Medical Image Analysis (2023)

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Identifying autism spectrum disorder based on individual-aware down-sampling and multi-modal learning

Li Pan[a], Jundong Liu[b], Mingqin Shi[d], Chi Wah Wong[e], Kei Hang Katie Chan[b,c,f,*]

[a]*Centre for Perceptual and Interactive Intelligence, Chinese University of Hong Kong, Hong Kong SAR, China*
[b]*Department of Biomedical Sciences, City University of Hong Kong, Hong Kong SAR, China*
[c]*Department of Electrical Engineering, City University of Hong Kong, Hong Kong SAR, China*
[d]*School of Basic Medical Sciences, Yunnan University of Traditional Chinese Medicine, Kunming, China*
[e]*Department of Applied AI and Data Science, City of Hope, Duarte CA 91010, USA*
[f]*Department of Epidemiology, Brown University, Providence RI 02912, USA*

ARTICLE INFO

Article history:
Received 1 May 2013
Received in final form 10 May 2013
Accepted 13 May 2013
Available online 15 May 2013

Keywords:
Autism spectrum disorder
Brain networking
fMRI
Computer-assisted diagnosis

ABSTRACT

Autism Spectrum Disorder(ASD) is a set of neurodevelopmental conditions that affect patients' social abilities. In recent years, many studies have employed deep learning to diagnose this brain dysfunction through functional MRI (fMRI). However, existing approaches mainly focused on the abnormal brain functional connections but ignored the impact of regional activities. Due to this biased prior knowledge, previous diagnosis models suffered from inter-site measurement heterogeneity and inter-individual phenotypic differences. To address this issue, we propose a simple unsupervised downsampling method for fMRI that can perform a personalized lower-resolution representation of the entire brain networking regarding both the functional connections and regional activities. Specifically, we abstract the brain imaging as a graph structure and straightforwardly downsample it to substructures by self-attention graph pooling. To further recalibrate the distribution of the extracted features under phenotypic information, we subsequently embed the sparse feature vectors into a population graph, where the hidden inter-subject heterogeneity and homogeneity are explicitly expressed as inter- and intra-community connectivity differences, and utilize Graph Convolutional Networks to learn the node embeddings. By these means, our framework can extract features directly and efficiently from the entire fMRI and be aware of implicit inter-individual variance. We have evaluated our framework on the ABIDE-I dataset with 10-fold cross-validation. The present model has achieved a mean classification accuracy of 86.45% and a mean AUC of 0.93, better than the state-of-the-art methods. The source code is available at https://github.com/jhonP-Li/ASD_GP_GCN.

1. Introduction

Autism spectrum disorder (ASD), a range of brain developmental disorders, has commonly been studied worldwide. In 2020, **?** reported that approximately 1 in 45 children in the U.S. was diagnosed with this disease caused by both genetic and environmental factors. This mental disorganization, which will result in difficulties with social interaction and communication, can be noticed at an early age of a child. However, another study, conducted in the U.K., showed that the current time-consuming diagnosis process could lead to a delay of around 3.5 years from the point at which parents first consult a doctor to the confirmation of an ASD diagnosis **?**, which results in unnecessary panic and delayed intervention.

Similar to physical disease diagnosis, this brain dysfunction can be detected with pathological manifestations. In **?**, the authors discovered that there existed structural differences in certain areas in the brain between autism patients and control subjects. **?** reported that abnormal brain function connections were found in ASD subjects. However, the study evaluated only a few samples, therefore these diagnosis methods cannot be generalized. In recent years, brain imaging analysis based on deep learning and machine learning, tested on large datasets, has been widely studied. **?** employed Long Short-Term Memory (LSTM) to analyze the time-series data of fMRI automatically. **?** applied Deep Neural Networks (DNN) directly on the fMRI and reported performance improvement compared to Support

*Corresponding author: Tel.: +852 3442 6661;
e-mail: kkhchan@cityu.edu.hk (Kei Hang Katie Chan)

Vector Machine (SVM) and Random Forest (RF). In **?**, the authors designed a Convolutional Neural Networks (CNN) architecture with fMRI as input and achieved slightly better performance than the DNN. **?** reached the best performance of end-to-end CNN models by 3D-CNN and the ensemble brain atlas.

Limited by the ability to process structured brain networking, the naive end-to-end implementations of CNN models have reached the bottleneck. On the one side, the functional connections of brain regions do not follow the spatial distributions of the areas, e.g., a node may interact with another node far away from it **?**. However, the convolution kernels can only extract features from spatial neighborhoods for each pixel. To be aware of those cross-space connections, CNN needs more convolution layers to form a wider receptive field, which backfires to overfitting due to the lack of samples. On the other side, to downsample the raw inputs, many methods have focused on selecting a certain number of functional connections. These methods represent brain imaging as a correlation coefficient matrix of which the elements denote the covariances between every two regions based on their time-series signals (**???????**). Specifically, in **?**, the authors elaborately constructed a workflow to extract features from functional connections and achieved state-of-the-art performance on the ABIDE I dataset using a linear classifier. However, the existing feature selection methods mainly extract features from the pairwise regional correlation matrix. This inflexible Euclidean representation of brain imaging did not only ignore the details of regional signals, which are believed to relate to ASD (**????**), but also omit the latent interaction among the connections.

To leverage the functional connections and regional activities, the brain imaging needs to be described as a graph structure, which perspicuously expresses the functional interactivity among regions as edges among nodes. Some studies have employed this non-Euclidean form to simulate brain networking and discover group-level brain biomarkers of ASD (**??????**), e.g., **?** constructed a personalized brain connectivity graph for each individual and measured the inter-individual graph structural difference using graph convolutional networks. **?** utilized graph convolutional networks on the geometric representation of the brain and tried to interpret the learned connections as ASD biomarkers. To find the abnormal brain networks that may interact with ASD, those graph convolutional methods have barely downscaled the input brain imaging, unlike the present graph embedding method. In other words, they focused on directly analyzing the graph-level information but did not extract higher-order features from it like the above selection methods, which makes the methods susceptible to inter-individual differences and even temporary brain activities **??**. Hence, limited by the surfeit model complexity and individual brain differences, the ASD identification accuracy of those models is not promising, making the inference of biomarkers unconvincing.

In the context of this ASD disease prediction problem, another challenge is the non-imaging difference between individuals, i.e., gender, handedness, IQ, and so on. Though this information is not present in the fMRI, it does affect the probability of suffering from ASD. For example, **?** indicated that one in every 42 males and one in 189 females in the United States is diagnosed with an autism spectrum disorder. **?** reported the correlation between the handedness and ASD. Besides, the fMRI scanning devices and measurement parameters from different data collection sites are not strictly the same **?**. Those hidden factors have caused the non-identity distribution of features and lowered the generalization ability of models. To address this, some authors manually forced those settings to be the same by hard clustering strategy on samples. For example, in (**????**), the authors elaborately selected training and testing samples from a certain data collection site. Hence, the implicit differences among samples were further narrowed, and these methods achieved much better classification performance than the models evaluated on the entire dataset. Although this hard clustering strategy did prove the feasibility of ASD diagnosis based on deep learning, the generalization ability of the models cannot be guaranteed as the number of samples has been dramatically reduced in that way.

To address the above issue, Graph Convolutional Networks (GCN) can be adopted to recalibrate the features extracted from brain imaging, according to the non-imaging data. Unlike assigning each subject into a cluster, we embed each subject into a population graph, where nodes denote individuals and edges represent the phenotypic similarity between every two nodes. Thus, the hidden inter-subject heterogeneity and homogeneity are explicitly expressed as inter- and intra-community connectivity differences. Some methods succeeded in fusing imaging and text information in this way but achieved lower classification accuracy than unimodal methods owing to inefficient brain imaging feature extraction (**???**). Namely, in **?**, the authors employed 3D CNN to extract features from fMRI and Variational Autoencoder to extract features from MRI, which are not efficient as previously discussed. Although this framework has considered functional, structural, and phenotypic information, it achieved lower performance than the method that directly utilized 3D CNN on fMRI **?**.

In this study, we propose a novel framework that incorporates self-attention graph pooling and graph convolutional networks. We explicit graph pooling to downsample the structured form of brain imaging, whereas previous brain modeling methods mainly extract features from functional connections. Unlike existing graph-level analysis, we individually downsample and flatten the brain imaging to sparse vectors, and implement Multilayer Perceptron to extract higher-level information from the selected subgraphs. To fuse the imaging and non-imaging information, we then initialize a population graph, where nodes represent individuals and edges denote phenotypic similarities. Assigning each subject with the extracted brain imaging features, we employ Graph Convolutional Networks to learn the node embeddings on the population graph, which succeed in regularizing the individual features according to phenotypic property. Having merged functional connections, regional activities, and non-imaging data, our framework presents its superior in ASD diagnosis, reaching an accuracy of 86.45% on ABIDE I dataset. The main contributions of our work are four-fold:

1) We have developed a self-attention downsampling method on fMRI. The unsupervised graph pooling efficiently downsamples the non-Euclidean brain networking. By

Table 1. Overview of the ABIDE I dataset preprocessed by CPAC

Sites	Age(year)		Gender		handedness				Diagnostic group	
	Min	Max	Male	Female	Left	Right	Ambi*	Mixed	ASD	Control
CALTECH	17.0	56.2	10	5	1	13	1	0	5	10
CMU	19.0	33.0	7	4	1	10	0	0	6	5
KKI	8.2	12.8	24	9	1	27	0	5	12	21
LEUVEN	12.1	32.0	49	7	7	49	0	0	26	30
MAX_MUN	7.0	58.0	42	4	2	44	0	0	19	27
NYU	6.5	39.1	136	36			N/A		74	98
OHSU	8.0	15.2	25	0	1	24	0	0	12	13
OLIN	10.0	24.0	23	5	5	23	0	0	14	14
PITT†	9.3	35.2	43	7	4	45	0	0	24	26
SBL†	20.0	49.0	26	0	1	0	0	0	12	14
SDSU	8.7	17.2	21	6	2	25	0	0	8	19
STANFORD	7.5	12.9	18	7	3	20	2	0	12	13
TRINITY	12.0	25.7	44	0	0	44	0	0	19	25
UCLA	8.4	17.9	74	11	9	76	0	0	48	37
UM†	8.2	28.8	93	27	15	97	0	0	47	73
USM	8.8	50.2	67	0			N/A		43	24
YALE	7.0	17.8	25	16	7	34	0	0	22	19
Total	6.5	58.0	727	144	59	531	3	5	403	468

* Ambi: Ambidextrous
† The handedness information of some subjects is unavailable.

weighing both the functional connections and regional signals simultaneously, this downsampling method retains the crucial information for diagnostic classification.

2) We have fused the multi-modal information through graph convolutional networks and intuitively illustrated its efficiency.
3) Different from hard selecting universal biomarkers of ASD, our framework tends to select a personalized substructure of brain networking for each individual. This novel strategy has detected inter-group heterogeneity and intra-group homogeneity regarding brain activities.
4) We have constructed an ASD diagnosing framework, which outperforms state-of-the-art methods on the ABIDE I dataset, reaching a classification accuracy of 86.45%. This clinically meaningful method could contribute to early detection and intervention for ASD.

The rest of the paper is organized as follows: Section 2 introduces the datasets and illustrates the details of the proposed model. In Section 3, we present the experimental setup, evaluation metrics, experimental results and comparison with other methods, ablation study, and intuitive exhibition of model mechanisms. In section 4, we discussed the limitation of the current model and future works. Finally, we draw the conclusion in Section 5.

2. Materials and Methods

2.1. ABIDE Dataset

Constructed by **?**, the ABIDE I dataset contains a variety of information of 1,112 subjects, i.e., MRI. fMRI, and phenotype data, collected from 17 international sites. To reduce the fMRI measurement error, current studies are focusing on the preprocessed data. **?** performed four different preprocessing pipelines on the original material. To compare with other methods (**?????**), in the current paper, we have used the data preprocessed by the Configurable Pipeline for the Analysis of Connectomes (CPAC). Built by **?**, the chosen functional preprocessing pipeline includes time slicing, motion correction, skull-stripping, global mean intensity normalization, and nuisance signal regression. Thus, the noise caused by unrelated motions, like the heart beating, is reduced. To further regularize the input sample features, we have also employed band-pass filtering and global signal regression.

As shown in the table 1, the processed data contains 403 ASD subjects and 468 typical control subjects. Caused by the measurement difference among different sites, some categories of the phenotype data are not or partially collected, like handedness information. Moreover, the distribution of the dataset is unbalanced on some features. For example, the 17 sites only have collected 144 female samples but collected 727 male samples. According to the view of **?**, gender is a rather important factor affecting the probability of ASD. This unbalanced feature distribution, which is not present in the MRI or fMRI data, has caused the non-identical distribution of features and thus affected the performance of unimodal learning models.

To further reduce the dimensionality of input data, fMRI is separated logically as signals of regions of interest (ROIs). The voxel-wise time series is thus paraphrased as the time series of regional signals. Proposed by (**????**), the Harvard-Oxford atlas is split into cortical and subcortical structural probabilistic atlases. The HO atlas, which has also been selected by other works (**???**), is filtered with a 25% threshold and subsequently

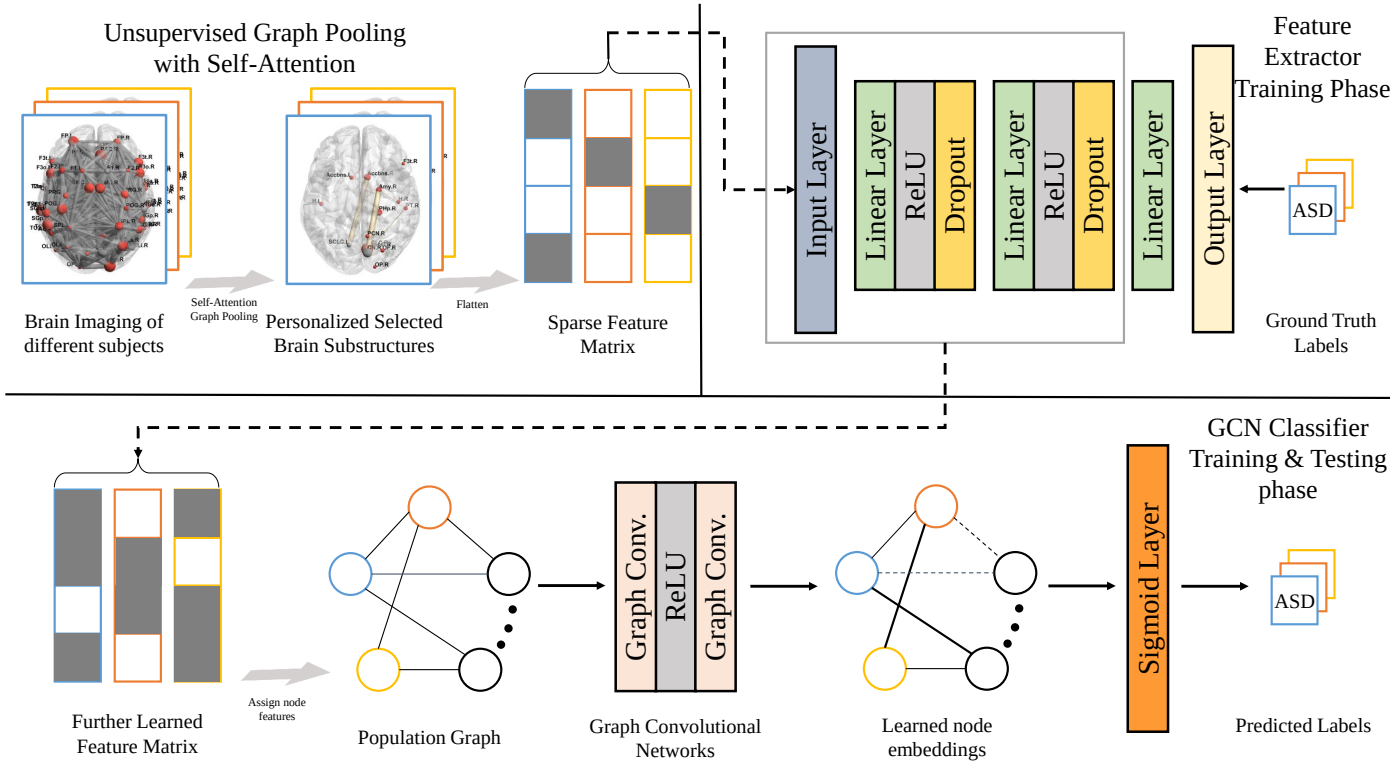


Fig. 1. Overview of the proposed framework; The top-left part illustrates self-attention graph pooling in section 2.3. The top-right part indicates the training phase of feature extractor, where we train the MLP using the pooling output. In the bottom part, we construct a population graph, where nodes denote subjects and edges represent interindividual phenotypic similarity. We then train a GCN model by assigning each node with dense features extracted by MLP.

divided into left and right hemispheres at the midline. The ROIs represent 110 functional brain regions, i.e., left and right Hippocampus, left and right Cuneal Cortex, left and right Planum Temporale, left and right Occipital Pole, etc. Thus, the original 4D brain imaging is further downsampled to a 2D data structure, containing 110 regions and the corresponding time series for each area.

2.2. *Model Overview*

As shown in the figure 1, the whole pipeline consists of three main parts. First, the unsupervised graph pooling directly downsamples the graph representation of the brain to a sparse brain networking. We then train a multilayer perception using the flattened features and ground truth labels to extract higher-order features from the pooling results. Finally, we employ a two-layer graph convolutional network to learn the node embeddings by embedding every individual into the population graph and building edges according to the phenotypic information.

2.3. *Graph Pooling*

In this section, we develop a self-attention unsupervised graph pooling strategy inspired by **?** to select the crucial subgraphs of brain networking for each subject individually. It improves the existing brain imaging feature extractors in two main aspects. First, it can directly downsample the entire graph structure, while other methods mainly target functional connections. Second, this graph pooling operation downsamples graph without supervision. In other words, this step can be directly added to other related frameworks without any additional training cost.

Graph pooling, as a downsampling method for graph structure, is a central component of graph convolutional networks in graph classification tasks. For example, the intuitive idea is to average all node embeddings to represent the entire graph **?**. Compared to other graph pooling methods (**???**), the proposed graph pooling procedure is designed to preserve the information and connectivity of the graph and reduce the information redundancy. As shown in figure 4, this strategy can automatically select key brain subgraph without strong assumption about the brain networking.

Before this implementation of graph pooling, other models trained classifiers, like MLP, or constructed a specific feature extracting framework to extract features from the functional connections (**????**) which are represented as the correlation coefficients of every two regions. At this phase, those methods have directly ignored the details of regional brain activities. This intuitive method, mapping the two vectors to a float ranging from -1 to 1, has reduced the data complexity. However, regarding the time series of brain regions as node features and functional connections as edges, this edge-only feature extraction strategy has caused much more information loss to the entire graph structure. Moreover, studies have proved

the importance of the node features in the diagnosis of ASD. In (**????**), the authors reported the abnormal regional activities among ASD subjects. Thus, leveraging the information of both functional connections and regional activities, graph pooling has proved its superior in this brain disorder diagnosis as shown in section 3.3 and figure 3.

2.3.1. Graph Representation of Brain Imaging

After being labeled as 110 regions according to the HO atlases, fMRI can be abstracted as a graph structure, where nodes denote brain regions and edges indicate functional connections. Initially, every node is assigned with a feature vector that represents the time series of regional activity. In **?**, the authors defined 6105 brain functional connections by connecting right-side regions to the left side and left-side regions to the right. Inspired by it, we construct a graph representation of brain imaging, where regions are connected according to the same strategy and all regions are connected with the global mean time series to reduce measurement error further. In short, the input brain graph structure contains 111 nodes and 6215 edges. In figure 4, we have also tested the graph pooling with other brain networking initialization.

2.3.2. Node Selection

The self-attention graph pooling consists of two parts: First, it selects the nodes based on the criterion of minimizing graph information loss. Subsequently, to connect the probably isolated subgraph caused by the node selection and recorrect the initial brain regional connections to some degree, an unsupervised edge prediction method is employed between the two-hop neighbors of each node and itself. At the first component, a node information score is defined as the $\mathcal{L}_1$ norm of the Manhattan distance between the node features itself and the one constructed from its neighbors **?**:

$$S \;=\; \gamma(g) \;=\; \left\|(I-(D^{(l)})^{-1}A^{(l)})H^{(l)}\right\|_1 \tag{1}$$

where $A^{(l)}$ and $H^{(l)}$ are the adjacency and node features matrices of the l-th layer. The information of edges is present implicitly as the connections among nodes. I represents the identity matrix and $D^{(l)}$ denotes the l-th layer diagonal degree matrix of $A^{(l)}$. $\|\cdot\|_1$ performs the $\mathcal{L}_1$ norm row-wisely. The vector S contains the information score of each node, which indicates its importance at this selection stage. The nodes are then selected by ranking and selecting the top-K ones regarding the information score:

$$\begin{aligned} idx \;&=\; top(S, \lceil r * n^{(l)} \rceil) \\ H^{(l+1)} \;&=\; H^{(l)}(idx,:) \\ A^{(l+1)} \;&=\; A^{(l)}(idx,idx) \end{aligned} \tag{2}$$

where r is the pooling ratio which is set manually and will be discussed in the section 3.3. The function $top(\cdot)$ returns the indices of top $n^{(l+1)} = \lceil r * n^{(l)} \rceil$ values of the information scores S. $H^{(l)}(idx,:)$ and $A^{(l)}(idx,idx)$ performs the element selection according to the indices of top information scores. Thus, in the l-th layer, $n^{(l+1)}$ nodes are remained and others are removed.

Intuitively, the information score of a node is the feature difference between the average value of its neighbors and itself. The greater the difference, the higher the information score, and the less likely to remove the node. For example, if the feature of a node is equal to the average feature of its neighbors, it may be safe to drop this node without further information loss to the entire graph. On the other side, this selection method simulates a probable universal strategy for removing the information redundancy of fMRI: If the blood oxygen level activity is close to its neighbors, the area may be regarded as coactivated with its neighbors. After Removing all those nodes, the remaining ones may be activated in the first order, which may act like a trigger that has launched the sequence of regional brain activities.

2.3.3. Edge Prediction

At the same time, the node selection method may isolate some subgraphs and be susceptible to the initialization of brain graph structure, as it can not learn new connections beyond the given ones. To preserve the completeness of the subgraph, **?** developed a differentiable edge detection method based on the node features, which involved superfluous training overhead. Instead, we have designed a self-attention approach to predict underlying links among selected nodes:

$$E^{(l)}(p,q) \;=\; \frac{H^{(l)}(p,:)\cdot H^{(l)}(q,:)}{\left\|H^{(l)}(p,:)\right\|\left\|H^{(l)}(q,:)\right\|} + A^{(l)}(p,q) \tag{3}$$

where $E^{(l)}(p,q)$ represents the similarity score between the two nodes, $H^{(l)}$ denotes the feature matrix at the l-th layer. The corresponding element of adjacency matrix $A^{(l)}(p,q)$ is added to the cosine similarity of the two feature vectors to assign a more significant similarity score to the directly connected nodes. The similarity score is then normalized by the sparse attention mechanism proposed by **?**:

$$\begin{aligned} Sim^{(l)}(p,:) \;&:=\; \underset{\mathbf{n}\in\Delta^{K-1}}{argmin}\left\|\mathbf{n}-E^{(l)}(p,:)\right\|^2 \\ where\;\; \Delta^{K-1} \;&=\; \{\mathbf{n}\in\mathbb{R}^K \mid 1^T\mathbf{n}=1, \mathbf{n}\geq 0\} \end{aligned} \tag{4}$$

where Δ^{K-1} is a (K-1)-dimensional probability simplex and K denotes the number of nodes in the brain graph. For arbitrary node p, the softmax function normalizes the similarity scores between it and other nodes to probability distributions in Δ^{K-1}. However, this probabilistic approach retains small non-zero values of normalized similarities, which increases the complexity of downsampled subgraph. **?** projects the target vector onto simplex Δ^{K-1} and achieves sparsity when hitting the boundary. Specifically, the adjacency matrix is updated as the optimum of this quadratically constrained optimization problem. The closed-form solution is as follows:

$$A^{(l+1)}(p,q) \;=\; \left[E^{(l)}(p,q) \;-\; \tau(E^{(l)}(p,:))\right]_+ \tag{5}$$

$$\begin{aligned} \tau(\mathbf{n}) \;&=\; \frac{\left(\sum_{j\in Q(\mathbf{n})}\mathbf{n}_j\right)-1}{|Q(\mathbf{n})|} \\ where\;\; Q(\mathbf{n}) \;&=\; \{j\in[K] \mid \mathbf{n}_j > 0\} \end{aligned} \tag{6}$$

where $[K] = \{1,...,K\}$ and $[t]_+ := max\{0,t\}$. All the coordinates that below threshold function $\tau(\cdot)$ will be truncated to

zero. Thus, this piecewise function maintains the sparsity of adjacency matrix $A^{(l+1)}(p,q)$. Moreover, as shown in figure 5, this link prediction is able to detect some underlying connections that are not given in the initialization step, which makes the model more robust.

2.3.4. Individual-aware downsampling

To conclude a universal ASD clinical diagnosis suggestion, (**????**) tried to manually select top-N critical functional connections from the fMRI of all subjects. Ideally, this procedure would return some functional connections from which we could quickly tell if someone is suffering from this brain disorder. However, as discussed in **?**, the results are not satisfactory: The highest mean accuracy is 70.40%, and the smallest number of selected functional connections is 250, which is not adequate nor efficient for clinical diagnosis.

Different from that ambitious universal key features selection, we develop a personalized feature extraction strategy. Instead of selecting a set of universal biomarkers of ASD, we treat each subject separately and downsample the graph modeling according to its characters. As illustrated in the figure 1, we performed graph pooling onto every individual and stored the results as sparse vectors. In terms of storage and computation cost, like the above hard selection, that strategy has successfully downsampled the input brain imaging. On the other hand, the feature matrix has preserved more information than the previous methods. The two main benefits of this sparse feature fusion are as follows: First, it downsamples the brain imaging to extraordinarily few key components even without a performance decrease, which will be discussed in the section 3.3. Second, the difference in selected features between individuals has clinical meanings. It may indicate the varied brain regional activities and connections among different groups and will be discussed in the section 3.4.

2.4. Graph Convolutional Networks

To fuse brain imaging data and non-imaging data, we constructed a population graph where nodes represent subjects and edges indicate the similarity degree regarding phenotypic information. The non-imaging information similarity among subjects is characterized as the connectivity degree among nodes, i.e., nodes with similar phenotypic properties are more likely to be in the same community. We employ Graph Convolutional Networks to process the population graph structure with every node associated with a feature vector extracted from brain imaging. Proposed by **?**, GCN extends convolution operations onto graph structures and is able to learn the node embeddings. At each layer of GCN, the node feature vector is then recalculated as the weighted sum of its and its neighbors' features, that is, the node embedding. Hence, the features of nodes that are in the same community tend to follow a similar distribution.

2.4.1. Population Graph Construction

As stated in section 2.1, we use the data of 871 subjects preprocessed by CPAC. The connection between two nodes is decided by their phenotypic similarity, i.e., gender, age, handedness, etc. However, caused by inconsistent measurement among different data collection sites, some categories of data are not or partly collected. For example, nearly 30% of the handedness data are not available as illustrated in the table 1. Suggested by **?**, we consider a subset of the whole non-imaging data, which contains gender, age, and data collection sites information. Intuitively, the similarity is computed as the cosine similarity between two phenotypic feature vectors M_u and M_v.

$$Sim(u,v) = |\frac{M_u \cdot M_v}{\|M_u\| \, \|M_v\|}| \tag{7}$$

where $M = \{Age, Gender, Site\}$ denotes the selected subset of non-imaging data. A threshold of 0.5 is then applied to the derived similarity values to decide whether the two nodes u, v are connected or not.

$$A(u,v) = \begin{cases} 1, & if \; Sim(u,v) > 0.5 \\ 0, & otherwise \end{cases} \tag{8}$$

where $Sim(u,v)$ is the similarity score of the two subjects. A represents the adjacency matrix of the graph. Two nodes are connected if their cosine similarity value is above 0.5. By these means, the population graph is initialized as an undirected graph containing 871 nodes. In figure 6, we have also tested the GCNs with different population graph initialization.

2.4.2. GCNs

We have implemented two kinds of GCN in this part. The first layer is the same as the one proposed by **?**. The second layer is the Cluster-GCN presented by **?**, which has accelerated the basic GCN block.

Extending convolution operations to non-Euclidean space, GCNs have achieved promising performance on arbitrarily structured graphs. Though there exist different forms of GCN block, the universal core task is to learn a non-linear function $f(H^{(l)}, A)$ which aggregates the feature vectors of connected nodes to generate features for next layer:

$$H^{(l+1)} = f(H^{(l)}, A) = \sigma(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}H^{(l)}W^{(l)}) \tag{9}$$

with $\tilde{A} = A + I$, where I is the identity matrix and $\tilde{D}$ is the diagonal node degree matrix of $\tilde{A}$. For the l-th layer of the GCN, the graph can be represented as the feature matrix $H^{(l)}$. $H^{(0)} = X$ and $H^{(L)} = Z$ denote the input and final output feature matrix respectively. $W^{(l)}$ is the learnable weight matrix and $\sigma(\cdot)$ is the non-linear activation function, ReLU. In this way, the features are aggregated to form features of the next layer. **?** reduced the computational cost by clustering nodes into multiple bathes:

$$\frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}} \nabla loss(y_i, z_i^L) \tag{10}$$

where $\mathcal{B}$ indicates the subset of nodes and z_i^L represents the final prediction label of the i-th subject. Hence, at the loss backpropagation phase, the model only needs to calculate the gradient for the mini-bath. The Binary Cross-Entropy Loss is defined as the loss function:

$$loss(y_i, z_i^L) = -[y_i * log(z_i^L) + (1 - y_i) * log(1 - z_i^L)] \tag{11}$$

Table 2. Comparison with state-of-the-art methods on ABIDE I dataset. Best and second results are highlighted with BOLD and underline

References	Methods	Performance		
		Accuracy	Sensitivity	Specificity
?	GCN	69.50	-	-
?	LSTM	66.80	-	-
?	SVC	66.80	61.00	72.30
?	DNN	70.00	74.00	63.00
?	FCs selection and GCN	70.40	-	-
?	3D-CNN	73.30	-	-
?	Autoencoder	67.50	68.30	72.20
?	FCs selection and LDA	<u>77.70</u>	-	-
?	Joint learning	73.10	71.40	74.60
?	CNN	70.20	77.00	61.00
?	FCs selection and SVM	76.80	72.50	<u>79.90</u>
?	FCs selection and MLP	74.52	<u>80.69</u>	66.71
?	MC-NFE	68.42	70.05	63.64
Present study	**Graph pooling and GCN**	**86.45**	**82.70**	**89.62**

GCN: Graph Convolutional Networks
LSTM: Long Short-Term Memory
SVC: Support Vector Classification
MLP: Multi-Layer Perceptron
LDA: Linear Discriminant Analysis
LR: Logistic Regression
FCs: Functional Connections
MC-NFE: Multi-site Clustering and Nested Feature Extraction

After the graph convolutional layers, a linear classifier is applied on each node. The final outputs of the classifier represent the probability of ASD. By filtering the probabilities with a 0.5 threshold, the model finally outputs the predicted diagnostic group of each subject.

3. Experimental analysis

3.1. Experimental Settings

To make this experiment consistent with other studies (**??????**), we have performed the 10-fold cross-validation on the 871 samples and repeated it ten times. The multilayer perceptron and graph convolutional networks are trained separately but strictly on the same train set. At the training stage of the multilayer perceptron, we have employed the nested 10-fold cross-validation and repeated the inner loop five times every outer one. The whole framework is trained and tested on an NVIDIA TESLA V100S. During the optimization, we have utilized the Adam optimizer, of which the parameters are set as follows: learning rate = 0.0001, weight decay = 0.01. We have also used the dropout to enhance the generalization of GCN with a dropout ratio of 0.01.

3.2. Results

In the table 2, we have compared our framework with other models on the same ABIDE I dataset preprocessed by CPAC **?**. In general, those methods can be categorized into two types: single-stage and multi-stage. The single-stage methods directly deploy deep learning methods, like CNN, to deal with this ASD VS Control binary image classification problem (**??????**). However, limited by the number of available training samples, these naive implementations of neural networks have not achieved promising performance. On the other hand, multi-stage methods usually consist of two components: feature extraction and classification. Previous works trained feature extractors to extract features from brain functional connections (**????**). A classifier, like SVC, is then trained with extracted features as inputs. This kind of framework has successfully downsampled the high-dimensional brain imaging and thus made obvious performance improvement even with linear classifiers compared to the straightforward CNN models. For example, **?** constructed a specific workflow to select the key features from brain imaging and achieved an accuracy of 77.7% with linear discriminant analysis.

Like the above multi-stage methods, the present framework includes feature extraction and classification parts. We have employed graph pooling to downsample the given brain networking and trained a multilayer perceptron to further extract features, whereas previous feature extractors mainly focused on functional connections. Inspired by **?**, we employ GCN in the final classification part, which has leveraged imaging and non-imaging information. By these means, our framework outperforms the state-of-the-art method, reaching an accuracy of 86.45% and a mean AUC of 0.93. The efficiency of graph pooling is discussed in the following part.

3.3. Efficiency of Graph Pooling

As previously mentioned, studies have reported abnormal brain functional connections found on ASD subjects**?**. With

Fig. 2. Example of framework performance on the outer loop of the nested cross-validation. The backbone is set as Graph pooling and GCN. The random seed of outer loop = 13.

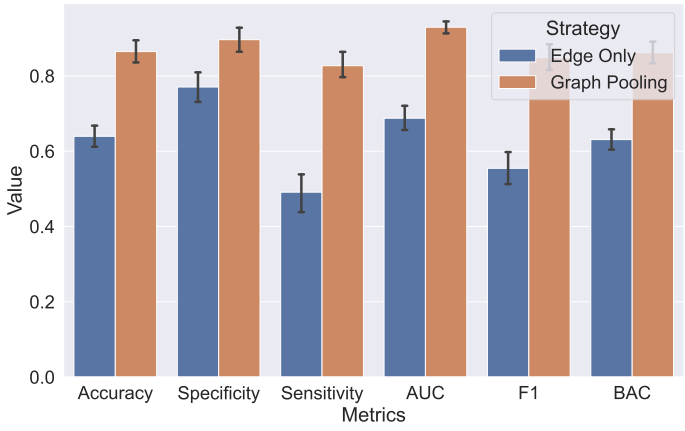


Fig. 3. Comparison between two strategies: edge-only brain representation and graph pooling. The brain imaging downsampled by the two strategies is fed into the same backbone, i.e., MLP and GCN, to assure other settings are unchanged.

this prior knowledge, in (**????**), the authors solely simplified brain imaging as functional connections. To directly and fairly compare these two downsampling strategies, we have fed the results of the two into our framework with strictly the same settings. As shown in figure 3, the graph pooling outperforms the edge-only brain networking representation with the proposed framework in terms of all the current metrics. The framework with graph pooling results has achieved an accuracy of 86.45%, while it with functional brain connections reached an accuracy of 63.96%. Considering that the only difference between the two models in figure 3 is the preprocessing method, we can conclude the superior of graph pooling.

To further evaluate the efficiency of graph pooling, we have tested the graph pooling for different brain networking initialization, including Erdős–Rényi model **?**, Watts-Strogatz model **?**, Barabási–Albert model **?**, Bipartite graph **?**, and fully connected graph. As shown in figure 4, the graph pooling is not sensitive to the initialization of brain networking details, which indicates that the proposed self-attention graph pooling is efficient and robust to downsample brain networking. Even if the edges are randomized according to Erdős–Rényi model with a probability of 0.3, the proposed framework can still achieve promising performance. Besides, the bipartite graph model **?** has reached similar or better performance in terms of some metrics compared to the fully connected graph.

The main advantages of graph pooling are 3-fold: First, it can be easily generalized to other related problems. Previous methods work under a strong assumption that it is only the abnormal functional connections that cause ASD. This presupposition is not only biased in the current task (**????**) but also limit generalizing these methods to other brain disorder diagnosing problems. On the contrary, graph pooling requires less prior knowledge about brain functions as it straight receives the entire graph representation of brain imaging and can learn connections beyond the given brain networking. Besides, as discussed in section 2.3.2, this unsupervised downsampling method needs no training overhead. Second, it is more efficient for brain networking downsampling. As shown in table 3, by using the downsampling results of graph pooling, the framework has achieved a remarkable improvement. Third, it can be aware of individual characters to some degree, which is benefited from the self-attention mechanism and sparse representation of extracted features, as discussed in section 2.3.4. We have observed variance in brain imaging pooling results of subjects from different groups, as shown in figure 5, and found even more obvious variance when we further split the groups. This inter-group heterogeneity may have clinical and biomedical meanings.

3.4. Key Brain Substructures

Interpreting the ASD diagnosis framework has been being a hot topic as it may indicate the brain biomarkers of autism spectrum disorder and direct the early intervention. To intuitively exhibit the pooling results, i.e., the most critical substructures selected at that stage, we have plotted them by BrainNet Viewer proposed by **?**. We have applied graph pooling onto the brain imaging of all subjects. Thus, for every individual, the pooling result is six selected nodes and their connections. Subsequently, we have split 871 subjects into different groups regarding their non-imaging properties, including age, gender, and the data collection site they belong to. We have surveyed the top 15 most frequently selected regions and connections from the pooling results of all subjects inside this group for each one.

Unlike figuring out universal brain biomarkers of ASD, the outputs of self-attention pooling only have specified the importance of regions and edges of individual brain imaging. However, we can still draw some conclusions by summarizing the pooling results in different groups. The illustrated substructures of the brain, as shown in figure 5, may indicate a common brain activity mechanism inside a specific group.

Ideally, as discussed in section 2.3.2, the remained regions are the first activated ones regarding external stress or active internal activities. They act like a trigger that has launched a sequence of regional activities. Based on this knowledge about model working principles, we have observed some inter-group heterogeneity in essential substructures selected by graph pooling. That finding indicates that self-attention graph pooling along can be aware of individual phenotypic properties to some degree. In figure 5, few differences have been found between ASD and Typical Control subjects as the heterogeneity caused by individual characteristics may be averaged. We further divide the two groups into four by incorporating gender into consideration, as shown in figure 7. According to the basic posist above, we may not conclude that it is the illustrated brain key substructures differences between different groups that have caused ASD. But these personalized subgraphs, serving as the inputs of MLP, are sufficient for ASD diagnosis.

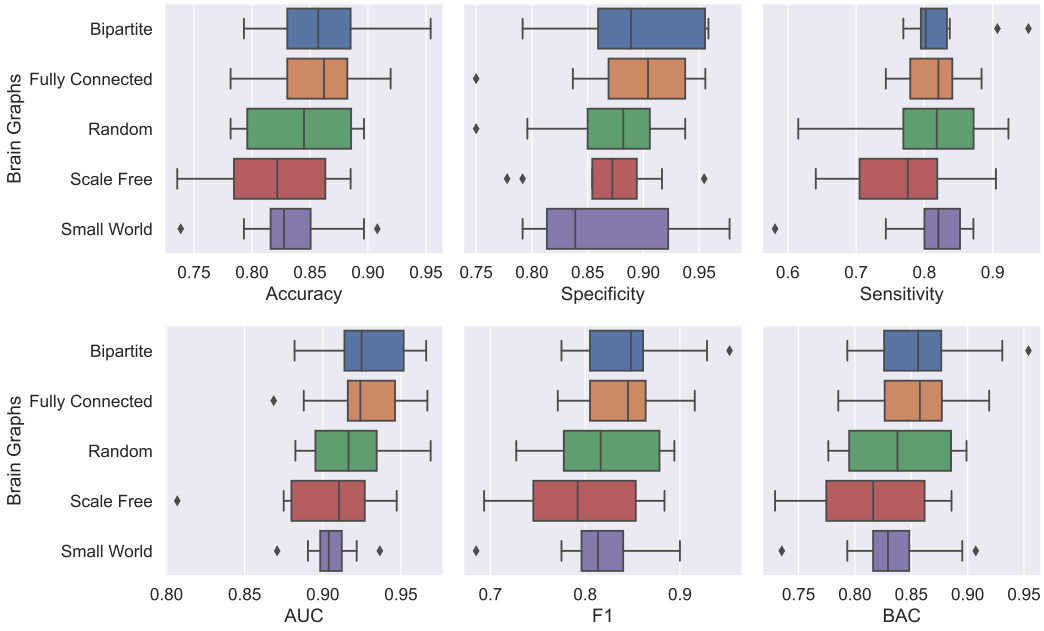
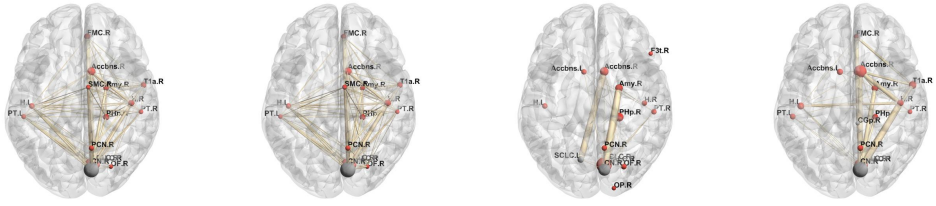
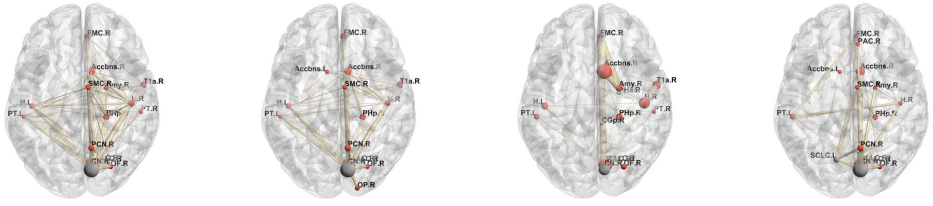


Fig. 4. Study of the impact of different brain graph initializations. The small world graph is generated based on the Watts-Strogatz model. The scale-free graph is initialized using BA Scale-Free Network Model. The Erdős–Rényi random graph is created with probability of linking two nodes=0.3.



(a) ASD (c) Male (e) Age >= 40 (g) Collected on PITT



(b) Typical Control (d) Female (f) Age <= 8 (h) Collected on Max_Mun

Fig. 5. Results of individual-aware downsampling. The pooling ratio is set as 0.05, i.e., 6 regions are selected out of 111. Each axial view of the brain shows the top 15 nodes most frequently selected from those who are inside the corresponding group. The width of edges and size of nodes indicate the relative frequency of being selected. Number of subjects inside each group are as follows: a: 403; b: 468; c: 727; d: 144; e: 14; f: 15; g: 50; h: 46.

3.5. Efficiency of Graph Convolutional Networks

? incorporated graph convolutional networks with brain functional connection selections and obtained accuracy improvement compared to **?**. The classification accuracy was increased from 66.80% to 70.40% by leveraging both imaging and non-imaging information with GCN. To figure out the efficiency of GCN in our framework, we have tested its efficiency in fusing imaging and text information. As shown in figure 6, we have compared the population graph of which the edges are built regarding the non-imaging information with a random graph and a fully connected graph. The GCN run on the random graph reached a mean accuracy of 53.50%, while the text-information population graph, considering age, site, and gender similarities, achieved an accuracy of 86.45%. This comparison proved the efficiency and importance of the provided non-imaging data. Besides, even compared to the fully connected graph, the framework run on the text-information graph achieved better performance in some metrics. We would also like to highlight that when the information of data collection sites is not given, the framework can also reach a classification accuracy of 83.24%, which indicates the feasibility of deploying the framework onto a single site.

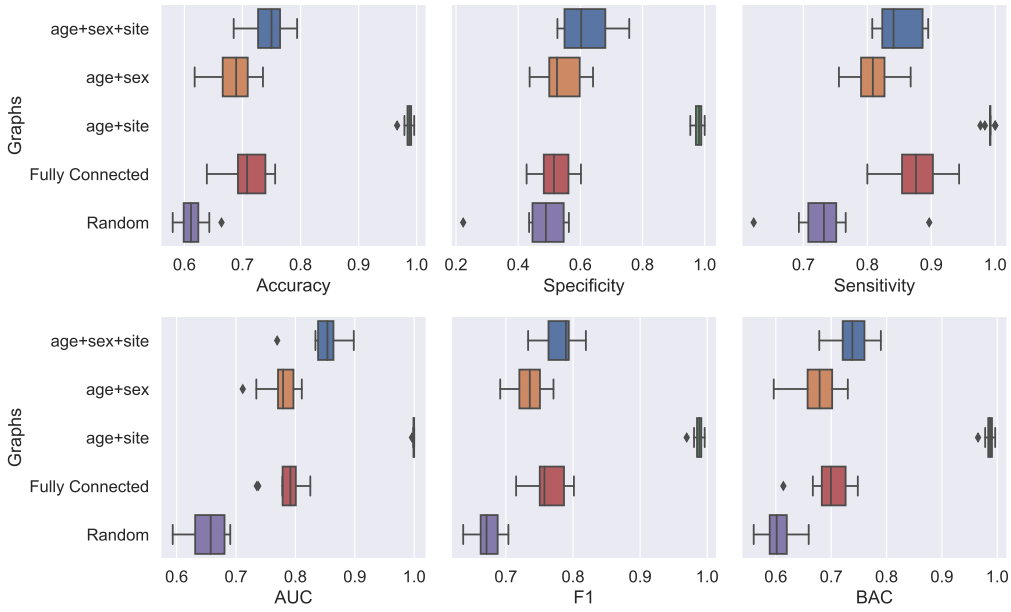
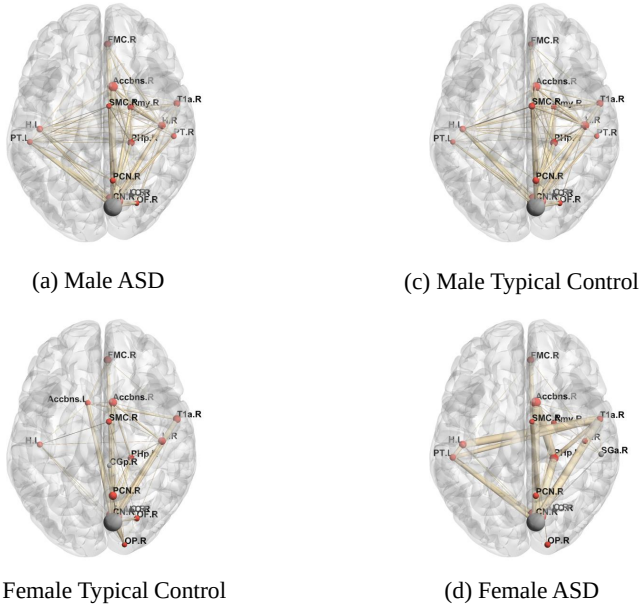


Fig. 6. Comparison between different population graph initializations. The four strategies are tested with the same node features.



(a) Male ASD

(c) Male Typical Control

(b) Female Typical Control

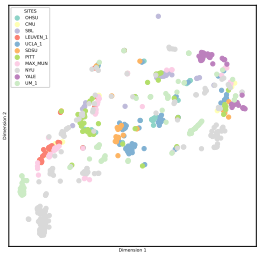
(d) Female ASD

Fig. 7. Additional brain view for figure 5 with same settings. The illustrated subgraphs indicate selection preference in graph pooling.

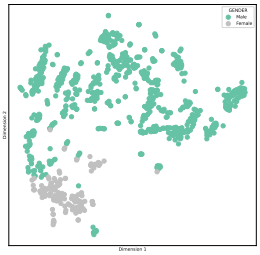
All the previous works, which have employed GCN for the same purpose of leveraging non-imaging information and fMRI (**???**), have assumed the ability of GCN to be aware of inter-individual phenotypic differences and to regularize raw features based on the former. Even though the efficiency of fusing imaging and non-imaging data is proved as discussed above, there is no clear conclusion that the GCN really has learned the inter-individual non-imaging differences. To intuitively present the learned node embeddings, we have downsampled them onto the 2D plane with t-SNE proposed by **?**. As shown in the left three plots of figure 8, even though graph pooling has detected some implicit inter-group heterogeneity, which is discussed in section 3.4, the features subsequently learned by MLP have not performed the relative feature distribution difference in respect of phenotypic information. The inevitable information loss during the feature dimensionality reduction may have caused this inconsistency, as the dimension of the basic features is up to 128. Still, the node embeddings learned by Graph Convolutional Networks have shown obvious clustering even in the 2D space. As exhibited in the right three diagrams of figure 8, the distance among subjects that are identical regarding a certain kind of phenotypic information is relatively close in the feature space compared to those who are not. This clustering is more evident when only considering the genders, which indicates that the learned edge weights of the population graph may depend more on it.

4. Discussions

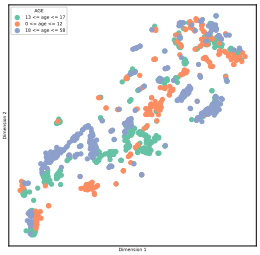
First, we have proved the superiority of the proposed self-attention graph pooling with the current framework. As we have highlighted that the unsupervised graph pooling has no training cost, it is worthy of applying more models to further improve the diagnosis accuracy and prove the efficiency of the proposed downsampling method. Second, we have used GCN on the population graph to narrow the inter-site heterogeneity. It is interesting to deploy federated learning in the hospitals to build similar individual connections, which may be more practical in the real world. In addition, the individual-aware downsampling is different from usual fixed universal biomarkers. Personalized brain networkings cannot directly indicate the cause of autism. But it can provide a new way of downsampling brain imaging and improving the model performance. For neural scientists, it



(a) (b) Node Embeddings; Sites Raw Extracted Features; Sites



(c) (d) Node Embeddings; Genders Raw Extracted Features; Genders



(e) (f) Node Embeddings; Ages Raw Extracted Features; Ages

Fig. 8. 2D view of the node embeddings learned by Graph Convolutional Networks. The nodes, which represent subjects, are colored according to different phenotypic properties: Sites, genders, and ages

may also indicate some brain activity patterns, as discussed in section 3.4.

5. Conclusion

In this paper, we have proposed a framework to identify Autism Spectrum Disorders. First, we have proposed a simple self-attention downsampling method for fMRI. This end-to-end, unsupervised, and flexible graph pooling method has successfully considered brain functional connections and regional activities simultaneously, which can also be aware of individual differences in brain function. Besides, we have exploited Graph Convolutional Networks to incorporate imaging with phenotypic information and illustrated its efficiency in recalibrating the feature distribution. Our framework has achieved a mean accuracy of 86.45% ± 0.05 and a mean AUC of 0.93 ± 0.03 on ABIDE I dataset. The superior performance of our model indicates its ability to detect Autism Spectrum Disorders and contribute to early intervention.

Declaration of Competing Interest

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

CRediT authorship contribution statement

Li Pan: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Writing - review & editing. Jundong Liu: Conceptualization, Methodology, Formal analysis. Mingqin Shi: Conceptualization. Chi Wah Wong: Conceptualization, Methodology, Formal analysis, Writing - review & editing. Kei Hang Katie Chan: Conceptualization, Methodology, Writing - review & editing, Supervision, Funding acquisition.

Acknowledgments

This work was supported by City University of Hong Kong New Research Initiatives/Infrastructure Support from Central (APRC) and Centre for Perceptual and Interactive Intelligence (CPII), Chinese University of Hong Kong.