frontiers in Neuroscience

ORIGINAL RESEARCH
published: 04 February 2022
doi: 10.3389/fnins.2021.828512

An Invertible Dynamic Graph Convolutional Network for Multi-Center ASD Classification

Yueying Chen[1,2], *Aiping Liu*[1,2]*, *Xueyang Fu*[1,2], *Jie Wen*[3] and *Xun Chen*[1,2]

[1] School of Information Science and Technology, University of Science and Technology of China, Hefei, China, [2] USTC IAT-Huami Joint Laboratory for Brain-Machine Intelligence, Institute of Advanced Technology, University of Science and Technology of China, Hefei, China, [3] Division of Life Sciences and Medicine, Department of Radiology, The First Affiliated Hospital of USTC (Anhui Provincial Hospital), University of Science and Technology of China, Hefei, China

Autism Spectrum Disorder (ASD) is one common developmental disorder with great variations in symptoms and severity, making the diagnosis of ASD a challenging task. Existing deep learning models using brain connectivity features to classify ASD still suffer from degraded performance for multi-center data due to limited feature representation ability and insufficient interpretability. Given that Graph Convolutional Network (GCN) has demonstrated superiority in learning discriminative representations of brain connectivity networks, in this paper, we propose an invertible dynamic GCN model to identify ASD and investigate the alterations of connectivity patterns associated with the disease. In order to select explainable features from the model, invertible blocks are introduced in the whole network, and we are able to reconstruct the input dynamic features from the network's output. A pre-screening of connectivity features is adopted to reduce the redundancy of the input information, and a fully-connected layer is added to perform classification. The experimental results on 867 subjects show that our proposed method achieves superior disease classification performance. It provides an interpretable deep learning model for brain connectivity analysis and is of great potential in studying brain-related disorders.

Keywords: fMRI, graph convolutional networks, invertible networks, brain connectivity networks, autism spectrum disorder, disease classification

OPEN ACCESS

Edited by:
Yu Zhang,
Lehigh University, United States

Reviewed by:
Qibin Zhao,
RIKEN, Japan
Yulun Zhang,
ETH Zürich, Switzerland

***Correspondence:**
Aiping Liu
aipingl@ustc.edu.cn

Specialty section:
This article was submitted to Brain Imaging Methods, a section of the journal Frontiers in Neuroscience

Received: *03 December 2021*
Accepted: *23 December 2021*
Published: *04 February 2022*

Citation:
Chen Y, Liu A, Fu X, Wen J and Chen X (2022) An Invertible Dynamic Graph Convolutional Network for Multi-Center ASD Classification. Front. Neurosci. 15:828512. doi: 10.3389/fnins.2021.828512

1. INTRODUCTION

As one of the most common neurodevelopmental disorders, the exact etiology of Autism Spectrum Disorder (ASD) remains unknown. In the past 50 years, ASD has gone from a narrowly defined, rare disorder of childhood to a well-publicized disease, and recognized as a very common and heritable brain disorder. The major characteristic of ASD is being deficit in social interaction and social communication with repetitive and unusual behaviors and activities (Lord et al., 2018). Despite medical progress, the diagnosis of ASD still depends on the symptom-based clinical criteria with complex diagnostic steps. However, with increasing recognition of the importance of early diagnosis for effective intervention, more effort has been made on exploring other possible modalities and biomarkers for ASD identification.

With the development of neuroimaging technologies, resting-state functional Magnetic Resonance Imaging (rs-fMRI) has attracted increasing interest in ASD studies, which enjoys advantages of superior spatial resolution to accurately locate the active areas in the whole brain, overcoming the limitations of earlier tools such as positron emission tomography (PET),

electroencephalography (EEG), and magnetoencephalography (MEG). By computing the correlation between fMRI time series of different regions of interests (ROIs), we can construct a functional connectivity network and many disorders may lead to the alterations in it (Li et al., 2016; Miller et al., 2016; Bachmann et al., 2018; Chandra et al., 2019; Zhang et al., 2020). For example, a widespread decrease of functional connectivity strengths was reported in patients with Alzheimer's Disease (AD) (Demirtaş et al., 2017). Studies showed that regional connectivity changes (both increase and decrease) of dopaminergic cortico-striatal and mesolimbic-striatal loops have been found in PD subjects (Filippi et al., 2018). ASD has also been suggested to be related to altered brain connectivity in the development of disease and has been extensively investigated (Kleinhans et al., 2008; Monk et al., 2009; Yerys et al., 2015; Dajani and Uddin, 2016; Xu et al., 2020). While a wide range of connectivity changes are reported, inconsistent conclusions have been observed in studies of functional connectivity in ASD, indicating the importance to thoroughly investigate the connectivity patterns with a large population of ASD.

Based on brain connectivity networks, machine learning, especially deep learning methods have further provided powerful tools to extract representative features associated with ASD and have greatly deepened our understanding of the disease (Chan et al., 2020). The classical machine learning techniques such as Support Vector Machines (SVM) are most widely used to identify patients from healthy controls in various studies (Subbaraju et al., 2017). For instance, Abraham et al. (2017) achieved 66.8% classification accuracy on 871 subjects obtained from ABIDE dataset.

Neural networks and deep learning methods such as autoencoder, Deep Neural Network (DNN) (Guo et al., 2017), Long Short Term Memory (LSTM) (Dvornek et al., 2018), and Convolutional Neural Network (CNN) (Haweel et al., 2021) have generated better performance in ASD classification. For instance, Yin et al. (2021) applied a DNN model and achieved the classification accuracy of 76.2% on 871 subjects of ABIDE dataset, and further improved the performance to an accuracy of 79.2% by combining DNN with an autoencoder.

Compared with traditional deep learning models, Graph Convolutional Network (GCN) can deal with data of non-Euclidean structure, which may be more suitable, and more interpretable for brain connectivity graph generated by fMRI. GCN has been used to classify ASD and select biomarkers from typical developing subjects (Ktena et al., 2018; Parisot et al., 2018). Recently, with a connectivity-based GCN model, 70.7% accuracy for classifying 1057 subjects (525 ASD and 532 healthy controls) has been reported (Wang et al., 2021). It's worth noting that when integrating information from more modalities, we may obtain higher classification accuracy. For instance, 85.06% of accuracy in ASD classification has been reported in Rakić et al. (2020) based on both structural MRI (sMRI) and fMRI features of 368 ASD and 449 healthy control subjects using an autoencoder model. While more modalities are beneficial to disease identification, it requires extra resources on data collection. In this paper, we are more interested in resting-state fMRI and focus on the ASD classification using brain connectivity features based on fMRI signals.

However, most deep learning models are limited in interpretation because of their black box representation. Although the classification performances of most deep learning networks are superior to those of traditional or interpretable methods, the features they finally generate can hardly be corresponded to the inputs, challenging the selection of helpful biomarkers. To overcome this shortcoming, Jacobsen et al. (2018) proposed an invertible network using a fully-connected layer as an inner trainable network, which can accurately reconstruct the inputs to a layer from its outputs without any degradation of classification accuracy. Given its superiority, Zhuang et al. (2019) proposed an invertible network for ASD classification, and gained 71% accuracy on the whole ABIDE dataset.

To improve the model interpretability and to better utilize structural, spatial, and temporal characteristics of brain connectivity networks, in this paper, we propose an invertible dynamic GCN (ID-GCN) model for ASD classification. More specifically, invertible blocks are utilized in the whole network, capable of reconstructing the input features from the output of the network, followed by a fully-connected layer to perform classification. Additionally, we select the connectivity features with a pre-screening operation to reduce the redundancy of the input information. The proposed method is verified on multi-center ABIDE datasets and the results demonstrate its effectiveness for disease classification and potential for studying the disease-related connectivity features. The contributions of this paper are summarized as:

- An invertible graph convolutional network is designed for disease classification based on brain connectivity networks. It is capable of generating disease-related interpretable connectivity features and improving classification accuracy.
- The proposed model integrates the structural, spatial, and dynamic information of the brain connectivity networks, and a prior selection of the features is adopted to reduce the redundancy of the input features.
- The proposed method has been validated on ABIDE dataset with superior performance.

2. METHODS

In this section, we first provide the notations and their definitions used in this paper, then we introduce our proposed invertible dynamic GCN model in detail.

2.1. Notations and Definitions

In this paper, we use $G(V, E)$ to represent a graph, where $V = \{v_1, v_2, ..., v_n\}$ is the set of nodes, and $E = \{e_{ij}\}$ is the set of edges. In the spatial connectivity graph, e_{ij} represents the Euclidean distance of two connected nodes, and in the functional connectivity graph, e_{ij} represents their connectivity strength. Additionally, let A denote the adjacency matrix of the graph and X denote the correlation matrix, in which every row represents a node's features.

2.2. Graph Convolutional Network

Graph Convolutional Network is a deep learning architecture, which can not only use the data itself but also the relationship between data represented as a graph. Through the adjacency matrix A of the graph, we can first calculate the normalized Laplacian matrix of X, which calculation formula is:

$$L = I - D^{-\frac{1}{2}} A D^{-\frac{1}{2}} \tag{1}$$

Where I is an identity matrix and D is the diagonal degree matrix of X. Then, we get an eigendecomposition of the Laplacian matrix, $L = U \Lambda U^T$, where U is a set of orthonormal eigenvectors, and $\Lambda = \text{diag}(\lambda_0, ..., \lambda_{n-1})$ is the matrix's non-negative eigenvalues. Based on these formulas, we get the propagation rule of graph convolution layers is:

$$X^l = \sigma(U \Theta(\Lambda) U^T X^{l-1}) \tag{2}$$

Where σ is the activation function of the layer, and $\Theta(\cdot)$ is the GCN convolution kernel. To simplify the calculation, we then fit the kernel by Chebyshev polynomials of order k (Hammond et al., 2011), which can be derived from:

$$T_k(c) = 2cT_{k-1}(c) - T_{k-2}(c) \tag{3}$$

$$T_0(c) = I, T_1(c) = c \tag{4}$$

And the fitting formula is:

$$\Theta(\Lambda) = \sum_{k=0}^{K-1} \beta_k T_k(\tilde{\Lambda}) \tag{5}$$

$$\tilde{\Lambda} = \frac{2}{\lambda_{max}} \Lambda - I \tag{6}$$

Where β_k is the weight coefficient of the kth Chebyshev polynomial, and λ_{max} is the max eigenvalue of the Laplacian matrix. Since the calculation of Chebyshev polynomials is performed only on eigenvectors Λ, it does not affect other matrix operations like doing eigendecomposition. So the Equation (2) can be expressed as:

$$X^l = \sigma((\sum_{k=0}^{K-1} \beta_k T_k(\tilde{L})) X^{l-1}) \tag{7}$$

Where $\tilde{L}$ is defined as $\tilde{L} = \frac{2}{\lambda_{max}} L - I$. Then we substitute the trainable weight matrix W for β_k, and get the final propagation rule of graph convolution layers as:

$$X^l = \sigma(\sum_{k=0}^{K-1} T_k(\tilde{L})(X^{l-1}) W) \tag{8}$$

2.3. Invertible Block

The architecture of the invertible block is shown in **Figure 1**, where the inputs are x_1 and x_2, and the outputs are denoted as z_1 and z_2. Those feature maps have the same shape, and φ and ω can be defined as any functions. In this model, we define φ and ω as independent GCN modules using different graphs as their inputs, which will be introduced in detail in the next section. In order to fully blend the advantages of the two GCN modules, the outputs of the first block y_1 and y_2, are then calculated to their average and half of their difference as z_1 and z_2. This invertible block can reconstruct the input from its output, where the forward pass and inverse are:

$$\begin{cases} y_1 = x_1 + \varphi(x_2) \\ y_2 = x_2 + \omega(y_1) \end{cases} \quad \begin{cases} z_1 = 0.5(y_1 + y_2) \\ z_2 = 0.5(y_2 - y_1) \end{cases} \tag{9}$$

$$\begin{cases} x_2 = y_2 - \omega(y_1) \\ x_1 = y_1 - \varphi(x_2) \end{cases} \quad \begin{cases} y_1 = z_1 - z_2 \\ y_2 = z_1 + z_2 \end{cases} \tag{10}$$

2.4. Invertible Dynamic GCN

In order to incorporate additional spatial and temporal characteristics of the brain functional connectivity network constructed by rs-fMRI data with better interpretability, we propose an invertible dynamic GCN (ID-GCN) model, which uses two different GCN as the function φ and ω in the invertible blocks to encode the functional connectivity graph and spatial connectivity graph of samples, respectively. The functional GCN, i.e., ω in the invertible block, uses the functional graph of each subject obtained by the correlation matrix. Meanwhile, the skeleton of the spatial graph is calculated directly according to the spatial distance between ROIs, and the connection weights are their correlation values. It is represented as φ for spatial GCN. The whole model includes three invertible blocks to extract explainable high dimensional features, and the inputs x_1 and x_2 of

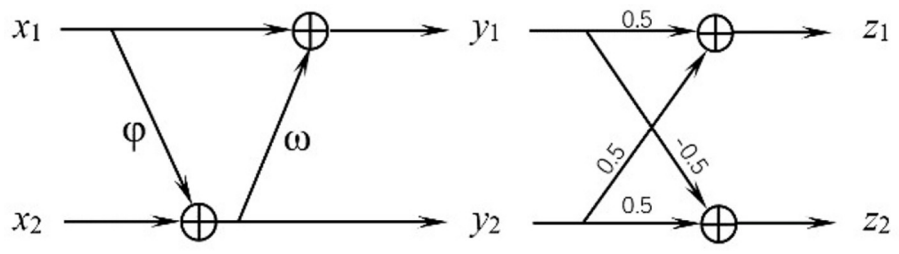

FIGURE 1 | Structure of the invertible block.

the first block are the same features that we send into the model. The proposed ID-GCN architecture for disease classification in this work is demonstrated in **Figure 2**.

To improve the computational efficiency and simplify the training process, for each node, the k connected nodes with the largest Pearson correlation coefficients in the functional graph or the smallest distance in the spatial graph are retained to construct a k-nearest graph. The correlation coefficients between each node and all other nodes are used as the sample's features which serve as input into the ID-GCN model. A fully connected layer with softmax is applied to perform the classification and the source of the collection site is included as an additional covariate. The cross-entropy is adopted in this model as loss function as:

$$L = \frac{1}{N}\sum_{i} -y_i * log(\hat{y}_i) - (1 - y_i) * log(1 - \hat{y}_i) \qquad (11)$$

where y_i is the label of the ith subject, $\hat{y}_i$ is the output of the network, and N is the number of subjects we use.

While there are usually hundreds of ROIs defined from the atlas, for a certain disease, it usually involves the changes of a portion of brain regions. Additionally, with a great individual variance of connectivity patterns, a large number of connectivity features may be easily disturbed by noise, affecting subsequent analysis and interpretation. However, reducing the number of ROIs in the input model may inevitably cause the loss of information. Therefore, rather than reducing the entire number of ROIs, we reduce the dimension of the input features of each ROI individually by selecting the M most important features for disease classification using random forest.

As our brains are a dynamic system, time-varying connectivity features have been suggested to be related to the functioning of our brain. Thus, in this model, we further utilize the dynamics of connectivity as additional features for ASD classification. The time sliding window is applied to sample the time-dependent signals and get the correlation matrix X_t of each time window. The temporal variations of dynamic connectivity are then calculated as the auxiliary feature represented as Ft, which is concatenated with other connectivity features. After the pre-selection of random forest, the reserved feature matrix $\{F_t\}$ is combined with the selected feature F of the original correlation matrix X as the final input features. The overview of the proposed model is shown in **Figure 3**.

3. EXPERIMENTS AND RESULTS

3.1. Real Dataset and Experimental Setting

We validated the proposed method on the publicly available ABIDE dataset (Martino et al., 2014), and chose 416 ASD subjects and 451 healthy controls (HC) from 13 acquisition sites. The phenotypical information of each acquisition site can refer to **Table 1**. The dataset was preprocessed with the Configurable Pipeline for the Analysis of Connectomes (C-PAC) (Sikka et al., 2013), which includes skull striping, slice timing correction, motion correction, global mean intensity normalization, nuisance signal regression, and band-pass filtering (0.01–0.1 Hz). The fMRI images were registered to the standard anatomical space (MNI152). To define brain areas, the Harvard Oxford (HO) atlas was chosen, consisting of 110 ROIs. More details of the dataset may refer to ABIDE Preprocessed.

We implemented the proposed model in a 5-fold cross-validation setting, using 80% of the data for training and 20% for testing. We set the pre-selected feature number M as 48, combined with $J = 10$ auxiliary dynamic features. Additionally, the Chebyshev polynomial order was chosen as 3, and $k = 3$ nearest nodes were selected to generate our graphs.

To test the proposed method, we compared it with other methods including siamese GCN (Ktena et al., 2018), Random Forest, SVM, and GCN, evaluating its performance improvement induced by the combination of spatial and dynamic connectivity features, and testing the effectiveness of pre-screening on the features. In these models for comparison, features input in siamese GCN is the paired subject features as implemented

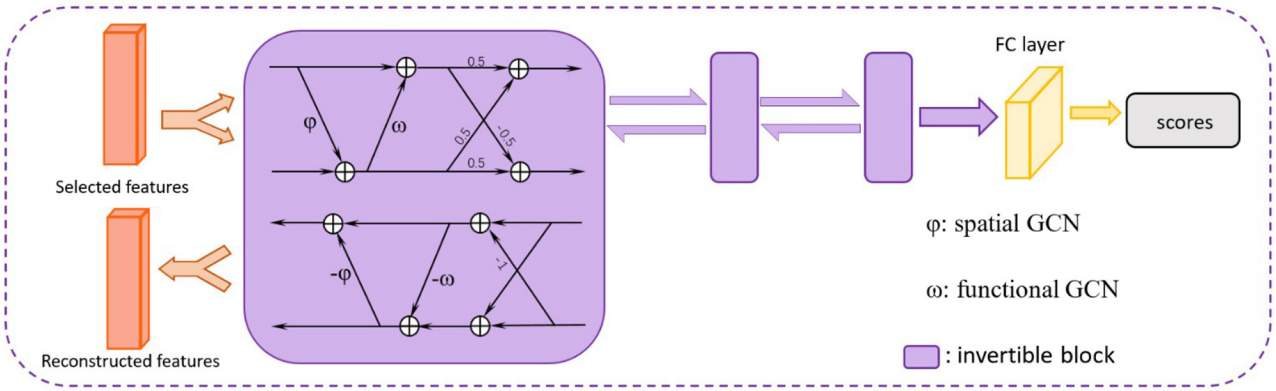

FIGURE 2 | The proposed ID-GCN architecture. The selected features are trained in three invertible blocks. A fully connected (FC) layer is finally used to obtain the output scores for ASD classification. The whole network is reversible before the FC layer, meaning that we can reconstruct the informative disease-related brain connectivity patterns by selecting important output features of the network.

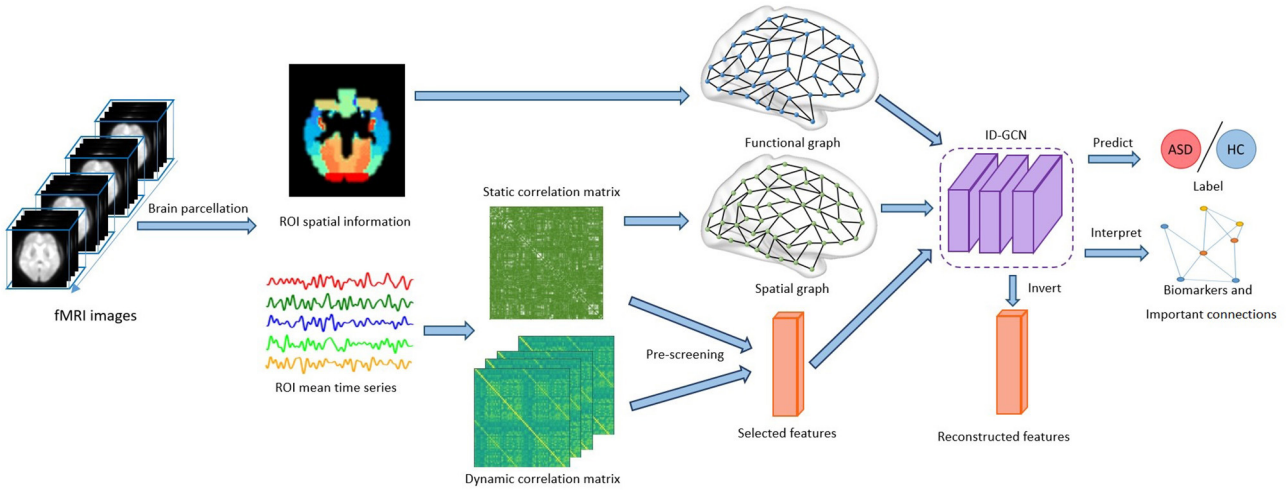

FIGURE 3 | Overview of the proposed framework. The brain connectivity features inferred from the fMRI time series and brain parcellation are fed to the model. After training using the ID-GCN model, we obtain the predictions for ASD classification, and important brain connectivity features are selected accordingly.

TABLE 1 | Phenotypical information summary of ABIDE data.

Site	ASD	HC	Gender (M/F)	Total	Age (mean±std)
PITT	30	27	49/8	57	18.9±6.8
TRINITY	24	25	49/0	49	17.2±3.6
UM_1	55	55	84/26	110	13.4±2.9
UM_2	13	22	33/2	35	16±3.3
USM	58	43	101/0	101	22.1±7.6
YALE	28	28	40/16	56	12.7±2.9
LEUVEN_1	14	15	29/0	29	22.6±3.5
LEUVEN_2	15	20	27/8	35	14.2±1.4
KKI	22	33	42/13	55	10.1±1.3
NYU	79	105	147/37	184	15.3±6.6
UCLA_1	41	32	63/10	73	13.2±2.4
UCLA_2	13	13	24/2	26	12.5±1.5
MAX_MUN	24	33	50/7	57	26.2±11.9
TOTAL	416	451	738/129	867	16.4±7.1

in the study (Ktena et al., 2018), while the other models use the whole connectivity matrix of a single subject as inputs. All the methods were evaluated in terms of accuracy, AUC value, precision, recall, and F1-score. The definitions of them are as follows:

$$Accuracy = (TP + TN)/n \tag{12}$$

$$Precision = TP/(TP + FP) \tag{13}$$

$$Recall = TP/(TP + FN) \tag{14}$$

$$F1 - score = 2 * Precision * Recall/(Precision + Recall) \tag{15}$$

where n is the total number of our subject, TP is true positive subject's number, TN is true negative, FP is false positive, FN denotes false negative, and AUC means the area under the ROC curve. We additionally performed ablation experiments to demonstrate the effects of each step of our method, including (1) GCN using the functional graph as input (GCN); (2) GCN using the spatial and functional graph in different layers (GCN adding spatial information); (3) ID-GCN with principal component analysis (PCA) for feature selection (ID-GCN with PCA); and (4) ID-GCN without dynamic features.

4. RESULTS

The classification results are shown in **Figure 4** and **Table 2**. It's noted that our proposed model, ID-GCN achieves the highest classification accuracy as 76.3%. Specifically, our model demonstrates great improvement in all the evaluation metrics compared with traditional SVM and Random Forest models and obtains 3.1% gains in accuracy compared with GCN using the same hyperparameters. Siamese GCN used paired subject features as input and generated classification results by multiplying two feature matrices from shared weight GCN. However, it's noticed that siamese GCN demonstrated worse performance on the given dataset where the paired features didn't successfully distinguish the subjects in this case.

Considering that the classification performance depends on the number of subjects, in order to have a fair comparison, we have tested our algorithm on the different number of subjects and show comparison with other state-of-the-art methods in **Table 3**. More specifically, we chose the number of subjects as 95, 459, 867, and 1,066, respectively. As they were examined on a different number of subjects, we didn't repeat their experiments but reported their datasets and results, only using same order of magnitude of subjects to run our model for better comparison. It can be seen that our results outperform other methods on the

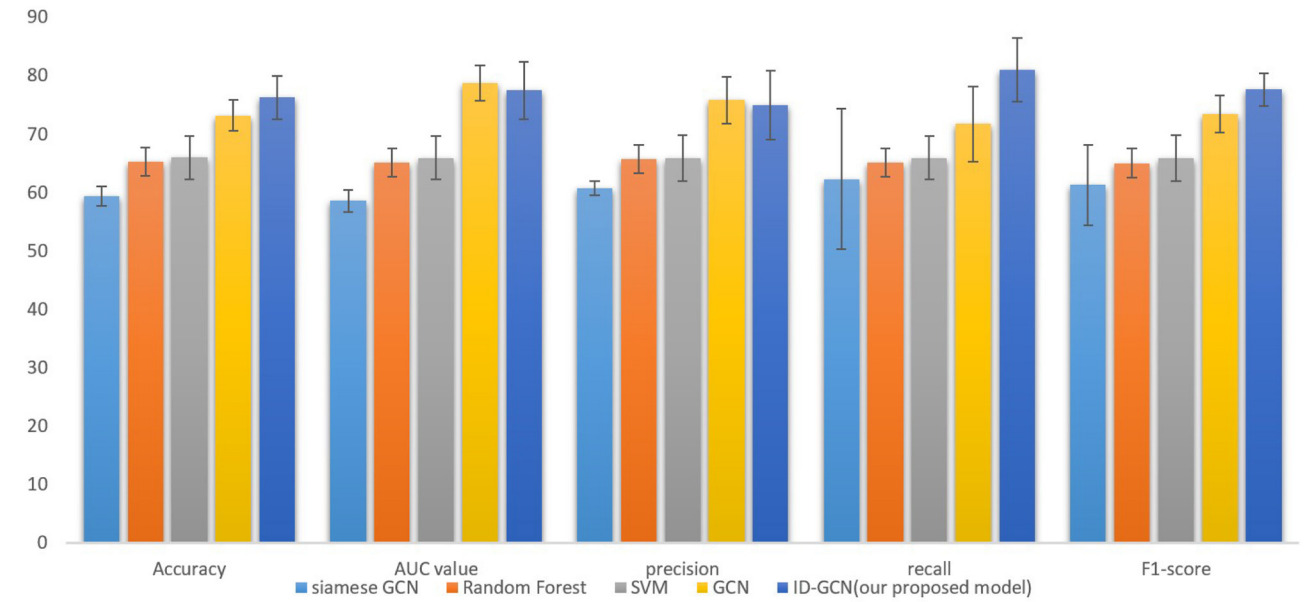

FIGURE 4 | Comparison with traditional and GCN models including siamese GCN (Ktena et al., 2018), Random Forest, SVM, and GCN.

TABLE 2 | Comparisons of different methods.

Model	Accuracy	AUC	Precision	Recall	F1-score
SVM	66.0±3.7%	65.9±3.7%	65.9±3.9%	65.9±3.7%	65.9±3.9%
Random forest	65.3±2.4%	65.1±2.4%	65.7±2.4%	65.1±2.4%	65.0±2.5%
GCN	73.2±2.7%	**78.7**±3.0%	**75.8**±4.0%	71.7±6.5%	73.4±3.2%
Siamises GCN	59.4±1.7%	58.6±1.9%	60.7±1.2%	62.3±12.0%	61.3±6.9%
ID-GCN(our model)	**76.3**±3.7%	77.5±4.9%	75±5.9%	**81.0**±5.5%	**77.6**±2.8%

TABLE 3 | Comparison with other SOTA methods.

Model	Number of subjects	Accuracy
DNN (Li et al., 2018)	95	85.3%
Combined MCNNEs (Aghdam et al., 2019)	459	70.45%
CNN-EW (Xing et al., 2018)	1096	66.88%
ASD-DiagNet (Eslami et al., 2019)	1035	70.1%
cGCN (Wang et al., 2021)	1057	70.7%
3D CNN (Thomas et al., 2020)	1162	64%
ID-GCN(our model)	95	**87.38**%
	459	**77.42**%
	867	**76.3**%
	1066	**71.44**%

same order of magnitude of data. It's worth noting that with data from different centers, the accuracy may vary. As demonstrated in **Table 3**, we can notice that more subjects do not guarantee better performance which is partially due to the great inter-center and inter-subject variability. When using 95 subjects from the same acquisition center, both (Li et al., 2018) and our method achieve high classification accuracy, and our proposed method obtains better classification performance compared with that of Li et al. (2018). Furthermore, the model performance of every single center is provided in **Table 4** that we train all the subjects and test the proposed method for each center separately. It shows that the classification accuracy varies across the centers, indicating great inter-center variability.

Additionally, the studies with multimodality data often demonstrate better performances using the same method. For example, Rakić et al. (2020) gained 85.06% of accuracy using both sMRI and fMRI features in the classification of 817 subjects. However, in this paper, we focus on the functional connectivity features. Although the proposed method has improved the classification accuracy compared with other GCN models and has interpretability, it still has several limitations. The temporal variations of brain connectivity have been utilized to represent the dynamics of brain connectivity. However, it's unable to fully delineate the time-varying connectivity. The classification accuracy of our interpretable model is limited compared with some networks without interpretability. For better performance, RNN model with temporal connectivity networks will be explored in our future work. Additionally, the biological interpretation of the biomarkers selected from our invertible network has been limited investigated. The effective

center-invariant biomarkers with sufficient biological meanings are warranted in future studies.

The results of ablation experiments are demonstrated in **Table 5**. It can be seen from the table that after adding spatial information as graph input, the accuracy of the model increased by over 1%, indicating the importance of the spatial information. As the number of connectivity features is large, great individual variation and noise may disturb the robust feature learning and degrade the classification performance. The feature selection, therefore, contributed to a significant improvement in the classification accuracy. We also evaluated other dimension reduction approach, i.e., Principal Component Analysis, for feature selection. As shown in **Table 5**, PCA led to less improvement in the classification accuracy. It may be due to the difficulty in the alignment of principal components across the subjects. Moreover, the temporal dynamics benefited the GCN model with a small accuracy gain.

In order to better understand ASD, we further identified the disease-related features by sorting the importance of each node's features extracted under the 5-fold cross-validation. The top 10% important connectivity edges were reconstructed as demonstrated in **Figure 5** and **Table 6**. It's noted that the connections between Right Pallidum and Right Inferior Frontal Gyrus, Left Frontal Orbital Cortex and Left Central Opercular Cortex, and connections involving Left Supramarginal Gyrus and Right Inferior Temporal Gyrus greatly contributed to the classification accuracy. Additionally, we evaluated the impacts of nodes by excluding each node and examining its influence on classification performance. With such lesion operation, we were able to assess the importance of each node. As shown in **Figure 6**, the highly-rated ROIs include Right Pallidum, Right Inferior Frontal Gyrus (triangle part), Right Inferior Temporal Gyrus (anterior division), Left Frontal Orbital Cortex, Left Temporal Fusiform Cortex (posterior division), and Right Temporal Occipital Fusiform Cortex, indicating their potential ROIs for ASD.

TABLE 4 | Model performance in each single center.

Site	Number of subjects	Accuracy
PITT	57	71.7±6.7%
TRINITY	49	72.0±11.7%
UM_1	110	75.5±3.6%
UM_2	35	82.9±10.7%
USM	101	84.8±7.0%
YALE	56	80.0±6.7%
LEUVEN_1	29	73.3±8.3%
LEUVEN_2	35	74.3±10.7%
KKI	55	74.5±8.9%
NYU	184	76.2±5.2%
UCLA_1	73	74.7±8.8%
UCLA_2	26	83.3±18.2%
MAX_MUN	57	66.6±11.9%
TOTAL	867	76.3±3.7%

TABLE 5 | Ablation study on the effects of different components.

Model	Accuracy
GCN	73.2%
GCN adding spatial information	74.5%
ID-GCN with PCA	74.2%
ID-GCN without dynamic features	76.1%
ID-GCN(our model)	**76.3**%

5. DISCUSSION AND CONCLUSION

The early diagnosis of ASD is a challenging task as great variations exist in the symptoms. In addition to the clinical criterion, researchers have tried to identify the effective neuroimaging biomarkers for the better diagnosis of ASD. Brain connectivity features are promising for studying ASD as

TABLE 6 | Important connectivity edges selected by feature reconstruction.

ROI1	ROI2
Right Pallidum	Right Inferior Frontal Gyrus
Left Frontal Orbital Cortex	Left Central Opercular Cortex
Left Temporal Fusiform Cortex (posterior division)	Left Heschl's Gyrus (includes H1 and H2)
Left Supramarginal Gyrus (anterior division)	Right Temporal Occipital Fusiform Cortex
Left Supramarginal Gyrus (posterior division)	Left Frontal Orbital Cortex
Right Inferior Temporal Gyrus (anterior division)	Left Supramarginal Gyrus (anterior division)
Right Inferior Temporal Gyrus (anterior division)	Left Lateral Occipital Cortex (inferior division)

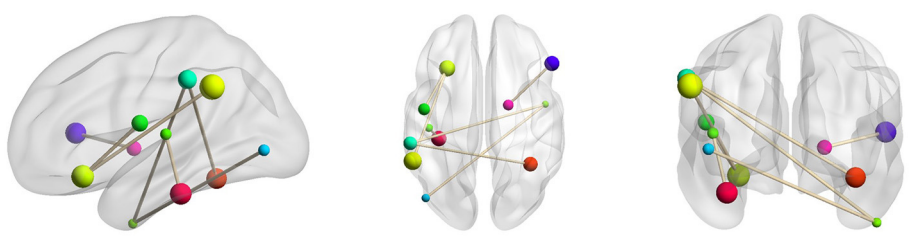

FIGURE 5 | Selected key connectivity features for ASD classification.

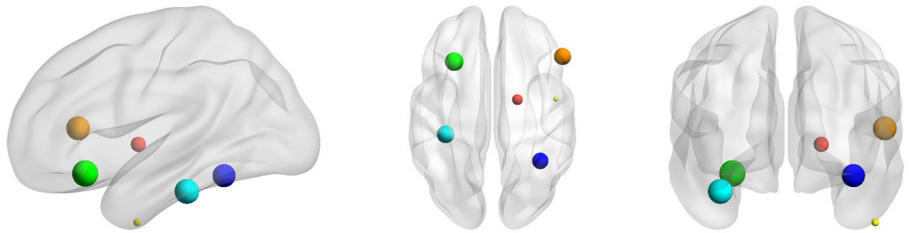

FIGURE 6 | Selected key ROIs for ASD classification, including Right Pallidum (red), Right Inferior Frontal Gyrus (triangle part) (orange), Right Inferior Temporal Gyrus (anterior division) (yellow), Left Frontal Orbital Cortex (green), Left Temporal Fusiform Cortex (posterior division) (cyan), and Right Temporal Occipital Fusiform Cortex (blue).

TABLE 7 | The classification accuracy with different k.

The value of k	2	3	4	5	6	8	10	15	20
Accuracy	73.7%	**76.3**%	75.1%	76.0%	75.0%	75.0%	75.8%	74.7%	75.3%

TABLE 8 | The classification accuracy with different M.

The value of M	10	30	48	50	70	90	110
Accuracy	72.1%	73.6%	**76.3**%	76.0%	75.7%	74.6%	73.9%

widespread connectivity changes have been observed in ASD. With various statistical and machine learning methods, we have largely expanded our understanding of the disease. However, the classification performance based on brain connectivity features is still limited, partially due to the insufficient representation ability for multi-center ASD data. It's, therefore, critical to learn the robust connectivity features for better representing the disease population. While the deep learning-based methods are promising, mos of them are designed in a black-box principle, challenging their biological interpretability.

In this study, we propose an explainable graph convolutional network, namely ID-GCN for multi-center ASD data classification and investigation by incorporating the functional, spatial and temporal information of the connectivity networks and using the invertible network to select interpretable biomarkers. The use of GCN aims to integrate the high-dimensional features of each node, and the invertible network is capable of reconstructing the extracted disease-related features back to the original connectivity graph. The proposed model contains two different GCN for brain functional connectivity and spatial connectivity, respectively. A random forest is adopted to narrow the feature space and reduce the redundancy of the data. We further integrate the dynamics of brain connectivity as important features for ASD classification. The experimental results on ABIDE dataset suggest the efficacy of our model. It is a potential classifier for large multi-center datasets despite their variations.

When classifying the ASD subjects, several connectivity features reconstructed by the model are assigned with higher importance. Those connections involve Right Pallidum, Right Inferior Frontal Gyrus, Left Frontal Orbital Cortex, Left Central Opercular Cortex, Left Temporal Fusiform Cortex, Right Temporal Occipital Fusiform Cortex, Left Supramarginal Gyrus and Right Inferior Temporal Gyrus, which are mostly consistent with the prior studies. For instance, the altered connectivity of Temporal Pole, Pallidum, and Frontal Orbital Cortex in ASD has been reported in Yerys et al. (2015); Dajani and Uddin (2016); Monk et al. (2009). In another line of studies, the changes of connectivity patterns in Fusiform Gyrus and Inferior Frontal Gyrus have been investigated for ASD subjects (Kleinhans et al., 2008; Xu et al., 2020). We additionally performed lesion analysis that sequentially removed each ROI and examined its impact on the classification accuracy. According to their contributions to the classification performance, eight ROIs including Right Superior Temporal Gyrus, Right Superior Frontal Gyrus, Right Pallidum, Right Inferior Frontal Gyrus (triangle part), Right Inferior Temporal Gyrus (anterior division), Left Frontal Orbital Cortex, Left Temporal Fusiform Cortex (posterior division), and Right Temporal Occipital Fusiform Cortex were chosen which are mostly involved in the connectivity features reconstructed by ID-GCN. It further substantiates the explainable features learned by the proposed method.

There are several parameters that need to be determined in the proposed model, and we have evaluated the impacts of different parameters on classification performance. **Table 7** demonstrates the classification accuracy as a function of the numbers of neighbors. It's observed that classification performance depends on the values of k, and when k=3, we obtained the highest classification accuracy. It indicates that there may be only a few connected areas that are most robust across the subjects. We have also chosen the number of features M using the grid search in **Table 8**, and when M=48, it achieved the best performance. If the number of M is too small or too large, the performance of the model will decline greatly.

While the proposed method is capable to identify the disease-related features and achieves a competitive classification

performance, it still has several limitations. The temporal variations of brain connectivity have been utilized to represent the dynamics of brain connectivity. However, it's unable to fully delineate the time-varying connectivity patterns, which can be further extended in our future work. The classification accuracy of our interpretable model is limited compared with some recent networks without interpretable modules. To further improve the performance, RNN models with temporal connectivity networks can be potential. Additionally, the biological interpretation of the biomarkers selected from our invertible network has been limited investigated. The effective center-invariant biomarkers with sufficient biological meanings are warranted in future studies.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

YC, AL, and XF worked on the method and analyzed the data. JW interpreted the results. AL and XC supervised the project. The manuscript was drafted by YC and AL. All the authors have reviewed and revised the manuscript.

FUNDING

This study was supported in part by the National Natural Science Foundation of China (Grants 61922075, 61701158 and 22077116), and in part by the USTC Research Funds of the Double First-Class Initiative (Grants YD2100002004 and YD9110002011).

ACKNOWLEDGMENTS

The authors thank the publicly available dataset provider and the professors of the University of Science and Technology of China.

REFERENCES

Abraham, A., Milham, M. P., Martino, A. D., Craddock, R. C., Samaras, D., Thirion, B., et al. (2017). Deriving reproducible biomarkers from multi-site resting-state data: an autism-based example. *NeuroImage* 147, 736–745. doi: 10.1016/j.neuroimage.2016.10.045

Aghdam, M. A., Sharifi, A., and Pedram, M. M. (2019). Diagnosis of autism spectrum disorders in young children based on resting-state functional magnetic resonance imaging data using convolutional neural networks. *J. Digit. Imag.* 32, 899–918. doi: 10.1007/s10278-019-00196-1

Bachmann, K., Lam, A. P., Sörös, P., Kanat, M., Hoxhaj, E., Matthies, S., et al. (2018). Effects of mindfulness and psychoeducation on working memory in adult adhd: A randomised, controlled fmri study. *Behav. Res. Therapy* 106, 47–56. doi: 10.1016/j.brat.2018.05.002

Chan, H.-P., Samala, R. K., Hadjiiski, L. M., and Zhou, C. (2020). Deep learning in medical image analysis. *Adv. Exp. Med. Biol.* 1213, 3–21. doi: 10.1007/978-3-030-33128-3_1

Chandra, A., Dervenoulas, G., Politis, M., and Initiative, A. D. N. (2019). Magnetic resonance imaging in alzheimer's disease and mild cognitive impairment. *J. Neurol.* 266, 1293–1302. doi: 10.1007/s00415-018-9016-3

Dajani, D. R., and Uddin, L. Q. (2016). Local brain connectivity across development in autism spectrum disorder: a cross-sectional investigation. *Autism Res.* 9, 43–54. doi: 10.1002/aur.1494

Demirtaş, M., Falcon, C., Tucholka, A., Gispert, J. D., Molinuevo, J. L., and Deco, G. (2017). A whole-brain computational modeling approach to explain the alterations in resting-state functional connectivity during progression of alzheimer's disease. *NeuroImage Clin.* 16, 343–354. doi: 10.1016/j.nicl.2017.08.006

Dvornek, N. C., Yang, D., Ventola, P., and Duncan, J. S. (2018). "Learning generalizable recurrent neural networks from small task-fmri datasets," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, Vol. 11, eds A. F. Frangi, J. A. Schnabel, C. Davatzikos, C. Alberola-López, and G. Fichtinger (Cham: Springer International Publishing), 329–337.

Eslami, T., Mirjalili, V., Fong, A., Laird, A. R., and Saeed, F. (2019). Asd-diagnet: a hybrid learning approach for detection of autism spectrum disorder using fmri data. *Front. Neuroinformat.* 13:70. doi: 10.3389/fninf.2019.00070

Filippi, M., Elisabetta, S., Piramide, N., and Agosta, F. (2018). Functional mri in idiopathic parkinson's disease. *Int. Rev. Neurobiol.* 141, 439–467. doi: 10.1016/bs.irn.2018.08.005

Guo, X., Dominick, K. C., Minai, A. A., Li, H., Erickson, C. A., and Lu, L. J. (2017). Diagnosing autism spectrum disorder from brain resting-state functional connectivity patterns using a deep neural network with a novel feature selection method. *Front. Neurosci.* 11:460. doi: 10.3389/fnins.2017.00460

Hammond, D. K., Vandergheynst, P., and Gribonval, R. (2011). Wavelets on graphs via spectral graph theory. *Appl. Comput. Harm. Anal.* 30, 129–150. doi: 10.1016/j.acha.2010.04.005

Haweel, R., Shalaby, A., Mahmoud, A., Seada, N., Ghoniemy, S., Ghazal, M., et al. (2021). A robust dwt-cnn-based cad system for early diagnosis of autism using task-based fmri. *Med. Phys.* 48, 2315–2326. doi: 10.1002/mp.14692

Jacobsen, J.-H., Smeulders, A., and Oyallon, E. (2018). "i-revnet: deep invertible networks," in *International Conference on Learning Representations*, Vancouver, BC.

Kleinhans, N. M., Richards, T., Sterling, L., Stegbauer, K. C., Mahurin, R., Johnson, L. C., et al. (2008). Abnormal functional connectivity in autism spectrum disorders during face processing. *Brain* 131, 1000–1012. doi: 10.1093/brain/awm334

Ktena, S. I., Parisot, S., Ferrante, E., Rajchl, M., Lee, M., Glocker, B., et al. (2018). Metric learning with spectral graph convolutions on brain connectivity networks. *NeuroImage* 169, 431–442. doi: 10.1016/j.neuroimage.2017.12.052

Li, X., Dvornek, N. C., Zhuang, J., Ventola, P., and Duncan, J. S. (2018). "Brain biomarker interpretation in asd using deep learning and fmri," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Vol. 11072 (Granada: Springer International Publishing), 206–214.

Li, Y., Liang, P., Jia, X., and Li, K. (2016). Abnormal regional homogeneity in parkinson's disease: a resting state fmri study. *Clin. Radiol.* 71, e28–e34. doi: 10.1016/j.crad.2015.10.006

Lord, C., Elsabbagh, M., Baird, G., and Veenstra-Vanderweele, J. (2018). Autism spectrum disorder. *The Lancet* 392, 508–520. doi: 10.1016/S0140-6736(18)31129-2

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

Miller, R. L., Yaesoubi, M., and Calhoun, V. D. (2016). Cross-frequency rs-fmri network connectivity patterns manifest differently for schizophrenia patients and healthy controls. *IEEE Signal Process. Lett.* 23, 1076–1080. doi: 10.1109/LSP.2016.2585182

Monk, C. S., Peltier, S. J., Wiggins, J. L., Weng, S.-J., Carrasco, M., Risi, S., et al. (2009). Abnormalities of intrinsic functional connectivity in autism spectrum disorders. *NeuroImage* 47, 764–772. doi: 10.1016/j.neuroimage.2009.04.069

Parisot, S., Ktena, S. I., Ferrante, E., Lee, M., Guerrero, R., Glocker, B., et al. (2018). Disease prediction using graph convolutional networks: application to autism

spectrum disorder and alzheimer's disease. *Med. Image Anal.* 48, 117–130. doi: 10.1016/j.media.2018.06.001

Rakić, M., Cabezas, M., Kushibar, K., Oliver, A., and Lladó, X. (2020). Improving the detection of autism spectrum disorder by combining structural and functional mri information. *NeuroImage Clin.* 25:102181. doi: 10.1016/j.nicl.2020.102181

Sikka, S., Cheung, B., Khanuja, R., Ghosh, S., Gan Yan, C., Li, Q., et al. (2013). Towards automated analysis of connectomes: the configurable pipeline for the analysis of connectomes (c-pac). *Front. Neuroinf.* 7. doi: 10.3389/conf.fninf.2013.09.00042

Subbaraju, V., Suresh, M. B., Sundaram, S., and Narasimhan, S. (2017). Identifying differences in brain activities and an accurate detection of autism spectrum disorder using resting state functional-magnetic resonance imaging : a spatial filtering approach. *Med. Image Anal.* 35, 375–389. doi: 10.1016/j.media.2016.08.003

Thomas, R. M., Gallo, S., Cerliani, L., Zhutovsky, P., El-Gazzar, A., and van Wingen, G. (2020). Classifying autism spectrum disorder using the temporal statistics of resting-state functional mri data with 3d convolutional neural networks. *Front. Psychiatry* 11:440. doi: 10.3389/fpsyt.2020.00440

Wang, L., Li, K., and Hu, X. P. (2021). Graph convolutional network for fmri analysis based on connectivity neighborhood. *Netw. Neurosci. (Cambridge, Mass.)* 5, 83–95. doi: 10.1162/netn_a_00171

Xing, X., Ji, J., and Yao, Y. (2018). "Convolutional neural network with element-wise filters to extract hierarchical topological features for brain networks," in *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Madrid, 780–783.

Xu, J., Wang, C., Xu, Z., Li, T., Chen, F., Chen, K., et al. (2020). Specific functional connectivity patterns of middle temporal gyrus subregions in children and adults with autism spectrum disorder. *Autism Res.* 13, 410–422. doi: 10.1002/aur.2239

Yerys, B. E., Gordon, E. M., Abrams, D. N., Satterthwaite, T. D., Weinblatt, R., Jankowski, K. F., et al. (2015). Default mode network segregation and social deficits in autism spectrum disorder: evidence from non-medicated children. *NeuroImage Clin.* 9, 223–232. doi: 10.1016/j.nicl.2015.07.018

Yin, W., Mostafa, S., and Wu, F.-X. (2021). Diagnosis of autism spectrum disorder based on functional brain networks with deep learning. *J. Comput. Biol.* 28, 146–165. doi: 10.1089/cmb.2020.0252

Zhang, Z., Peng, P., and Zhang, D. (2020). Executive function in high-functioning autism spectrum disorder: a meta-analysis of fmri studies. *J. Autism Develop. Disor.* 50, 4022–4038. doi: 10.1007/s10803-020-04461-z

Zhuang, J., Dvornek, N. C., Li, X., Ventola, P., and Duncan, J. S. (2019). "Invertible network for classification and biomarker selection for asd," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, Vol. 11766 (Shenzhen), 700–708.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.