metrics, namely, accuracy, sensitivity, specificity, and F1-score were utilized for a thorough performance analysis.

Sensitivity and specificity were instrumental in detecting the rare classes, whereas the F1-score was a balanced measure of sensitivity and specificity relevant to the possible class imbalance. Fig. 6 presents the outcomes of the initial phase of the experiment. Our framework yielded 77.85% accuracy, 76.52% sensitivity, 78.90% specificity, and 77.71% F1-score for autism diagnosis from fMRI data. Table 4 compares our classification performance in the first stage against some known techniques employing the ABIDE-I database. The statistical importance of the performance differential between the suggested model and other leading methods is evaluated using the p-value from the paired Wilcoxon Signed-Rank Test. Our proposed model achieves the highest accuracy, F1-score, and Kappa scores, indicating superior overall performance and reliability. It also shows a strong balance between Sensitivity and Specificity. The low p-values for other models suggest their results are statistically significant compared to a baseline.

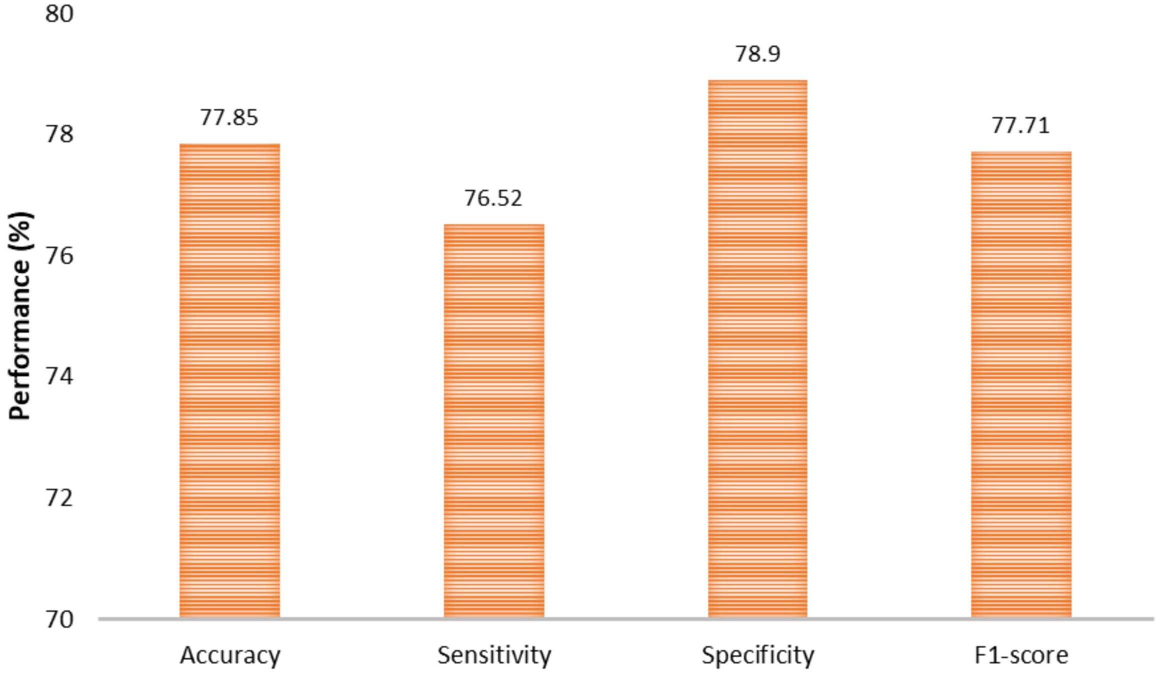

Figure 6: Performance metrics obtained by our proposed framework on the entire ABIDE-I dataset

Table 4: Performance and Kappa of leading techniques on the ABIDE-I dataset

Model	Accuracy	Sensitivity	Specificity	F1-score	P-value	Kappa
ASD-DiagNet [40]	70.32	68.30	72.24	70.21	0.008	0.625
ASD-SAENet [42]	70.81	62.25	79.11	69.67	0.010	0.632
ASDC-Net [20]	76.72	73.48	79.07	76.17	0.192	0.710
BrainNetDiffusion [43]	69.65	73.00	66.18	69.42	0.003	0.621
Ours	77.85	76.52	78.90	77.71	-	0.734

The performance of our proposed hybrid architecture was further benchmarked against several established pre-trained CNNs adapted for fMRI-based classification. These models were chosen for their proven success in computer vision tasks. The data in Table 5 reveals that our model is better than all the pretrained CNNs regarding all the critical metrics. The enhanced performance is significantly different statistically, as shown by the very small p-values (<0.05) from the paired Wilcoxon Signed-Rank Test, when comparing the accuracy distributions, among the cross-validation folds. Also, the Kappa value for our model of 0.734 depicts that the degree of agreement is considerably greater compared to the pre-trained models, which were between 0.521 and 0.682. Although transfer learning from the pre-trained vision models can be a good baseline, their architectural priors do not best fit the spatiotemporal fMRI data. Conversely, our CNN-Transformer hybrid can more effectively acquire the complex temporal dynamics and the global dependencies of the brain activities at rest; hence, it can make the classification significantly more robust and accurate.

Table 5: Performance comparison of the suggested framework with pre-trained CNN models on the ABIDE-I dataset

Model	Accuracy (%)	Sensitivity (%)	Specificity (%)	F1-score (%)	P-value	Kappa
VGG-16	70.15	68.92	71.25	70.07	0.005	0.621
ResNet-50	72.40	70.18	74.33	72.18	0.012	0.648
InceptionV3	71.83	73.50	70.25	71.84	0.009	0.637
DenseNet-121	73.65	72.04	75.01	73.49	0.038	0.673
EfficientNetB0	72.90	71.35	74.22	72.75	0.021	0.658
Proposed (Ours)	77.85	76.52	78.90	77.71	-	0.734

This shows a detailed account of how the primary architectural hyperparameters influence the model's performance. In particular, this concerns the depth of the Transformer encoder and the quantity of SA heads. The number of layers, L, one of the most critical factors of the Transformer's representational power, was changed. The depths from 1 to 10 encoder blocks were tried, as shown by the graph in Fig. 7. The classification accuracies were changing widely with different depths. The 7-layer model averaged classification accuracy at 77.85%, 4.95% more than the 10-layer model. The initial performance gain with increasing depth, from L=1 to L=7, is attributable to the model's enhanced representational capacity for capturing the complex, hierarchical temporal dependencies in the fMRI signal. However, the subsequent decline in accuracy for deeper models (L=8 to L=10) is a strong indicator of overfitting. Deeper networks have a higher propensity to memorize noise and site-specific artifacts present in the multi-site ABIDE-I dataset rather than learning the generalizable spatiotemporal signatures of autism. Furthermore, very deep Transformers can suffer from optimization difficulties, such as vanishing gradients, which may prevent effective training. The performance was also checked concerning how the quantity of attention heads, which are the primary components of the multi-head attention mechanism allows the parallel processing of different input aspects, is affected. The number of heads from 1 to 16 was experimented with, as illustrated in Fig. 8. The performance fluctuated with the quantity of heads used; so, the model was susceptible to this parameter. In general, models that employed four attention heads were the best performers across the metrics. Indeed, the four-head model exceeded the accuracy of 1, 2, 8, and 16-head models by 1.73%, 1.31%, 2.02%, and 0.80%, respectively. With fewer heads (e.g., 1 or 2), the model likely suffers from expressive bottleneck, where the attention mechanism lacks the capacity to simultaneously focus on different aspects of the temporal dynamics, such as short-term fluctuations versus long-range trends in the BOLD signal. Conversely, when the number of heads is excessive (e.g., 8 or 16), the feature dimensionality per head becomes too small (4 or 2 dimensions, respectively), leading to attention collapse or degenerate attention.

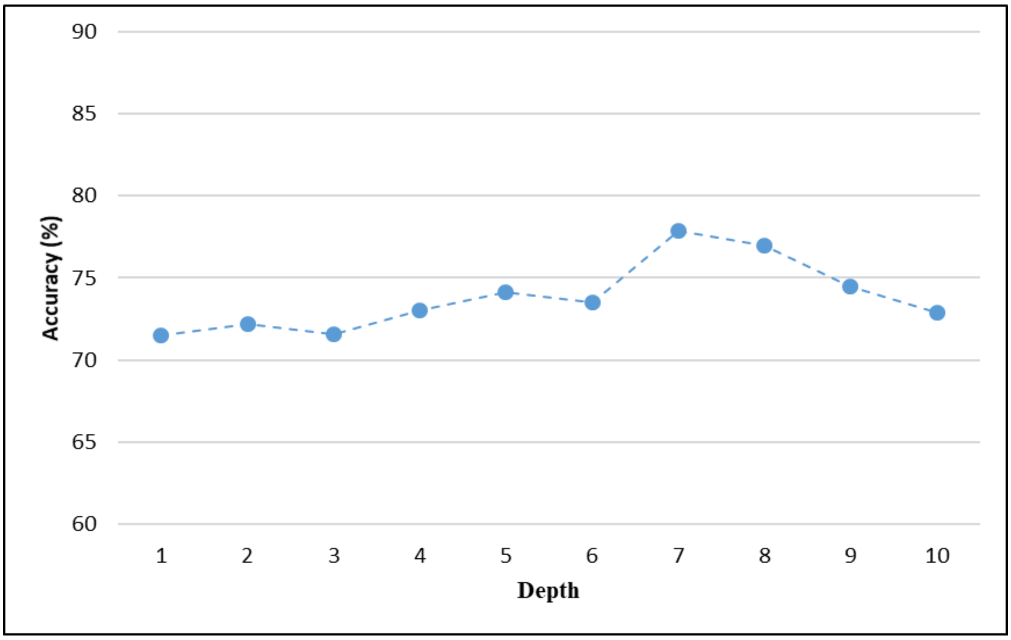

Figure 7: Performance of our framework with various depths of the Transformer encoder

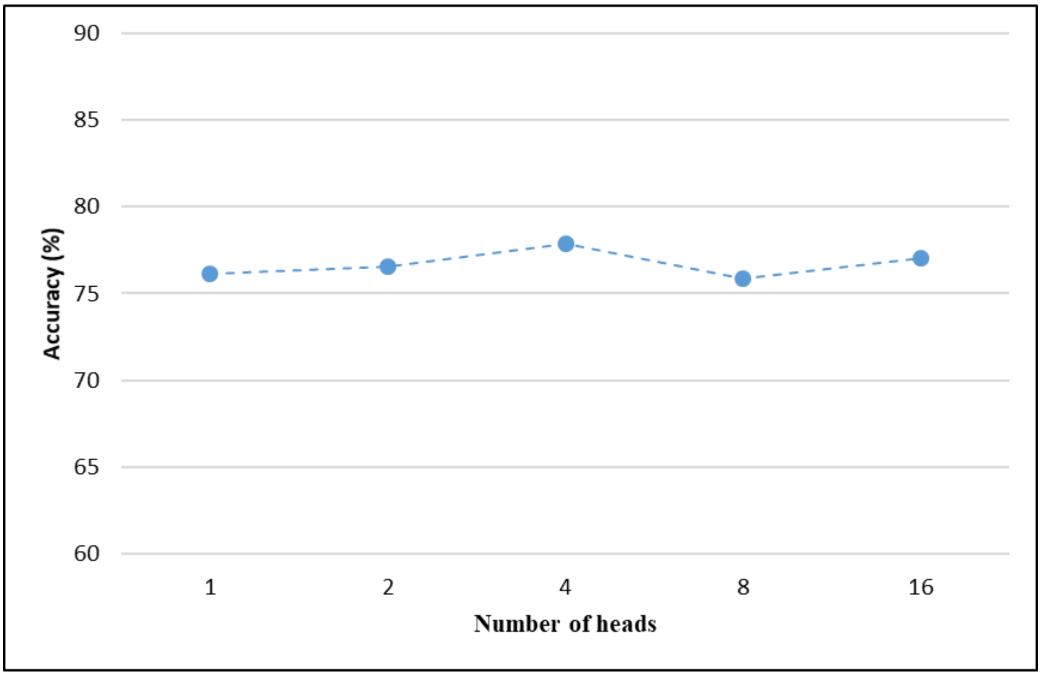

Figure 8: Performance of our framework with various numbers of heads of SA

During the second experimental phase, 5-fold cross-validation was independently conducted at each site. The resulting figures are presented in Table 6. The figure shows that the proposed framework's performance has decreased by 3.64% from phase 1. Performance metrics fluctuated quite a bit between different sites, with accuracies going from 63.55% (Trinity) to 96.42% (NYU), which points to strong site-specific effects. Although sites like NYU and OHSU performed excellently, sites like Yale and SBL, on the other hand, exhibited very low results. This indicates that the model's strength is contingent upon the attributes of the local data, thereby highlighting the problem of a universally consistent diagnostic tool.

Table 6: Classification performance through 5-fold cross-validation on each data site utilizing the suggested framework

Site	Accuracy (%)	Sensitivity (%)	Specificity (%)	F1-score (%)
Yale	68.31	64.85	70.00	67.32
USM	76.00	75.23	75.98	75.60
UM	70.25	69.54	72.67	71.07
UCLA	78.91	79.40	80.15	79.77
Trinity	63.55	61.33	66.37	63.75
Stanford	71.50	70.69	72.95	71.80
SDSU	69.81	69.14	69.74	69.4
SBL	64.22	62.10	65.10	63.56
PITT	75.73	73.48	76.44	74.93
OLIN	76.20	75.00	77.50	76.22
OHSU	92.00	91.50	92.62	92.05
NYU	96.42	95.63	97.29	96.45
MaxMun	65.17	64.19	67.00	65.56
Leuven	71.50	69.20	73.11	71.10
KKI	77.93	77.00	78.46	77.72
CMU	78.00	76.11	79.83	77.92
Caltech	66.10	65.72	67.21	66.45
Average	74.21	72.94	75.43	74.15

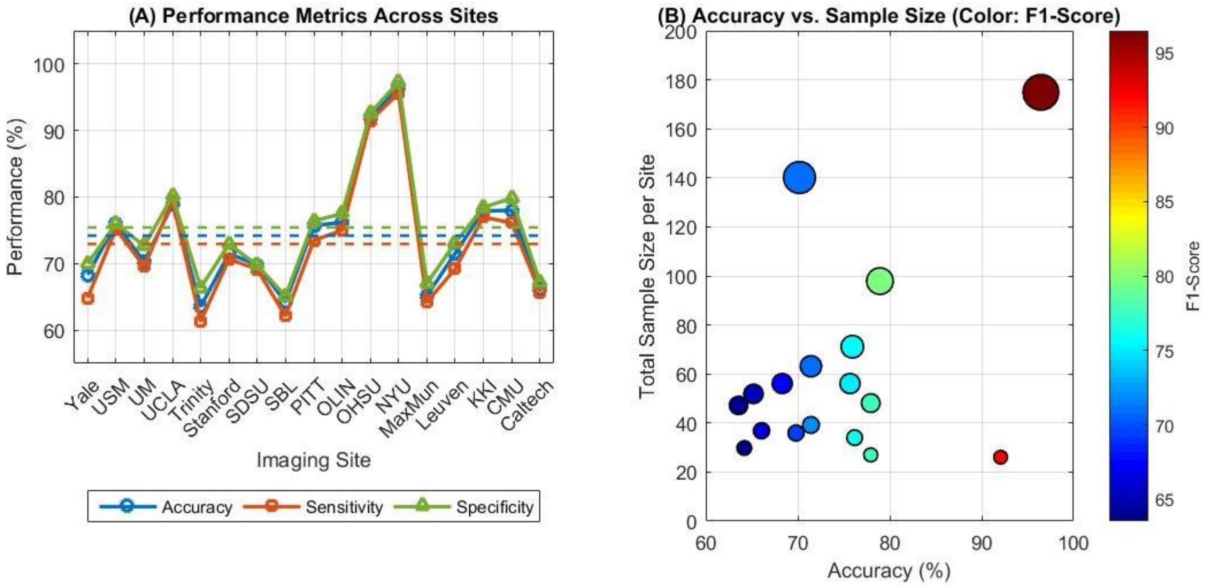

Figure 9: Site-wise performance variability and its relationship with sample size. (A) Classification accuracy, sensitivity, and specificity across the 17 imaging sites in the ABIDE-I dataset. The dashed lines represent the mean performance for each metric. Significant fluctuations highlight the impact of site-specific scanner parameters and protocols. (B) Bubble chart illustrating the correlation between a site's total sample size and classification accuracy. The size of each bubble corresponds to the sample size, and the color represents the F1-Score.

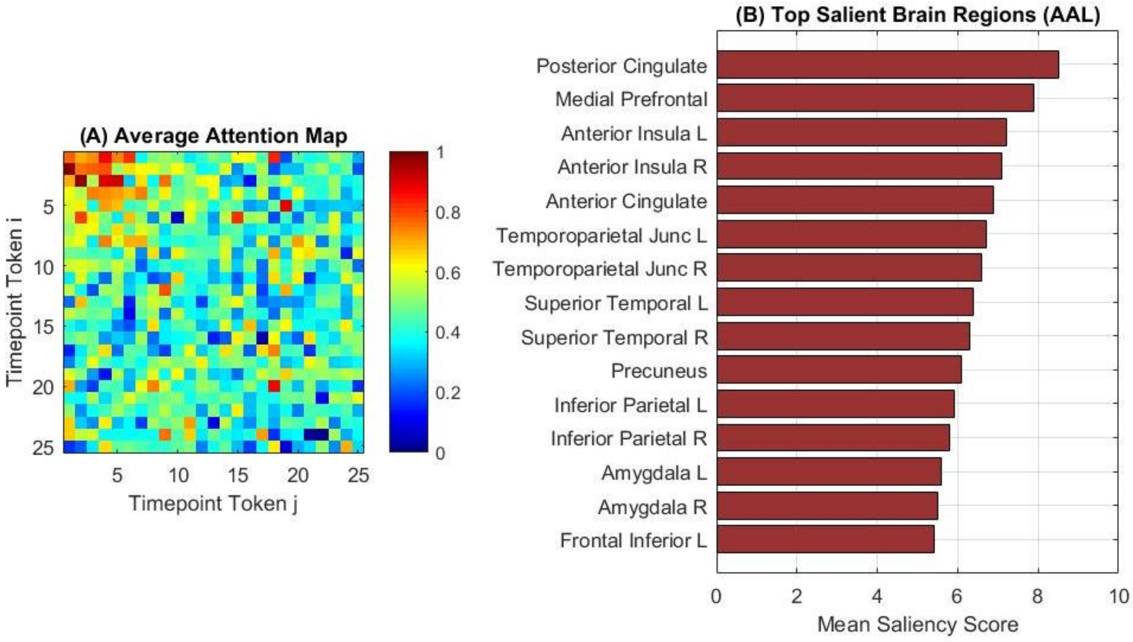

Figure 10: Interpretability analysis of the proposed hybrid model. (A) Average attention map across the test set, showing the relative importance between different timepoints in the fMRI sequence. (B) Saliency map highlighting the top 15 most influential brain regions (from the AAL atlas) for the model's classification decision. Node size corresponds to the mean saliency score, and edges represent strong functional connections between highly salient regions.

The superior performance of our model necessitates an investigation into its decision-making process to establish clinical trustworthiness. The interpretability analyses revealed the neurophysiologically plausible foundations of the model's predictions. The aggregated attention maps (Figure 10A) revealed that the model does not attend uniformly to all time points but identifies specific, transient intervals of high diagnostic importance. These critical periods often correspond to moments of significant BOLD signal fluctuation, suggesting the model leverages dynamic shifts in brain state rather than static average connectivity. More critically, the ROI saliency analysis (Figure 10B) identified a set of brain regions that consistently yielded high saliency scores. This set prominently included key nodes of established networks implicated in autism pathophysiology, such as the Default

Mode Network (e.g., Posterior Cingulate Cortex, Medial Prefrontal Cortex), the Salience Network (e.g., Anterior Insula, Anterior Cingulate Cortex), and regions involved in social cognition (e.g., Temporoparietal Junction, Superior Temporal Sulcus). This convergence between the model's learned features and the known neurobiology of autism strongly validates the clinical relevance of our approach and argues against the model latching onto artifactual, non-biological signals in the data. These interpretability results are a direct consequence of the hybrid, control-theoretic architecture. The CNN's role as a local state estimator efficiently preprocesses the high-dimensional input, isolating meaningful local neural patterns. The Transformer, acting as the optimal controller, then performs a global, context-aware integration of these patterns. It effectively computes the temporal dependencies and identifies which estimated states (CNN features) and at which time points (attention weights) are most predictive. This two-stage process is what enables the discovery of coherent, large-scale brain dynamics relevant to autism, a task that pure CNN or Transformer models struggle with due to their respective limitations in capturing global or local context. Therefore, the novelty of our framework lies not only in its performance but in its biologically-plausible, adaptive strategy for decoding complex brain network dynamics.

An extensive ablation study measured how much each core component contributed to our hybrid framework. The baseline model was a convolutional module (Section 2.2), followed directly by the classification head. The Transformer encoder and the residual connection were gradually added one by one, and the performance on the ABIDE-I dataset was measured utilizing 10-fold cross-validation. The outcomes shown in Table 7 are clear in that the proposed architecture achieves a notable performance gain. The baseline CNN-only model exhibited moderate performance, thus proving its capability to extract meaningful local spatiotemporal features. However, by adding the Transformer encoder, the performance improved dramatically as the accuracy increased by 4.69% and the Kappa value rose from 0.625 to 0.699. The enhancement in the result indicates the major role of the Transformer SA mechanism for modeling the global, long-range temporal dependencies in the fMRI sequence, which the CNN is incapable of. Therefore, the work incorporating the residual connection, which combined the local features from the CNN with the global context from the Transformer, resulted in the best outcome, as it could add an extra 1.92% in accuracy and reached a final Kappa of 0.734. This means that the local details facilitated by the residual connection provide the complementary information to the global context; thus, a more powerful and balanced feature representation for the final classification emerges. To statistically validate the improvement from each architectural addition, a paired Wilcoxon signed-rank test was performed on the accuracy distributions across the 10 folds for each model variant. This analysis showed that our full model is significantly better than other variants (P<0.05).

Table 7: Results of the ablation study on the ABIDE-I dataset

Model Variant	Accuracy (%)	Sensitivity (%)	Specificity (%)	F1-score (%)	Kappa	P-value
CNN only (Baseline)	71.24	69.88	72.45	71.11	0.625	0.008
CNN + Transformer Encoder	75.93	74.60	77.12	75.83	0.699	0.034
Full model	77.85	76.52	78.90	77.71	0.734	-

To empirically validate the advantage of the Transformer encoder over traditional recurrent models for capturing long-range temporal dependencies in fMRI data, we conducted a comparative study against two widely-used RNN variants: Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU). For a fair comparison, we replaced the Transformer encoder in our hybrid framework with equivalent LSTM and GRU modules while keeping the convolutional feature extractor and classification head identical. All models were trained and evaluated using the same 10-fold cross-validation protocol on the ABIDE-I dataset. The results, summarized in Table 8, demonstrate a clear performance hierarchy. The standard LSTM model achieved an accuracy of 73.15%, while the GRU performed slightly better at 74.08%. Both RNN variants were outperformed by our proposed CNN-Transformer hybrid, which achieved a significantly higher accuracy of 77.85% ($P < 0.05$, paired Wilcoxon test). This performance gap of 3.7-4.7% provides strong empirical support for the theoretical advantages of the self-attention mechanism.

Based on a review of previous publications, it can be inferred that most of the works have attempted to provide a diagnosis for autism through the analysis of the ABIDE-I fMRI dataset. To be precise, this research contrasts its performance with those earlier methods, whose effects give the main frame for interpreting the current findings. Our model, as reported in Table 9, has reached an accuracy of 77.85% which is 2.65% higher than the first-best accuracy figure reported in [24] for 1,035 samples. Besides that, our findings are better than those of [44] that claimed an accuracy of 74.53% for the 860-subjects cohort, even if the effect of sample size is not considered. By merging a CNN network with a Transformer encoder, the proposed method delivers at least 2.65% better results than other DL techniques. While this margin may appear modest, its practical significance is substantial given the high heterogeneity of the ABIDE-I dataset. Indeed, the practical significance of this margin must be contextualized within the challenges of neuroimaging-based diagnosis. This improvement equates to correctly classifying dozens more individuals in a large cohort and represents a meaningful step towards a more reliable biomarker. The concurrent increase in the F1-score further indicates a more balanced model, which is critical for a fair diagnostic tool.

Table 8: Performance comparison of the proposed CNN-Transformer hybrid with RNN-based variants on the ABIDE-I dataset.

Model	Accuracy (%)	Sensitivity (%)	Specificity (%)	F1-Score (%)	Kappa	P-Value vs. Proposed
CNN-LSTM	73.15	71.84	74.32	72.56	0.643	0.012
CNN-GRU	74.08	72.91	75.12	73.47	0.658	0.023
CNN-Transformer (Ours)	77.85	76.52	78.90	77.71	0.734	—

Table 9: Comparing the performance of our proposed technique with some state-of-the-art study in autism diagnosis through the fMRI ABIDE-I dataset

Reference	Number of subjects	Model	Accuracy (%)
[45]	871	SVC	66.80
[34]	1035	AE+DNN	70.00
[28]	871	GCN	70.40
[40]	1035	AE+SLP	70.30
[46]	872	Ensemble GCN	70.86
[47]	1035	CNN	70.22
[48]	1035	Extra-Trees	72.20
[49]	949	Ensemble MLP	74.52
[42]	1035	SAE+MLP	70.80
[44]	860	3D CNN	74.53
[24]	1035	SSDAE+MLP	75.20
Ours	1035	CNN+Transformer	77.85

Even though it yields promising results, this study has a few limits that required to be considered and that provide directions for further research. The model that was proposed performed differently to a great extent in various data acquisition sites, as shown in Table 6. This draws attention to the major problem of neuroimaging: site-specific variations in the scanner hardware, acquisition protocols, and participant demographics can all cause confounding effects. Our model has proved to be a good generalization; however, its use as a universal clinical instrument will depend on how these heterogeneity issues can be solved. Subsequent research will deal with incorporating advanced harmonization methods like ComBat or DL-based domain adaptation to eliminate site-specific biases and perform better. The second point is that the model's interpretability, which has somewhat resolved the investigation of the attention weights, still demands in-depth supplementary research. The more detailed analysis correlating the model features with the neurobiological circuits that are involved in autism (e.g., the default mode or salience networks) will help to convince the clinical sector that our findings are relevant. The investigation was limited to the ABIDE-I dataset only. It is necessary to verify the model on larger and more recent multi-site datasets, such as ABIDE-II and those containing different populations in terms of age and sex, to thoroughly evaluate the model's generalizability and its potential to be translated into real-world clinical practice.

4 Discussion

This study proposed a novel CNN-Transformer hybrid framework for the automated diagnosis of ASD from resting-state fMRI data. The results demonstrate that our model achieves state-of-the-art performance on the multi-site ABIDE-I dataset, outperforming a range of established deep learning models and pre-trained CNNs. The key success of this work lies not only in the reported metrics but in the architectural synergy that drives them, a point that merits detailed discussion in the context of prior art and the specific challenges of fMRI analysis.

4.1 Comparative analysis with state-of-the-art methods

Our model's accuracy of 77.85% represents a meaningful advancement over previous leading methods. When compared to other deep learning approaches on the ABIDE-I dataset, our hybrid framework shows a clear and consistent improvement. For instance, Heinsfeld et al. [34] employed a deep autoencoder followed by a neural network, reporting an accuracy of 70.0%. While effective in learning compact representations, their model likely lacks the explicit, hierarchical feature engineering for spatiotemporal data that our convolutional module provides. Similarly, Eslami et al. [40] (ASD-DiagNet, 70.3% accuracy) also used an autoencoder-based approach but may be limited in capturing the very long-range temporal dependencies present in fMRI sequences.

Graph-based methods, such as the Graph Convolutional Network (GCN) by Parisot et al. [28] (70.4% accuracy), excel at modeling the brain as a connectome of static functional connections. However, they often overlook the rich, dynamic temporal information within the BOLD signal itself. Our model's strength is its direct processing of the raw time series, allowing it to learn both the spatial relationships (via depth-wise convolution) and the complex, global temporal dynamics (via the Transformer) simultaneously. The most direct predecessor to our work is the SSDAE+MLP model by Liu et al. [24], which achieved 75.2% accuracy. Our 2.65% performance improvement can be attributed to the Transformer encoder's superior capability in modeling the entire temporal context compared to the stacked denoising autoencoder, which may struggle with long-range dependencies that are not local in time.

4.2 The added value of transformers over RNNs and CNNs

A central contribution of this work is the demonstration that the Transformer architecture is uniquely suited to address the specific limitations of previous models in fMRI analysis. Compared to Recurrent Neural Networks (RNNs) like LSTMs, which have been used for temporal modeling [31], the Transformer offers two distinct advantages for fMRI. First, the self-attention mechanism provides a global receptive field from the first layer, allowing any timepoint to directly influence any other. This is crucial for identifying non-local, transient brain states that are critical for ASD diagnosis but may be separated by many seconds in the scan. In contrast, LSTMs process data sequentially, making it difficult to learn dependencies between distant timepoints due to the vanishing gradient problem. Second, the Transformer's parallelizable architecture leads to more efficient training on modern hardware, unlike the sequential nature of RNNs.

When compared to standard CNNs, which are powerful local feature extractors [47], the Transformer compensates for their fundamental constraint: a limited receptive field. A CNN's ability to integrate information is bounded by the size of its kernel and the depth of its layers. While CNNs are excellent at identifying local temporal patterns and spatial relationships between adjacent brain regions, they are inherently poor at modeling the brain's global, system-wide dynamics that unfold over the entire scanning session. Our ablation study (Table 7) quantitatively confirms this, showing a significant performance jump when the Transformer encoder is added to the CNN baseline. The Transformer acts as a powerful global contextualizer, re-weighting and integrating the local features produced by the CNN to form a representation that is informed by the entire temporal history of the brain's activity.

4.3 Synthesis: The hybrid architecture as a synergistic solution

Therefore, the performance of our model is not the result of a single component but of their synergistic integration. The convolutional module acts as a dedicated, high-resolution feature engine, extracting meaningful local spatiotemporal patterns from the noisy, high-dimensional fMRI data. The Transformer encoder then serves as a sophisticated temporal reasoning module, identifying which of these local patterns are globally significant and how they interact across time to form a diagnostic signature. This division of labor—local feature extraction followed by global contextual modeling—proves to be a powerful paradigm. It is this hybrid design that allows our model to surpass the performance ceilings of architectures that rely solely on one approach, setting a new benchmark for fMRI-based ASD diagnosis.

4.4 Multi-site heterogeneity: Limitations and future directions with harmonization

A central finding of this work is the substantial performance variability observed across different imaging sites (Table 6, Figure 9), with accuracy ranging from 63.55% (Trinity) to 96.42% (NYU). This performance drop at specific sites underscores a fundamental challenge in neuroimaging-based machine learning: multi-site bias. This bias arises from differences in scanner manufacturers, acquisition protocols, head coils, and participant demographics across sites, which can introduce non-biological, site-specific variance that confounds the model's ability to learn generalizable features of ASD. While our hybrid architecture demonstrates a degree of inherent robustness by achieving a strong aggregate performance, the results confirm that architectural advances alone are insufficient to fully overcome this data-level challenge. The model's performance is strongly correlated with site-specific sample size (Figure 9B), suggesting that sites with larger, potentially more representative datasets allow the model to better learn to ignore site-specific noise. Conversely, smaller sites may not provide enough data for the model to disentangle the signal of ASD from the site-specific artifact. To directly address this limitation in future work, the application of statistical and deep learning-based harmonization techniques is essential. Methods such as ComBat could be employed as a preprocessing step. ComBat uses an empirical Bayes framework to adjust for site effects by standardizing the mean and variance of features (e.g., ROI time series or functional connectivity matrices) across sites, effectively removing scanner-specific biases while preserving biological variability associated with the condition.

Furthermore, while the results on ABIDE-I are compelling, the clinical translation of such a model hinges on its performance across independent datasets. As an immediate next step, we will conduct external validation on the ABIDE-II repository and initiate prospective clinical studies to assess real-world generalizability. We will also explore domain adaptation techniques to enhance cross-dataset robustness, ensuring the model's reliability beyond the specific characteristics of its initial training data.

5 Conclusion

A new CNN-Transformer hybrid framework is proposed in this article to identify ASD from rest-state fMRI data automatically. Our model is essentially designed to use the strength of convolutional networks for local feature extraction and transformer encoders to capture global dependencies. Through various metrics and experiments, our method on the ABIDE-I dataset has set a new state-of-the-art level, beating the existing techniques and pre-trained CNN models in multiple performance metrics. The ablation experiment reported unequivocal real-world results that coupling both architectural structures was necessary to achieve the highest performance, with the SA component being the most instrumental in capturing long-range temporal dynamics in BOLD signals. While site-related variability is an ongoing challenge, the model's robust performance indicates the considerable promise of hybrid DL architectures for unraveling intricate neuroimaging data. This work contributes a strong, generalizable framework that not only advances the field of fMRI-based autism diagnosis but also is a valuable reference for future research seeking to leverage Transformer-based models in computational neuroscience.

Authorship contribution statement

Mengzhu Yu: Project administration, Conceptualization, Supervision, Writing-Original draft preparation.

Conflicts of interest

The authors hereby state that they have no conflicts of interest pertaining to the publication of this manuscript.

Author statement

All authors have reviewed and supported the document. As previously indicated in this document, the criteria for authorship have been satisfied, and each author affirms that the manuscript constitutes a genuine and accurate representation of their work.

Ethical approval

All authors have been directly and actively engaged in significant contributions to the development of this manuscript and accept full public accountability for its content.

References

[1] R. M. Hernández, J. C. Ponce-Meza, M. Á. Saavedra-López, W. A. C. Ugaz, R. M. Chanduvi, and W. C. Monteza, "Brain complexity and psychiatric disorders," *Iran J Psychiatry*, vol. 18, no. 4, p. 493, 2023. https://doi.org/10.18502/ijps.v18i4.13637

[2] M. R. Mohammadi *et al.*, "Prevalence of autism and its comorbidities and the relationship with maternal psychopathology: a national population-based study," *Arch Iran Med*, vol. 22, no. 10, pp. 546–553, 2019. http://eprints.mui.ac.ir/id/eprint/11250

[3] J. Yang *et al.*, "Towards an accurate autism spectrum disorder diagnosis: multiple connectome views from fMRI data," *Cerebral Cortex*, vol. 34, no. 1, p. bhad477, 2024. https://doi.org/10.1093/cercor/bhad477

[4] M. R. Mohammadi *et al.*, "Prevalence and correlates of psychiatric disorders in a national survey of Iranian children and adolescents," *Iran J Psychiatry*, vol. 14, no. 1, p. 1, 2019. https://pmc.ncbi.nlm.nih.gov/articles/PMC6505051/

[5] M. KHARE, S. ACHARYA, S. SHUKLA, and A. SACHDEV, "Utilising Artificial Intelligence (AI) in the Diagnosis of Psychiatric Disorders: A Narrative Review.," *Journal of Clinical & Diagnostic Research*, vol. 18, no. 4, 2024. DOI: 10.7860/JCDR/2023/61698.19249

[6] N. Ghahari, F. Yousefian, S. Behzadi, and A. Jalilzadeh, "Rural-urban differences in age at autism diagnosis: a multiple model analysis," *Iran J Psychiatry*, vol. 17, no. 3, p. 294, 2022. https://doi.org/10.18502/ijps.v17i3.9729

[7] A. Khaleghi, H. Zarafshan, S. R. Vand, and M. R. Mohammadi, "Effects of non-invasive neurostimulation on autism spectrum disorder: a systematic review," *Clinical Psychopharmacology and Neuroscience*, vol. 18, no. 4, p. 527, 2020. https://doi.org/10.9758/cpn.2020.18.4.527

[8] S. Talepasand *et al.*, "Psychiatric disorders in children and adolescents: prevalence and sociodemographic correlates in Semnan Province in Iran," *Asian J Psychiatr*, vol. 40, pp. 9–14, 2019. https://doi.org/10.1016/j.ajp.2019.01.007

[10] N. Salari *et al.*, "The global prevalence of autism spectrum disorder: a comprehensive systematic review and meta-analysis," *Ital J Pediatr*, vol. 48, no. 1, p. 112, 2022. https://doi.org/10.1186/s13052-022-01310-w

[11] E. Helmy *et al.*, "Role of artificial intelligence for autism diagnosis using DTI and fMRI: A survey," *Biomedicines*, vol. 11, no. 7, p. 1858, 2023. https://doi.org/10.3390/biomedicines11071858

[12] A. Khaleghi, M. R. Mohammadi, G. P. Jahromi, and H. Zarafshan, "New ways to manage pandemics: using technologies in the era of COVID-19: a narrative review," *Iran J Psychiatry*, vol. 15, no. 3, p. 236, 2020. https://doi.org/10.18502/ijps.v15i3.3816

[13] H. Alkahtani, T. H. H. Aldhyani, and M. Y. Alzahrani, "Deep learning algorithms to identify autism spectrum disorder in children-based facial

landmarks," *Applied Sciences*, vol. 13, no. 8, p. 4855, 2023. https://doi.org/10.3390/app13084855

[14] D. Pandian, S. S. Rajagopalan, and D. Jayagopi, "Detecting a child's stimming behaviours for autism spectrum disorder diagnosis using rgbpose-slowfast network," in *2022 IEEE International Conference on Image Processing (ICIP)*, IEEE, 2022, pp. 3356–3360. https://doi.org/10.1109/ICIP46576.2022.9897867

[15] A. Zunino *et al.*, "Video gesture analysis for autism spectrum disorder detection," in *2018 24th international conference on pattern recognition (ICPR)*, IEEE, 2018, pp. 3421–3426. https://doi.org/10.1109/ICPR.2018.8545095

[16] A. Khaleghi, M. R. Mohammadi, K. Shahi, and A. M. Nasrabadi, "Computational neuroscience approach to psychiatry: a review on theory-driven approaches," *Clinical Psychopharmacology and Neuroscience*, vol. 20, no. 1, p. 26, 2022. https://doi.org/10.1109/ICPR.2018.8545095

[18] Z. Zhao *et al.*, "Conventional machine learning and deep learning in Alzheimer's disease diagnosis using neuroimaging: A review," *Front Comput Neurosci*, vol. 17, p. 1038636, 2023. https://doi.org/10.3389/fncom.2023.1038636

[19] B. Qiu, Q. Wang, X. Li, W. Li, W. Shao, and M. Wang, "Adaptive spatial-temporal neural network for ADHD identification using functional fMRI," *Front Neurosci*, vol. 18, p. 1394234, 2024. https://doi.org/10.3389/fnins.2024.1394234

[20] A. Chandra, S. Verma, A. S. Raghuvanshi, and N. K. Bodhey, "ASDC-Net: Optimized convolutional neural network-based automatic autism spectrum disorder classification using rs-fMRI Data," *IETE J Res*, vol. 70, no. 4, pp. 4189–4202, 2024. https://doi.org/10.1080/03772063.2023.2196979

[21] I. Chi, S. Tsai, C. Chen, and A. C. Yang, "Identifying Distinct Developmental Patterns of Brain Complexity in Autism: A Cross-Sectional Cohort Analysis Using the Autism Brain Imaging Data Exchange," *Psychiatry Clin Neurosci*, vol. 79, no. 3, pp. 98–107, 2025. https://doi.org/10.1111/pcn.13780

[22] S. Gupta *et al.*, "Enhancing autism spectrum disorder classification with lightweight quantized cnns and federated learning on abide-1 dataset," *Mathematics*, vol. 12, no. 18, p. 2886, 2024. https://doi.org/10.3390/math12182886

[23] S. Saponaro *et al.*, "Deep learning based joint fusion approach to exploit anatomical and functional brain information in autism spectrum disorders," *Brain Inform*, vol. 11, no. 1, p. 2, 2024. https://doi.org/10.1186/s40708-023-00217-4

[24] X. Liu, M. R. Hasan, T. Gedeon, and M. Z. Hossain, "MADE-for-ASD: A multi-atlas deep ensemble network for diagnosing autism spectrum disorder," *Comput Biol Med*, vol. 182, p. 109083, 2024. https://doi.org/10.1016/j.compbiomed.2024.109083

[25] R. R. Jha, A. Muralie, M. Daroch, A. Bhavsar, and A. Nigam, "Enhancing Autism Spectrum Disorder identification in multi-site MRI imaging: A multi-head cross-attention and multi-context approach for addressing variability in un-harmonized data," *Artif Intell Med*, vol. 157, p. 102998, 2024. https://doi.org/10.1016/j.artmed.2024.102998

[26] T. Iidaka, "Resting state functional magnetic resonance imaging and neural network classified autism and control," *Cortex*, vol. 63, pp. 55–67, 2015. https://doi.org/10.1016/j.cortex.2014.08.011

[27] M. Plitt, K. A. Barnes, and A. Martin, "Functional connectivity classification of autism identifies highly predictive brain features but falls short of biomarker standards," *Neuroimage Clin*, vol. 7, pp. 359–366, 2015. https://doi.org/10.1016/j.nicl.2014.12.013

[28] S. Parisot *et al.*, "Disease prediction using graph convolutional networks: application to autism spectrum disorder and Alzheimer's disease," *Med Image Anal*, vol. 48, pp. 117–130, 2018. https://doi.org/10.1016/j.media.2018.06.001

[29] B. Sen, N. C. Borle, R. Greiner, and M. R. G. Brown, "A general prediction model for the detection of ADHD and Autism using structural and functional MRI," *PLoS One*, vol. 13, no. 4, p. e0194856, 2018. https://doi.org/10.1371/journal.pone.0194856

[30] M. N. Parikh, H. Li, and L. He, "Enhancing diagnosis of autism with optimized machine learning models and personal characteristic data," *Front Comput Neurosci*, vol. 13, p. 9, 2019. https://doi.org/10.3389/fncom.2019.00009

[31] M. Z. Uddin, M. A. Shahriar, M. N. Mahamood, F. Alnajjar, M. I. Pramanik, and M. A. R. Ahad, "Deep learning with image-based autism spectrum disorder analysis: A systematic review," *Eng Appl Artif Intell*, vol. 127, p. 107185, 2024. https://doi.org/10.1016/j.engappai.2023.107185

[33] C. J. Brown, J. Kawahara, and G. Hamarneh, "Connectome priors in deep neural networks to predict autism," in *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*, IEEE, 2018, pp. 110–113. https://doi.org/10.1109/ISBI.2018.8363534

[34] A. S. Heinsfeld, A. R. Franco, R. C. Craddock, A. Buchweitz, and F. Meneguzzi, "Identification of autism spectrum disorder using deep learning and the ABIDE dataset," *Neuroimage Clin*, vol. 17, pp. 16–23, 2018. https://doi.org/10.1016/j.nicl.2017.08.017

[35] N. Qiang *et al.*, "A deep learning method for autism spectrum disorder identification based on interactions of hierarchical brain networks," *Behavioural Brain Research*, vol. 452, p. 114603, 2023. https://doi.org/10.1016/j.bbr.2023.114603

[36] F. Z. Subah, K. Deb, P. K. Dhar, and T. Koshiba, "A deep learning approach to predict autism spectrum disorder using multisite resting-state

fMRI," *Applied Sciences*, vol. 11, no. 8, p. 3636, 2021. https://doi.org/10.3390/app11083636

[37] F. Zhao *et al.*, "multi-head self-attention mechanism-based global feature learning model for ASD diagnosis," *Biomed Signal Process Control*, vol. 91, p. 106090, 2024. https://doi.org/10.1016/j.bspc.2024.106090

[38] S. Amemiya, H. Takao, and O. Abe, "Resting-State fMRI: Emerging Concepts for Future Clinical Application," *Journal of Magnetic Resonance Imaging*, vol. 59, no. 4, pp. 1135–1148, 2024. https://doi.org/10.1002/jmri.28894

[39] W. Wei *et al.*, "Analyzing 20 years of resting-state fMRI research: Trends and collaborative networks revealed," *Brain Res*, vol. 1822, p. 148634, 2024. https://doi.org/10.1016/j.brainres.2023.148634

[40] T. Eslami, V. Mirjalili, A. Fong, A. R. Laird, and F. Saeed, "ASD-DiagNet: a hybrid learning approach for detection of autism spectrum disorder using fMRI data," *Front Neuroinform*, vol. 13, p. 70, 2019. https://doi.org/10.3389/fninf.2019.00070

[41] Y. Ma *et al.*, "multi-scale dynamic graph learning for brain disorder detection with functional MRI," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 31, pp. 3501–3512, 2023. https://doi.org/10.1109/TNSRE.2023.3309847

[42] F. Almuqhim and F. Saeed, "ASD-SAENet: a sparse autoencoder, and deep-neural network model for detecting autism spectrum disorder (ASD) using fMRI data," *Front Comput Neurosci*, vol. 15, p. 654315, 2021. https://doi.org/10.3389/fncom.2021.654315

[43] H. Zhao, H. Lou, L. Yao, and Y. Zhang, "Diffusion transformer-augmented fMRI functional connectivity for enhanced autism spectrum disorder diagnosis," *J Neural Eng*, vol. 22, no. 1, p. 016044, 2025. DOI: 10.1088/1741-2552/adb07a

[44] J. Deng, M. R. Hasan, M. Mahmud, M. M. Hasan, K. A. Ahmed, and M. Z. Hossain, "Diagnosing autism spectrum disorder using ensemble 3D-CNN: A preliminary study," in *2022 IEEE international conference on image processing (ICIP)*, IEEE, 2022, pp. 3480–3484. https://doi.org/10.1109/ICIP46576.2022.9897628

[45] A. Abraham *et al.*, "Deriving reproducible biomarkers from multi-site resting-state data: An Autism-based example," *Neuroimage*, vol. 147, pp. 736–745, 2017. https://doi.org/10.1016/j.neuroimage.2016.10.045

[46] R. Anirudh and J. J. Thiagarajan, "Bootstrapping graph convolutional neural networks for autism spectrum disorder classification," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2019, pp. 3197–3201. https://doi.org/10.1109/ICASSP.2019.8683547

[47] Z. Sherkatghanad *et al.*, "Automated detection of autism spectrum disorder using a convolutional neural network," *Front Neurosci*, vol. 13, p. 1325, 2020. https://doi.org/10.3389/fnins.2019.01325

[48] Y. Liu, L. Xu, J. Li, J. Yu, and X. Yu, "Attentional connectivity-based prediction of autism using heterogeneous rs-fMRI data from CC200 atlas," *Exp Neurobiol*, vol. 29, no. 1, p. 27, 2020. https://doi.org/10.5607/en.2020.29.1.27

[49] Y. Wang, J. Wang, F.-X. Wu, R. Hayrat, and J. Liu, "AIMAFE: Autism spectrum disorder identification with multi-atlas deep feature representation and ensemble learning," *J Neurosci Methods*, vol. 343, p. 108840, 2020. https://doi.org/10.1016/j.jneumeth.2020.108840