EAG-RS: A Novel Explainability-guided ROI-Selection Framework for ASD Diagnosis via Inter-regional Relation Learning

Wonsik Jung, Eunjin Jeon, Eunsong Kang, and Heung-Il Suk *Member, IEEE*

***Abstract*— Deep learning models based on resting-state functional magnetic resonance imaging (rs-fMRI) have been widely used to diagnose brain diseases, particularly autism spectrum disorder (ASD). Existing studies have leveraged the functional connectivity (FC) of rs-fMRI, achieving notable classification performance. However, they have significant limitations, including the lack of adequate information while using linear low-order FC as inputs to the model, not considering individual characteristics (*i.e.*, different symptoms or varying stages of severity) among patients with ASD, and the non-explainability of the decision process. To cover these limitations, we propose a novel explainability-guided region of interest (ROI) selection (EAG-RS) framework that identifies non-linear high-order functional associations among brain regions by leveraging an explainable artificial intelligence technique and selects class-discriminative regions for brain disease identification. The proposed framework includes three steps: (i) inter-regional relation learning to estimate non-linear relations through random seed-based network masking, (ii) explainable connection-wise relevance score estimation to explore high-order relations between functional connections, and (iii) non-linear high-order FC-based diagnosis-informative ROI selection and classifier learning to identify ASD. We validated the effectiveness of our proposed method by conducting experiments using the Autism Brain Imaging Database Exchange (ABIDE) dataset, demonstrating that the proposed method outperforms other comparative methods in terms of various evaluation metrics. Furthermore, we qualitatively analyzed the selected ROIs and identified ASD subtypes linked to previous neuroscientific studies.**

***Index Terms*— Autism spectrum disorder, resting-state fMRI, layer-wise relevance propagation, ROI selection**

I. Introduction

Autism spectrum disorder (ASD) is a neurological disability associated with brain development. Patients with ASD experience social communication and interaction difficulties in multiple contexts, and exhibit limited or repetitive behavioral patterns, interests, or activities [1]. Although patients with ASD incur considerable average medical expenses over their lifetime (*e.g.*, at least one million dollars per patient [2]), accurate clinical curative treatments are not available, forcing them to suffer from lifelong illnesses [3]. Therefore, it is crucial to identify the emergence of the disease as early as possible for accurate treatment [4].

Over the past few decades, several approaches based on resting-state functional magnetic resonance imaging (rs-fMRI) have been proposed to diagnose various brain diseases, including ASD [5], schizophrenia [6], and Alzheimer's disease [7]. Rs-fMRI is a non-invasive technique that identifies spatio-temporal scales of regional brain activation by measuring blood-oxygen-level-dependent (BOLD) signals [8]. Most existing rs-fMRI studies utilize raw time signals [9]–[11] or low-order brain functional connectivity (FC) [12], [13]. FC is typically constructed based on the temporal correlation between spatially remote brain regions—regions of interest (ROIs)—in a statistical manner [14], [15]. Therefore, FC not only provides information about functional communication in the human brain [16] but also employs it as a biotype for the disease diagnosis [17].

With recent advances in deep learning (DL), brain disease diagnostic methods based on rs-fMRI have garnered significant attention in neuroimaging research. Raw time signals of rs-fMRI have been used as inputs in recurrent neural networks (RNNs) [9], graph convolutional neural networks (GCNs) [18], *etc*, and FC has been used as input in a variant of auto-encoder (AE) architecture [19]–[25] to develop methods for diagnosing brain diseases. In addition, several approaches have utilized more discriminative features for improving diagnostic performance [21], [23], [26], [27]. Feature selection (FS) methods involve (i) ranking-based methods, in which all features are ranked and then curated based on their specific criteria [21], [26], and (ii) subset-based methods, which select features by optimizing a definite objective function [23], [27]. Moreover, FS methods help explore the pathology of brain diseases by considering the selected features as biomarkers [28].

Although these diagnostic methods exhibit remarkable classification performance, they continue to suffer from

This research was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (No. 2022R1A4A1033856) and the Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) No. 2022-0-00959 ((Part 2) Few-Shot Learning of Causal Inference in Vision and Language for Decision Making) and (No. 2019-0-00079, Artificial Intelligence Graduate School Program(Korea University)).

W. Jung, E. Jeon, and E. Kang are with the Department of Brain and Cognitive Engineering, Korea University, Seoul 02841, Republic of Korea (e-mail: {ssikjeong1, eunjinjeon, eunsong1210}@korea.ac.kr)

H.-I. Suk is with the Department of Artificial Intelligence, Korea University, Seoul 02841, Republic of Korea and also with the Department of Brain and Cognitive Engineering, Korea University, Seoul 02841, Republic of Korea (e-mail: hisuk@korea.ac.kr).

Corresponding author: Heung-Il Suk

certain limitations. Firstly, most existing methods use (partial) Pearson's correlation as their input FC, which represents the linear correlation of brain regions as connectivity strength and contains low-order information within brain regions or voxels. However, merely considering low-order information is not sufficient for capturing subtle changes in signal between normal and patient groups [29]. [30], [31] proposed a method for constructing high-order FC networks based on the similarity between the topographical profiles of pairs of FCs, which is referred to as "correlation's correlation". In contrast, existing studies on FC typically calculate low-order or high-order FCs separately. However, pairs of FC levels may exhibit intriguing relationships or functional associations. In this context, [32] integrated three types of FCs, encompassing low-order FC, high-order FC, and the inter-level associated FC, and proposed a hybrid high-order FC network for brain disease diagnosis tasks. Their method exhibited higher accuracy than methods based on a single type of FC. Based on the findings reported in prior studies, our study aims to enhance diagnostic performance by leveraging a combination of low-order FC and high-order features.

Ranking-based FS approaches focus on single levels of contribution, and therefore, do not consider complementary information between multiple features [28]. On the other hand, subset-based FS methods investigate the importance of various groups of features simultaneously and do not consider the individual characteristics of patients with ASD. Finally, a few other FS methods use refined inputs and ignore local-global structural information in terms of the entire population, making their decision-making processes difficult to explain. In practice, explainability is vital in the medical field (especially in neuroimaging) to improve reliability.

To address the aforementioned issues, we propose a novel explainability-guided ROI selection (EAG-RS) framework that selects informative features dynamically at the ROI-level for brain disease diagnosis. To this end, we estimate high-order information of FC based on a high-level representation obtained from the layer-wise relevance maps. We leverage the estimated information in conjunction with low-order information for ASD diagnosis learning. Further, prior studies [19], [24] have primarily focused on learning low-level inter-regional FC relationships based on computer vision tasks rather than from a neuroscientific perspective. Our earlier work [25] introduced a novel approach from a neuroscientific perspective to complement these limitations, emphasizing brain region-level considerations. We designed and used random ROI-level masking to facilitate robust and expressive feature learning. Given an ROI-masked FC, a stacked AE (SAE) inherently learns non-linear relations among remaining ROI connections to reconstruct or infer masked connections. Following model training, connection-wise relevance score estimation is performed based on the pre-trained SAE with layer-wise relevance propagation (LRP) [33] to explore the high-order relations between functional connections. The LRP transmits the output of the trained network back to the input level using a decomposition rule, which enables the identification of input connection features that contribute to the restoration of masked connections either positively or negatively. Thus, we estimate non-linear high-order relations among seed-based networks (*i.e.*, ROIs) in FC via the trained inter-regional non-linear relational learning model. Finally, given the estimated non-linear high-order FC, the pre-trained encoder and classifier are trained to discover ASD-informative ROIs at the sample level.

The proposed method is verified to exhibit superior classification performance than comparative methods on the publicly available Autism Brain Imaging Database Exchange (ABIDE) dataset [34]. The impacts of individual components of the proposed method are estimated using ablation studies and post-hoc analysis is performed to identify ASD subtypes. The main contributions of our study can be summarized as follows:

- We propose a novel method to derive the high-order information of FC using a connection-wise relevance score estimation module between each masked seed ROI and other neighboring ROIs.
- Two types of representative vectors (*i.e.*, mean and count) statistically estimated based on the connection-wise relevance score, contribute to the selection of disease-relative ROIs at the individual level.
- Our proposed framework achieves state-of-the-art diagnosis performance on the ABIDE dataset. The neuroscientific analysis is also conducted using our framework.

This study is an extension of our previous work [25], in which we introduced a self-supervised learning framework that considers inter-regional non-linear relations for rs-fMRI. In this study, the proposed framework is supplemented using connection-wise relevance score estimation and a diagnosis-informative ROI selection network, thereby improving its ASD diagnosis performance. We further conducted an ablation study to verify the capabilities of the constituent modules in our proposed framework. In addition, we utilized selected ROIs at the sample level for cluster subtyping of autism and analyzed them to acquire neuroscientifically reliable results by performing group-wise ROI comparisons.

II. Related Works

A. DL-based Brain Disease Diagnosis Methods in rs-fMRI

In recent decades, DL has been utilized to diagnose brain diseases using rs-fMRI [9], [11], [18], [19], [21]–[23], [35]. In [9] and [11], diagnosis tasks were performed using the raw time signals of rs-fMRI. In particular, [9] first trained a long short-term memory network (LSTM) to capture greater amounts of temporal information. [11] focused on extracting more disease-relevant intermediate features by combining a self-attention mechanism with mutual information maximization. Another approach is to use the FC of the rs-fMRI as the input for DNNs. For example, [19] and [22] learned the hidden representations

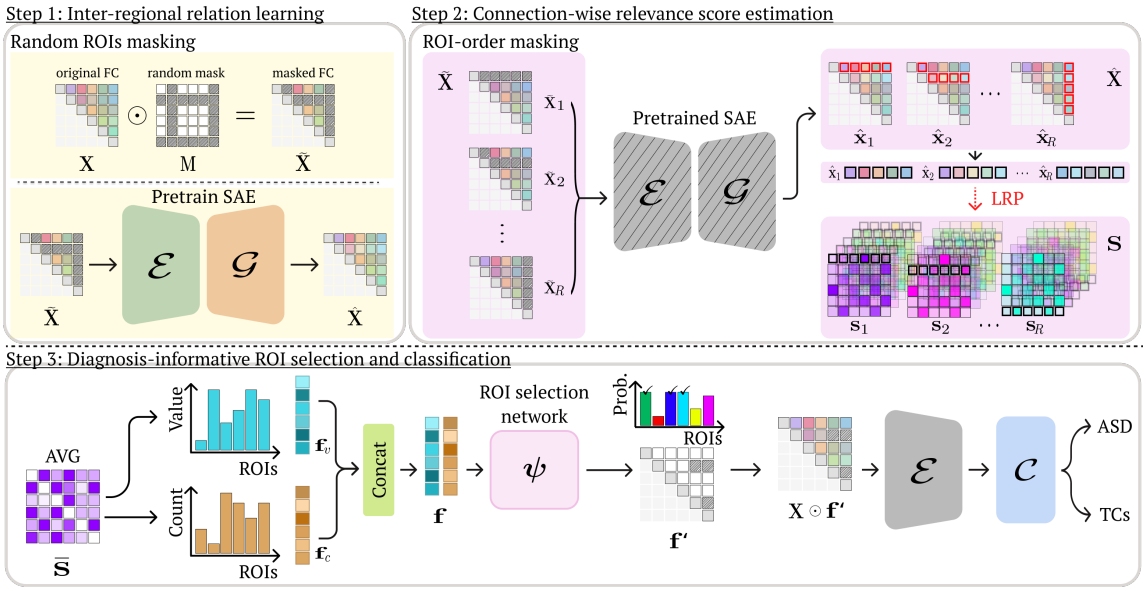

Fig. 1. Overview of the EAG-RS framework comprising a three-step learning strategy: i) inter-regional relation learning using seed (ROI)-based network masking that generates masked input FCs for ROIs, ii) estimating connection-wise relevance scores via LRP to investigate high-order information between functional connections, and iii) extracting ROI-level representative vectors from the estimated relevance scores for simultaneous diagnosis-informative ROI selection and brain disease diagnosis. Slashed modules (*i.e.*, pre-trained SAE) represent modules with frozen parameters (*i.e.*, fixed parameters). Given a masked FC, the SAE is trained to discern non-linear relations among low-order FC connections. Henceforth, we estimate connection-wise relevance scores using the pre-trained SAE combined with LRP. These scores signify the importance of a given connection in restoring other connections. Lastly, we integrate low-order and high-order FCs to individually select informative ROIs for ASD diagnosis.

of FC by training AE variants and using them for classification.

Several studies have demonstrated that the FS-based methods enhance classification performances. [21] proposed the DNN-FS method, in which the authors ranked outputs of multiple stacked sparse AEs in terms of their Fisher scores to determine more discriminative features. [23] devised support vector machine recursive feature elimination (SVM-RFECV). Further, [35] selected features based on a combination of elastic net and manifold regularization, referred to as MTFS-EM.

However, existing diagnosis methods result in suboptimal performance because of their single-level contribution to low-order information, without considering individual characteristics. In contrast, our proposed method presumes the high-order FC information dynamically and explores more discriminative features by incorporating estimations and explainable relevance maps.

B. Explainability for fMRI

Deep neural networks (NNs) suffer from a black-box problem, wherein their outcomes cannot be easily explained because of complex non-linear mechanisms. To address this issue, various explainable artificial intelligence (XAI) methods have been applied to neuroimaging models [36]–[39]. Among these, LRP is widely used to identify group-discriminative features/patterns by quantifying their relevance to the model outcome [25], [36], [38]. For example, [38] analyzed disease classification results by searching for the most group-discriminative patterns using LRP.

Besides providing neuroscientific explanations, LRP can also be used to construct novel FC-determination techniques. For example, a novel brain connectivity measurement based on a trained network with BOLD signals was reported using LRP [40]. The objective of this paper was to explain the relevance of one region in influencing another in a regression task using NNs.

Similar to [40], we explore brain connectivity using LRP in this paper. The main differences between [40] and our study are as follows: (i) We use a seed-based network mask as the input of the FC network, instead of BOLD signals, which include self-seed-ROI influences. Therefore, we focus on the inter-regional non-linear relationships to estimate the contributions of neighbors inherently restoring a seed-based network. (ii) We apply LRP results to explore high-order relations between functional connections using connection-wise relevance score estimation and leverage them to select ASD-discriminative ROIs dynamically during training. This is unlike any technique in any previously published LRP-based study.

III. Methods

The overall framework of the proposed EAG-RS is illustrated in Fig 1. It comprises three phases; i) inter-regional relation learning, ii) connection-wise relevance score estimation via LRP, and iii) brain disease diagnosis based on diagnosis-informative ROI selection. Given an FC randomly masked at the seed-based network level, the SAE is trained to learn a feature representation that reflects non-linear relations between low-order FC connections. The proposed framework estimates the connection-wise relevance score, which represents the relevance of a

connection to the restoration of other connections using the trained SAE model and LRP. This framework helps analyze FCs between spatially distinct regions. Then, ASD-informative ROIs are selected at an individual level based on statistical measures of the relevance scores (*i.e.*, mean and count). Finally, we diagnose ASD by considering only ASD-informative ROIs in FC.

A. ROI connection masking

Given an input FC matrix $\mathbf{X} \in \mathbb{R}^{R\times R}$, where R denotes the number of ROIs, we generate a mask matrix $\mathbf{M}$ at the ROI-level, where the set $\mathbf{R}_q$ includes randomly selected ROI indexes based on a q-ratio. The mask matrix $\mathbf{M} \in \mathbf{1}^{R\times R}$ is updated using the following rule: $\mathbf{M}(i,:) = 0$ and $\mathbf{M}(:,i) = 0$ for $i \in \mathbf{R}_q$, where $(i,:)$ and $(:,i)$ denote the elements of the i-th row and any column, and any row and the i-th column, respectively. Then, the masked input FC matrix $\tilde{\mathbf{X}}$ is obtained via element-wise multiplication of the input FC matrix $\mathbf{X}$ with the mask matrix $\mathbf{M}$, and it is denoted by $\tilde{\mathbf{X}} = \mathbf{M} \odot \mathbf{X}$, where $\odot$ represents an element-wise multiplication operator. The masked FC matrix $\tilde{\mathbf{X}}$ is flattened to a one-dimensional vector $\tilde{\mathbf{X}} \in \mathbb{R}^D$, where D indicates the number of elements in the upper triangle of the FC matrix without diagonal elements ($D = R \times (R-1)/2$). In this procedure, masks are generated arbitrarily for each input during each iteration. This enables augmentation of the training samples and helps in learning robust and enriched feature representations, thereby preventing overfitting [41].

B. Inter-regional relation learning

In contrast to our previous work [25], in this paper, we focus on inter-regional relation learning without considering classification to be the first step for examining high-order information of FCs. Initially, the flatten-masked FC $\tilde{\mathbf{X}}$ is embedded into a hidden space with $\mathbf{h}_1 \in \mathbb{R}^{D_1}$, as follows:

$$\mathbf{h}_1 = \mathcal{E}_1(\tilde{\mathbf{X}}) = \sigma(\mathbf{W}_1\tilde{\mathbf{X}} + \mathbf{b}_1), \tag{1}$$

where $\mathcal{E}_1$ denotes the first layer of the encoder with a weight matrix $\mathbf{W}_1 \in \mathbb{R}^{D_1\times D}$ and a bias vector $\mathbf{b}_1 \in \mathbb{R}^{D_1}$; and σ denotes the activation function. The first hidden represented features $\mathbf{h}_1$ is trained to reconstruct the original FCs $\mathbf{X}$ using the corresponding layer of the decoder (*i.e.*, generator) $\mathcal{G}_1(\mathbf{h}_1) = \tanh(\mathbf{W}_1'\mathbf{h}_1 + \mathbf{b}_1') = \hat{\mathbf{X}}$ by minimizing the reconstruction loss, as follows:

$$\min_{\mathbf{W}_1,\mathbf{W}_1',\mathbf{b}_1,\mathbf{b}_1'} \mathcal{L}_{rec}(\mathbf{X},\hat{\mathbf{X}}) = \sum_{i=1}^{N} ||\mathbf{X}^{(i)} - \mathcal{G}_1(\mathbf{h}_1^{(i)})||, \tag{2}$$

where N and $\hat{\mathbf{X}}$ represent the total numbers of training and reconstructed samples, respectively, which are outputs of $\mathcal{G}_1$; and $\mathbf{W}_1' \in \mathbb{R}^{D\times D_1}$ and $\mathbf{b}_1' \in \mathbb{R}^D$ denote a weight matrix and a bias vector, respectively. Note that $\Theta_\mathcal{E}$ and $\Theta_\mathcal{G}$ are learnable parameters; thus, $\{\mathbf{W}_1,\mathbf{b}_1\} \subset \Theta_\mathcal{E}$,$\{\mathbf{W}_1',\mathbf{b}_1'\} \subset \Theta_\mathcal{G}$.

During the training process, to learn high-level feature representations of FCs, we sequentially train the encoder ($\mathcal{E}_\ell$) and generator ($\mathcal{G}_\ell$) pair for each layer $\ell \in \{2,\dots,L\}$ in the network. We freeze the previous layer(s) of both the encoder and the generator while estimating the ℓ-th level representation of the FC input, $\mathbf{h}_\ell$, which corresponds to the $(\ell+1)$-th level non-linear relations among ROIs. Subsequently, this is transmitted to $\mathcal{E}_\ell$. The encoder $\mathcal{E}_\ell$ and the subsequent generator $\mathcal{G}_\ell(\mathbf{h}_\ell) = \tanh(\mathbf{W}_\ell'\mathbf{h}_\ell + \mathbf{b}_\ell') = \hat{\mathbf{h}}_{\ell-1}$ are trained by minimizing the sum of the reconstruction losses of $\{\mathbf{h}_\ell\}_\ell^L$ and $\mathbf{X}$. In this regard, the proposed SAE is trained to reconstruct the removed connections as well as estimate the remaining connections based on relations inherently present in the neighboring connections. In this step, the proposed model learns an inter-regional non-linear representation that encompasses first-order connections of rs-fMRI as well as high-level relations among ROIs.

C. Connection-wise relevance score estimation

After training the SAE using the process stated in Section III.B, we utilize the LRP technique to estimate connection-wise relevance scores in the pre-trained SAE. The relevance score represents the influence of each connection on other connections. The LRP traces back from the final output layer to the input connection layer to calculate these scores. Specifically, we define the relevance score $S_j^{\ell+1}$, which represents a hidden unit j in the $(\ell+1)$-th layer. The relevance score $S_j^{\ell+1}$ is determined based on the contribution of all hidden units in the ℓ-th layer that affect the activation of the hidden unit j in the subsequent layer $\ell+1$. This ensures that the total relevance per layer is conserved [29] as $\sum_i S_{i\leftarrow j}^{\ell,\ell+1} = S_j^{\ell+1}$.

Given the original FC matrix $\mathbf{X}$, which includes a set of R seed-based networks, we generate a masked FC $\tilde{\mathbf{X}}$ by removing one of the seed-based networks (*i.e.*, ROI), resulting in a set of $(R-1)$ seed-based networks. To achieve this, the r-th seed-based networks are masked in the sequence of ROI indexes. Subsequently, the masked FC matrix is transmitted to the pre-trained SAE to reconstruct the original FC matrix, denoted by $\hat{\mathbf{X}}$. Via this reconstruction process, the masked FC reconstructs the masked seed-based network based on the remaining $(R-1)$ non-masked seed-based networks, as follows:

$$\hat{\mathbf{X}} = \text{pre-trained SAE}(\tilde{\mathbf{X}}) = \mathcal{G}(\mathcal{E}(\tilde{\mathbf{X}})), \tag{3}$$

where $\mathcal{E}$ and $\mathcal{G}$ represent the encoding and decoding layers of the pre-trained SAE, respectively. This process is repeated R times for each seed-based network in the FC, resulting in $\hat{\mathbf{X}} = \{\hat{\mathbf{x}}_r^\top\}_{r=1,\dots,R}$. Subsequently, we use the LRP technique to estimate the relevance scores, representing the contributions of other connections to the masked seed-based networks.

The reconstructed FC $\hat{\mathbf{X}} = \{\hat{\mathbf{x}}_1^\top,\dots,\hat{\mathbf{x}}_r^\top,\dots,\hat{\mathbf{x}}_R^\top\}$ is utilized to estimate the connection-wise relevance score $\mathbf{S} = \{\mathbf{s}_1,\dots,\mathbf{s}_r,\dots,\mathbf{s}_R\}$ via LRP. The relevance score represents the contributions of non-masked neighboring ROIs in restoring masked seed-ROI connections. To obtain this, we define the $\phi(\cdot)$ function, which assigns a value of

0 to masked regions (*i.e.*, the i-th row) and applies LRP to the remaining non-masked regions, as follows

$$\phi(\hat{\mathbf{x}}_{i,j}^{\top}, i, j) = \begin{cases} \mathbf{0} & \text{if } i\text{-th row,} \\ \text{LRP}(\hat{\mathbf{x}}_{i,j}^{\top}) & \text{otherwise} \end{cases} \tag{4}$$

where i and j represent the corresponding ROI indexes and $\dim(\mathbf{0}) = 1 \times R$, respectively. The original dimension of the LRP outcomes, obtained when the entire FC matrix is provided as input to the pre-trained SAE without masking, is $R \times R$. However, since we mask the i-th seed-based networks before transmitting them to the pre-trained SAE, the masked regions in the LRP outcomes can be disregarded. Therefore, the dimension of $\text{LRP}(\hat{\mathbf{x}}_{i,j}^{\top})$ in Eq. (4) is $(R-1) \times R$.

The connection-wise relevance score $\mathbf{s}_r$, where $r \in \{1, \ldots, R\}$ for reconstructing the r-th ROI based on other connections, is obtained using the $\phi(\cdot)$ function and defined as

$$\mathbf{s}_r = \Big\|_{j=1}^{R} \phi(\hat{\mathbf{x}}_{r,j}^{\top}, r, j), \tag{5}$$

where $\|$ indicates a concatenation operator and $\dim(\mathbf{s}_r) = R \times R \times R$. Note that the self-connections corresponding to the seed-ROI are excluded during the calculation of the relevance score. To estimate the local contributions of the r-th ROI, we simply aggregate the relevance scores for various connections [42]. This can be done using the following equation

$$\mathbf{s}'_r = \sum_{k=1}^{R} \mathbf{s}(\cdot, \cdot, k), \tag{6}$$

where $\mathbf{s}'_r \in \mathbb{R}^{R \times R}$ denotes the aggregated relevance score. Moreover, a global explanation can be represented by aggregating all the connections from the perspective of the r-th ROI.

By increasing the order of ROI indexes r from 1 to R as given by Eq. (5) and (6), we obtain the global explanation set $\mathbf{S}$, as follows:

$$\mathbf{S} = \Big\|_{r=1}^{R} \mathbf{s}'_r. \tag{7}$$

The detailed procedure is outlined in Algorithm 1. Finally, the ROI selection network takes the mean of the set $\mathbf{S}$ as the input.

Algorithm 1: Connection-wise relevance score estimation

input : Masked FC $\tilde{\mathbf{X}}$
output: Relevance score $\mathbf{S}$
Define a set $A = \{1, 2, \ldots, R\}$ # *ROI indices*
Initialize $\mathbf{S} = []$
for $r = 1, \ldots, R$ **do**
  Draw data $\tilde{\mathbf{X}}$ containing masked r-th ROI indexes
  $\hat{\mathbf{X}} \leftarrow$ pre-trained SAE($\tilde{\mathbf{X}}$)
  for $j = 1, \ldots, R$ **do**
    if $j = r$ **then**
      $\mathbf{s}_r$.append($\mathbf{0}$) # $\dim(\mathbf{0}) = R$ ▷ Eq. (5)
    else
      $A \leftarrow A - \{r\}$ # *excluding self-connection*
      $\mathbf{s}_r$.append($\text{LRP}(\hat{\mathbf{X}}_{r,j}^{\top})[A,:]$) ▷ Eq. (5)
      # $\dim(\text{LRP}(\hat{\mathbf{X}}_{r,j}^{\top})) = R \times R$
    end
    $\mathbf{s}'_r \leftarrow$ np.sum($\mathbf{s}_r$, dim=2) ▷ Eq (6)
    # $\dim(\mathbf{s}'_r) = R \times R$
  end
  $\mathbf{S}$.append($\mathbf{s}'_r$) # $\dim(\mathbf{S}) = R \times R \times R$ ▷ Eq (7)
end

Algorithm 2: Formulating ROI-level vectors

input : Dataset $\{\mathbf{X}, \mathbf{S}, \mathbf{Y}\}$
output: ROI-level representative vectors $\mathbf{f}_v, \mathbf{f}_c$
$\bar{\mathbf{S}} = \text{np.mean}(\mathbf{S}, \text{dim=2})$ # $\dim(\bar{\mathbf{S}}) = (N \times R \times R)$
Initialize a temporary mask $\mathbf{K}$ # $\dim(\mathbf{K}) = (N \times R \times R)$
Initialize $\mathbf{f}_v, \mathbf{f}_c = [], []$
for $n = 1, \ldots, N$ **do**
  for $r = 1, \ldots, R$ **do**
    if $\bar{\mathbf{s}}_r^n < \text{np.mean}(\bar{\mathbf{s}}_r^n)$ **then**
      $\mathbf{k}_r^n \leftarrow \mathbf{0}$
    else
      $\mathbf{k}_r^n \leftarrow \mathbf{1}$
    end
  end
  $\bar{\mathbf{s}}_r^n \leftarrow \bar{\mathbf{s}}_r^n \odot \mathbf{k}_r^n$
  $\mathbf{f}_v$.append(np.sum($\bar{\mathbf{s}}_r^n$, dim=2))
  $\mathbf{f}_c$.append(np.sum($\mathbf{k}_r^n$, dim=2))
end

D. ROI selection network and diagnostic classifier

1) ROI selection network (ψ): We perform statistical analysis to distinguish individual impacts and identify the most important effects [42]. Therefore, given the averaged relevance scores $\bar{\mathbf{S}} \in \mathbb{R}^{R \times R}$, we reformulate the ROI-level representative vectors (*i.e.*, $\mathbf{f}_v \in \mathbb{R}^{R \times 1}$ and $\mathbf{f}_c \in \mathbb{R}^{R \times 1}$) using statistical measures such as mean and count, as referenced in Algorithm 2.

Subsequently, these vectors are concatenated channel-wise as $\mathbf{f} = [\mathbf{f}_v \| \mathbf{f}_c] \in \mathbb{R}^{R \times 2}$, and transmitted into the ROI selection network ψ. The joint training of the ROI selection network and classifier reveals discriminative features for diagnosis. To maintain the information corresponding to each ROI, we use a convolutional layer (Conv1D) with a learnable 2×1 kernel, a stride of one in each dimension, and zero padding. Based on these configurations, we define the ROI selection network as follows:

$$\begin{aligned} \hat{\mathbf{f}} &= \psi(\mathbf{f}) \\ &= \text{Gumbel-softmax}(\mathbf{W}_{\psi_2} \sigma(\mathbf{W}_{\psi_1} \mathbf{W}_{\text{Conv1D}}(\mathbf{f}) + \mathbf{b}_{\psi_1}) + \mathbf{b}_{\psi_2}) \end{aligned} \tag{8}$$

where $\mathbf{W}_{\psi_1}$, $\mathbf{W}_{\psi_2}$, and $\mathbf{W}_{\text{Conv1D}}$ denote weight matrices, $\mathbf{b}_{\psi_1}$ and $\mathbf{b}_{\psi_2}$ denote bias vectors, σ denotes a Rectified Linear Unit (ReLU) activation function, and Θ_{ψ} denotes a learnable parameter.

2) Diagnostic classifier ($\mathcal{C}$): Information of individually selected ROIs, $\hat{\mathbf{f}} \in \mathbb{R}^{R \times 1}$, is reshaped and multiplied with the original FC $\mathbf{X}$. To this end, we perform the following operation: $\hat{\mathbf{f}} = \hat{\mathbf{f}} \odot \mathbf{1}^{\top}$, where $\mathbf{1}^{\top} \in \mathbb{R}^{1 \times R}$ represents a vector of size R containing a single value. In addition, to reflect the symmetrical characteristics of FCs, we perform

the following operation:

$$\mathbf{f}' = (\hat{\mathbf{f}} + \hat{\mathbf{f}}^{\top}) \odot \frac{1}{2}\mathbf{I}, \tag{9}$$

where $\mathbf{I}$ denotes the identity matrix. Subsequently, the original FC $\mathbf{X}$ and information of individually selected ROIs ($\mathbf{f}' \in \mathbb{R}^{R\times R}$) are element-wise multiplied and simultaneously transmitted to the encoders ($\mathcal{E}$) of the pre-trained SAE and the prediction network ($\mathcal{C}$) for the brain disease diagnosis task. Note that we remove the bias vector corresponding to the encoder layers to retain connections of zero values and prevent it from affecting other connections. The diagnostic classifier is trained to predict the clinical status $\hat{\mathbf{y}}$ by minimizing cross-entropy loss, as follows:

$$\mathcal{L}_{\text{cls}}(\mathbf{y}, \hat{\mathbf{y}}) = -\frac{1}{N}\sum_{n=1}^{N} \mathbf{y}^n \cdot \log(\hat{\mathbf{y}}^n). \tag{10}$$

where N and $\mathbf{y}$ denote the numbers of training samples and class labels, respectively.

E. Optimization

The objective function corresponding to each step comprises different losses, and it is given by

$$\text{Step 1}: \min_{\Theta_{\mathcal{E}},\Theta_{\mathcal{G}}} \alpha\mathcal{L}_{\text{rec}}(\mathbf{X}, \hat{\mathbf{X}}) + (1-\alpha)\sum_{\ell=2}^{L} \mathcal{L}_{\text{rec}}(\mathbf{h}_{\ell-1}, \hat{\mathbf{h}}_{\ell-1}),$$

$$\text{Step 3}: \min_{\Theta_{\psi}\Theta_{\mathcal{E}},\Theta_{\mathcal{C}}} \mathcal{L}_{\text{cls}}(\mathbf{y}, \hat{\mathbf{y}}),$$

where α is a hyperparameter used to control the ratio between two losses. The parameters of the encoder and generators are optimized by minimizing the combination of the reconstruction losses in Step 1. None of the parameters are updated during relevance score estimation via LRP in Step 2. In Step 3, the proposed model performs a classification task. To this end, we use a cross-entropy loss to train the pre-trained encoder of SAE, an ROI selection network, and a classifier.

IV. Experiments

A. Dataset & Pre-processing

We use pre-processed rs-fMRI data collected from the publicly available ABIDE[1] dataset [34]. The ABIDE dataset includes previously collected structural MRI, rs-fMRI, and phenotypic data for use by the broader scientific community. It consists of $1,112$ subjects, including 539 from individuals with ASD and 573 corresponding to typical development (TD) (ages 7–64 years, median 14.7 years across groups) from 17 international sites[2]. The ROIs fMRI series of all sites are downloaded from the pre-processed ABIDE dataset with a configurable pipeline for the analysis of connectomes (CPAC), band-pass filtering ($0.01 - 0.1$Hz), and no global signal regression, and it is parcellated using the Harvard-Oxford (HO) atlas. After downloading the pre-processed data, 110 ROIs are acquired using the HO atlas. At this stage, samples with missing filenames and incomplete data are excluded; and the remaining 880 samples across 17 international sites are utilized, which include 418 ASD subjects and 478 TD subjects. The Pearson correlation coefficient is used to estimate FC.

[1] `http://fcon_1000.projects.nitrc.org/indi/abide/`

[2] {UM, NYU, MAX MUN, OHSU, SBL, OLIN, SDSU, CALTECH, TRINITY, YALE, PITT, LEUVEN, UCLA, USM, STANFORD, CMU, and KKI}

B. Experimental Settings

To ensure a fair comparison, stratified five-fold cross-validation is conducted, where one fold is used for the validation set, another for the test set, and the remaining folds for the training set comprising all samples in the ABIDE dataset. Average performance is estimated in terms of the area under the receiver operating characteristic curve (AUC), accuracy (ACC), sensitivity (SEN), and specificity (SPEC). All proposed methods as well as competing methods are implemented using PyTorch and trained using a Titan RTX GPU on Ubuntu 18.04. All codes used in the experiments are available in a repository[3].

1) Training Settings: In the proposed SAE architecture, the encoder $\mathcal{E}$ comprises two fully-connected layers ($L = 2$) with the units of $\{9000, 1800\}$. The generator, $\mathcal{G}$, comprises two fully-connected layers with a reverse number of hidden units from the encoder. For the non-linear activation function (σ), the scaled exponential linear unit (SELU) is used for only the first intermediate layer in the encoder, and hyperbolic tangent (Tanh) is used for the remaining layers. The diagnosis-informative ROI selection network ψ comprises one Conv1D layer and three fully-connected layers with units of $\{512, 1650, 110\}$. The classifier $\mathcal{C}$ comprises two fully-connected layers with $\{10, 2\}$ hidden units. The ReLU activation function is used for all intermediate layers. In the meantime, we set the sigmoid and softmax functions as the activation functions of the last layers of ψ and $\mathcal{C}$, respectively.

In Step 1, 10% of the ROIs ($q = 0.1$) are randomly masked during every training iteration and the SAE is trained using Adam optimizer [43] with a learning rate of 10^{-3} and a mini-batch size of 50 over 300 epochs. In addition, ℓ_2 regularization is applied with a coefficient of 5×10^{-5}. We set α to be 0.5. All trainable parameters in Step 3 are optimized using the same settings except for the learning rate (10^{-4}). The Gumbel-softmax temperature is set to 0.01. Note that we take a grid search strategy for hyperparameter selection and select the best parameters based on the validation set results.

2) Competing Methods: The following six comparative methods are considered to evaluate the proposed method. First, a basic AE, dAE [44], and SAE are trained without any masking methods; they share the same architecture as that of EAG-RS. Further, EAG-RS is compared with the AE with $\mathbf{M}$, and SAE with Gaussian noise [45]. Henceforth, we denote these two baselines by AE (M) and SAE (G). To validate the effectiveness of ROI-level

[3] `https://github.com/ku-milab/EAG-RS`

masking, the SAE is trained using random FC connection masking, SAE (FC-M), inspired by [46]. Additionally, we demonstrate the statistical significance between our proposed EAG-RS and competing methods based on McNemar's test [47]. We also compare EAG-RS with other simple feature selection methods, including ranking-based approaches, such as the t-test ($p < 0.05$) [26] and recursive feature elimination (RFE) [48], as well as the subset-based approach LASSO [27]. In the case of these three methods, we utilize a linear SVM, which is a commonly used classifier in brain disease diagnosis [23]. Here, we adopt the hyperparameter C for SVM and λ for LASSO in the sets of $\{10^{-3}, 10^{-2}, \ldots, 10^{3}\}$ and $\{0.001, 0.002, \ldots, 0.01\}$, respectively. In the case of RFE-SVM, RFE iteratively assigns SVM weights to each feature based on its importance to the brain disease diagnosis by eliminating the least informative and redundant features. Further, ASD-DiagNet [22], which is a state-of-the-art approach that employs an autoencoder-based architecture with joint single-layer perception (SLP) training, is also considered. For feature selection in ASD-DiagNet, the 1/4 largest and 1/4 smallest Pearson's correlation values are used as input features based on the training data. The hyperparameter ranges for our experiments are derived from the values reported in [22].

In addition, we re-implement and compare the results of the state-of-the-art methods. First, we select BrainNetCNN [49], a convolutional neural network (CNN)-based model comprising edge-to-edge, edge-to-node, and node-to-graph convolutional filters, thereby utilizing the topological locality of brain network structures. Next, BrainGNN [50] is considered, which uses FC as a node feature and selects the top 10% positive partial correlations as edge features. The architecture of BrainGNN consists of ROI-aware graph convolutional layers and ROI-selection pooling layers, along with a regularization loss term that softens the distribution of the node pooling scores, facilitating the prediction of neurological biomarkers. Finally, BrainNetTF [51] is also considered. It exhibits a transformer-based architecture with an orthonormal clustering readout function that accounts for the similarity of ROIs within functional modules underlying brain regions. The hyperparameter configurations reported in our manuscript are adopted for each comparative method.

C. Experimental Results

1) AE-based classification results: The comparative methods, such as AE, AE (M), and dAE, use different masking methods and are trained in an end-to-end manner. On the other hand, the SAE-based methods adopt greedy layer-wise training strategies [52], and their masking configurations are different from those of the AE-based methods. Experimental results are summarized in Table I. SAE without a mask (SAE) is observed to outperform AE-based methods in terms of all metrics except for specificity. Meanwhile, SAE (G) outperforms SAE in terms of all metrics. In addition, SAE (FC-M) outperforms all competing methods in terms of AUC, ACC, and SPEC, but not sensitivity (SEN). Importantly, SAE with a random seed-based network mask (M), which was proposed in our previous work [25] and is referred to as Baseline in this paper, is observed to outperform SAE (FC-M) in terms of all metrics. However, it did not outperform comparative methods in terms of SEN adequately. Finally, our proposed EAG-RS is observed to outperform all competing methods in terms of all metrics.

TABLE I

AVERAGED CLASSIFICATION PERFORMANCE OVER ASD AND TD. (FC: FUNCTIONAL CONNECTIVITY, FC-M: RANDOM FC CONNECTION MASK, G: GAUSSIAN NOISE, M: RANDOM SEED-BASED NETWORK MASK, *: '$p < 0.05$').

Models	AUC	ACC (%)	SEN (%)	SPEC (%)
AE	0.676±0.043*	62.11±3.63*	59.41±7.26*	64.67±2.42*
AE (M)	0.685±0.034*	62.78±2.56*	54.71±11.50*	70.47±11.07*
dAE [44]	0.678±0.044*	62.58±3.95*	59.22±7.32*	65.79±4.70*
SAE	0.691±0.053*	63.64±3.35*	59.80±4.90	67.29±8.67
SAE (G) [45]	0.713±0.029*	65.36±2.34*	61.32±6.34	69.85±6.69
SAE (FC-M)	0.735±0.033*	66.35±3.70*	57.19±5.53*	74.65±7.54
Baseline [25]	0.757±0.040	69.76±3.45	57.82±6.64	80.22±4.54
EAG-RS	**0.760±0.033**	**73.71±2.83**	**64.56±8.36**	**80.74±8.28**

TABLE II

COMPARISON OF FEATURE SELECTION METHODS: AVERAGED CLASSIFICATION PERFORMANCE.

Methods	AUC	ACC (%)	SEN (%)	SPEC (%)
t-test+SVM	0.705±0.030	65.59±3.47	63.82±5.76	67.03±5.02
LASSO+SVM	0.727±0.030	66.88±3.80	63.01±5.18	70.93±4.95
RFE+SVM	0.706±0.010	65.00±2.25	63.04±1.09	67.60±1.45
ASD-DiagNet [22]	0.743±0.046	67.84±4.22	60.04±6.86	74.28±4.63
EAG-RS w/o ψ	0.757±0.040	69.76±3.45	57.82±6.64	80.22±4.54
EAG-RS	**0.760±0.033**	**73.71±2.83**	**64.56±8.36**	**80.74±8.28**

TABLE III

COMPARISONS OF ASD CLASSIFICATION PERFORMANCES OF THE PROPOSED METHOD AND STATE-OF-THE-ART METHODS ON THE ABIDE DATASET. NOTE THAT THE ENTRIES CORRESPONDING TO EACH METHOD ARE BASED ON THE RESULTS REPORTED IN THEIR RESPECTIVE MANUSCRIPT.

Models	AUC	ACC (%)	SEN (%)	SPEC (%)
LSTM [9]	-	68.50	-	-
DNN [19]	-	70.00	**74.00**	63.00
GCNs [18]	0.750	70.40	-	-
Extra-Trees+SVM [53]	-	72.20	68.80	75.40
MTFS-EM [35]	0.722	69.39	62.50	74.06
EAG-RS	**0.760**	**73.71**	64.56	**80.74**

2) Comparison of feature selection performance: Table II indicates that the subset-based method (*i.e.*, LASSO) outperforms the ranking-based methods (*i.e.*, t-test and RFE). However, ASD-DiagNet performed better than conventional FS approaches, except in terms of SEN. Although comparative FS methods are used to select important features and remove redundant ones to improve classification performance, their performances are still lower than the proposed method without feature selection (EAG-RS w/o ψ). The proposed method with

TABLE IV
COMPARISON OF ASD IDENTIFICATION PERFORMANCES OF THE PROPOSED METHOD AND STATE-OF-THE-ART METHODS ON THE ABIDE DATASET. NOTE THAT THE ENTRIES CORRESPONDING TO EACH METHOD ARE REIMPLEMENTED USING THE SAME EXPERIMENTAL CONFIGURATIONS AS THOSE IN OUR EXPERIMENTS.

Models	AUC	ACC (%)	SEN (%)	SPEC (%)
BrainNetCNN [49]	0.686±0.030	62.12±3.12	61.53±7.09	62.43±8.68
BrainGNN [50]	0.627±0.040	61.13±3.54	58.60±9.25	63.37±8.21
BrainNetTF [51]	0.700±0.026	65.60±3.34	64.38±3.26	66.42±4.69
EAG-RS	**0.760±0.033**	**73.71±2.83**	**64.56±8.36**	**80.74±8.28**

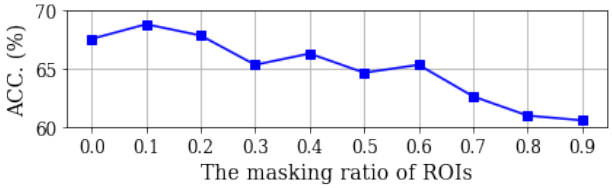

Fig. 2. Effectiveness of ROI-masking ratio, q, of the proposed framework on the ABIDE dataset.

FS module (EAG-RS) outperformed all other methods in terms of all metrics. Based on these promising results, we conclude that the steps adopted in the proposed EAG-RS play pivotal roles in classifying TD and ASD. The two ROI-representative vectors can extract appropriate features. Moreover, they are based on the connection-wise relevance score, which not only enables diagnosis based on the original feature representations of FC but also serves as criteria for the extraction of diagnosis-informative ROIs based on feature selection.

We further report and compare the classification results obtained from state-of-the-art methods on the ABIDE dataset to demonstrate the superiority of our proposed EAG-RS, as illustrated in Table III and Table IV. In Table IV, a fair comparison is ensured by re-implementing all the methods using the same experimental configurations as those used in our study.

V. DISCUSSION

A. The ratio of random ROI-level masking

First, the ratio of ROI-level masking is varied from 0 to 0.9 at intervals of 0.1. The corresponding performance results are presented in Fig. 4. When the ROI-level masking is $q = 0.1$ and $q = 0.2$, the performance is better than that obtained without ROI-level masking, indicating a positive influence of ROI-level masking on diagnostic performance. For this reason, $q = 0.1$ is selected for subsequent experiments, as it corresponds to the highest performance quality.

B. Ablation Study

We conduct additional experiments to validate the effectiveness of the proposed framework. We estimate features via six ablation cases in the context of a classification task. In Case I, the ROI selection network is removed and only FC features are used to classify the brain diseases using the multi-layer perceptron (MLP) (Case I = Baseline). In Case II, the ROI selection network is removed and an estimated ROI representative vector, $\mathbf{f}_v$, is used to classify brain diseases. Under this setting, the number of dimensions is different from that in the original FC; thus, brain diseases are classified using an SVM. In Case III, the other estimated ROI representative vector, $\mathbf{f}_c$, is used for classification. In Case IV, the original FC is used and the two ROI representative vectors (*i.e.*, $\mathbf{f} = [\mathbf{f}_v||\mathbf{f}_c]$) are concatenated. Henceforth, we use an ROI selection network. Therefore, the original FC is implemented along with $\mathbf{f}_v$ (Case V), $\mathbf{f}_c$ (Case VI), and the concatenation of the two ROI representative vectors, $\mathbf{f}$ (EAG-RS).

TABLE V
AVERAGED CLASSIFICATION PERFORMANCES FOR ASD AND TD IN THE ABLATION STUDY.

Case	AUC	ACC (%)	SEN (%)	SPEC (%)
I	0.757±0.040	69.76±3.45	57.82±6.64	80.22±4.54
II	0.552±0.020	48.53±4.58	55.35±2.19	55.75±2.41
III	0.666±0.029	59.35±2.91	50.32±9.62	68.59±7.33
IV	**0.760±0.018**	71.01±1.77	**65.60±7.50**	75.40±6.59
V	0.747±0.028	70.64±3.28	65.25±5.04	75.29±5.29
VI	**0.760±0.025**	71.09±3.29	64.46±5.74	78.26±6.01
EAG-RS	**0.760±0.033**	**73.71±2.83**	64.56±8.36	**80.74±8.28**

As reported in Table V, the proposed framework achieve the best diagnostic performance among all ablation cases. Cases without an ROI selection network (Case I, II, and III) are observed to exhibit lower classification performance. When independent representative vectors, $\mathbf{f}_v$ (Case II) and $\mathbf{f}_c$ (Case III), are used, slightly lower performance than that of the original FC (Case I) is observed. Therefore, the features are combined using concatenation to confirm the effectiveness of the representative features (Case IV), which improves the performance. However, the performance is observed to be degraded when each representative vector is used with an ROI selection network (Case V and VI). Thus, the count and value information aid the extraction of diagnosis-informative ROI information, which improves classification performance by removing redundant and irrelevant features of the original FC.

C. Analysis of ROI Selection Network

The best performing model on the validation dataset is analyzed further. On average, 22 ROIs (median, 23) are selected for the ASD group, and 40 ROIs (median, 43) from the TD group. To visualize the selection ratio (SR) of each ROI, the ROIs selected from the ψ module are enumerated and the total number is divided by the number of ROIs and the number of subjects in each group (Fig. 3). In particular, ROIs with SR exceeding 0.5 as considered (Fig. 3(a)). These ROIs are further categorized into two cases: i) $0.5 < \text{SR} < 0.75$ (Fig. 3(b)) and ii) $\text{SR} > 0.75$ (Fig. 3(c)). The majority of selected brain regions are associated with ASD. As depicted in Fig. 3(b), 12 brain

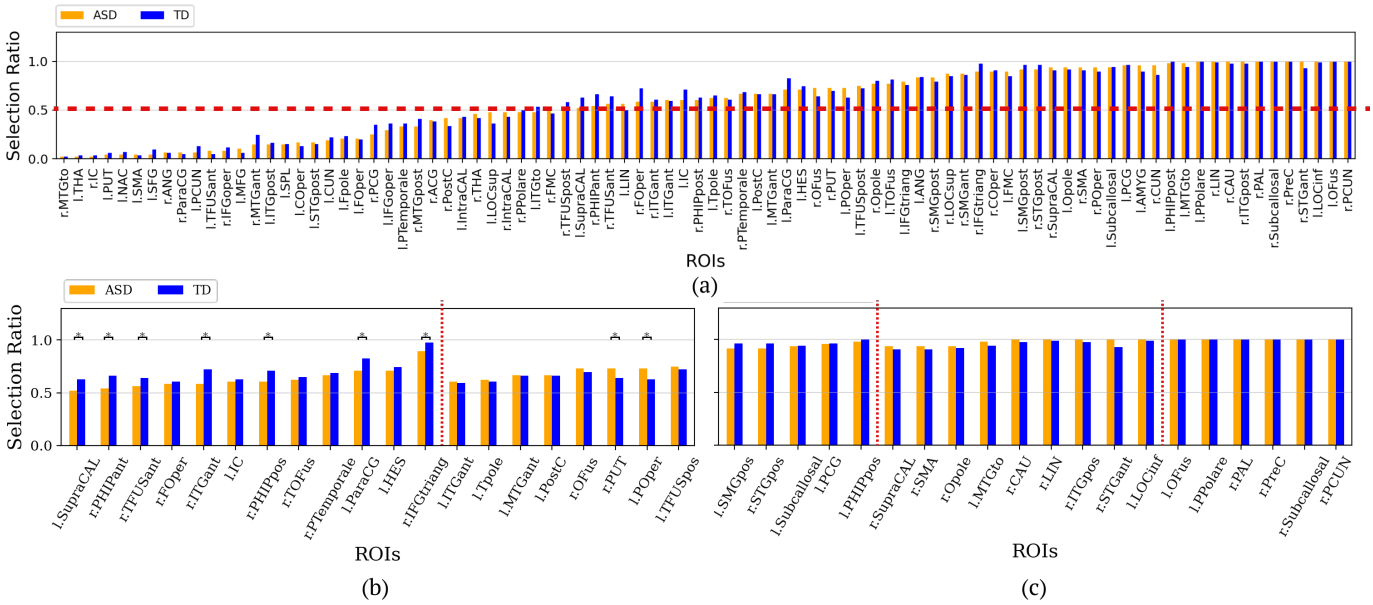

Fig. 3. (a) Visualization of the selection ratio (SR) of ROIs for each group at an ROI-level. Although we used 110 ROIs are used for this analysis, few had values of zero (*i.e.*, not selected). In addition, we set **0.5** as the threshold to consider the general patterns in SR. (b) The list of brain regions in each group ($0.5 < \text{SR} < 0.75$). The left and right sides of the red vertical line correspond to the TD and ASD groups, respectively. (c) The list of brain regions in each group ($\text{SR} > 0.75$). The left, middle, and right portions of the red vertical line correspond to the TD, ASD, and common groups, respectively.

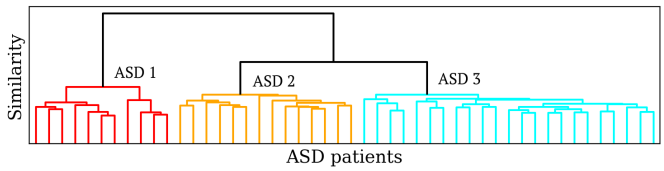

Fig. 4. The Y-axis represents the similarity between patients, *i.e.*, the shorter the distance, the greater the similarity between patients—the similarity threshold is set at 0.3. The yellow, green, and red lines represent the clustered subtypes of ASD.

regions lie on the left side of the red vertical line with slightly higher SR in the TD group, while eight regions lie on the right side of the vertical line with higher SR in the ASD group. In Fig. 3(c), five brain regions lie on the left side of the vertical line with slightly higher SR in the TD group, while nine regions lie on the middle side of the vertical line with higher SR in the ASD group. The SR patterns reveal that 12 specific brain regions are always selected, with six brain regions on the right side of the vertical line (Fig. 3(c)) chosen consistently in both the TD and ASD groups. In the TD group, the "l.PHIPpos" region is selected 100% of the time, while in the ASD group, five specific brain regions ("r.CAU," "r.LIN," "l.ITGpos," "r.STGant," and "r.LOCinf") are chosen consistently.

Group analysis is performed based on these selected ROIs to perform a neuroscientific analysis of ASD and TD groups at the ROI level. Subsequently, we identify nine brain regions with significantly different selection frequency between the two groups by measuring the difference (<0.05). Remarkably, the proposed framework, trained without prior knowledge, is observed to identify brain regions highly related to existing neuroscience studies. Specifically, it identifies the following brain regions: 'l.SupraCAL,' 'r.PHIPant,' 'r.TFUSant,' 'r.ITGant,' 'r.PHIPpos,' 'l.ParaCG,' 'r.IFGtriang,' 'r.PUT,' 'l.POper.' These regions are marked with asterisks (*) in Fig. 3(b). Notably, the left supracalcarine cortex (l.SupraCAL), situated in the visual cortex, is involved in various visual processes such as discerning object shape, size, and color, and motion perception [54]. Similarly, the right inferior temporal gyrus (r.ITGant) plays a similar role [55]. In addition, the right temporal fusiform cortex (r.TFUSant) in the temporal lobe is responsible for facial processing and recognition. It distinguishes facial features and supports social interactions and recognition skills [56]. The right parahippocampal gyrus (r.PHIPant and r.PHIPpos) in the medial temporal lobe is vital to memory, spatial navigation, and emotional processing [57]. The left paracingulate gyrus (l.ParaCG), which is a part of the cingulate cortex, contributes to various cognitive and emotional functions. The right putamen (r.PUT), a basal ganglia structure, participates in motor control, procedural learning, reward processing, attention, and cognition. The left parietal operculum cortex (l.POper) has broad involvement in somatosensory processing, language, and multisensory integration, supporting sensory perception, communication, and body awareness. Finally, the right inferior frontal gyrus triangular (r.IFGtriang) region contains Broca's area, essential for language processing like production, comprehension, and sophisticated inhibitory control. This region's crucial role in language-related functions and cognitive processes is well-documented [58]. This confirmation of the biological relevance and interpretability of our findings further

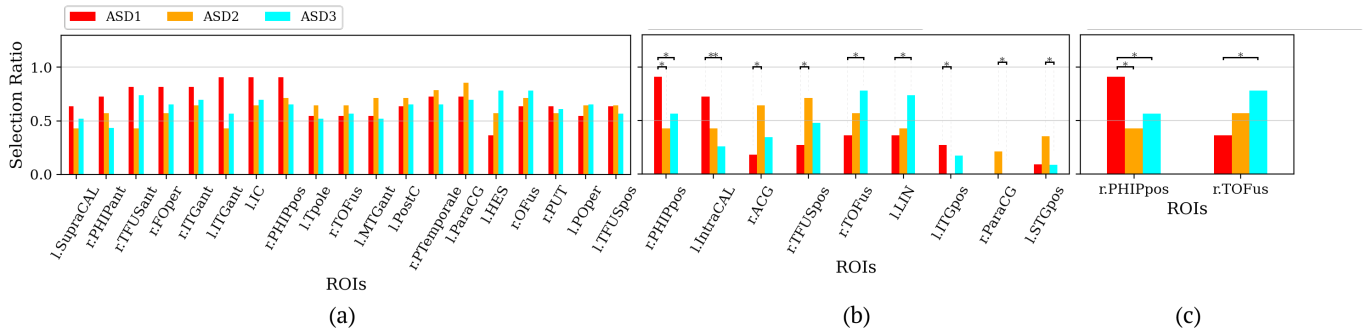

Fig. 5. (a) Visualization of the SR of ROIs corresponding to the three subtypes of ASD at ROI-level. The list of brain regions in the ASD group corresponds to $\mathbf{0.5} < \text{SR} < \mathbf{0.75}$. (b) Statistical test results for three subtypes of ASD analysis using ROIs selected by the ROI selection network. Only statistically significant ROIs are visualized. Note that $*, **$ represent $\boldsymbol{p} < \mathbf{0.05}$ and $\boldsymbol{p} < \mathbf{0.01}$, respectively. (c) The list of common brain regions captured between (a) and (b).

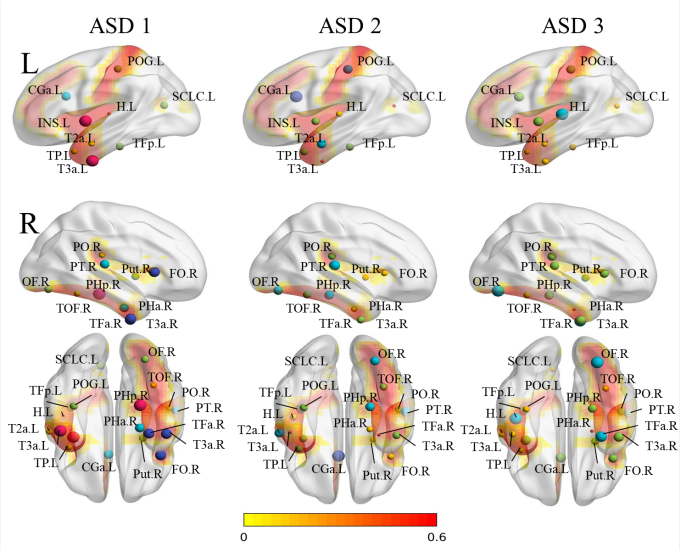

Fig. 6. Visualization of 3D brain image representing SRs for each ASD subtype. Each circular marker indicates a specific brain region, and it is assigned a color based on the range of SR (SR < 0.5 (red), 0.5 < SR < 0.6 (yellow), 0.6 < SR < 0.7 (light green), 0.7 < SR < 0.8 (cyan), 0.8 < SR < 0.9 (dark blue), SR > 0.9 (magenta)).

TABLE VI
CHARACTERISTICS OF DEMOGRAPHIC, PSYCHOLOGICAL, AND EDUCATIONAL ASSESSMENTS FOR TD AND ASD SUBTYPES.

Categories	TD	ASD1	ASD2	ASD3	F	p-value (>F)
Age (years)	15.46	12.90	15.69	15.24	0.75	0.528
Male (%)	76.92	81.82	92.86	87.50	0.80	0.497
Female (%)	23.08	18.18	7.14	12.50		
FIQ	113.18	97.95	99.61	94.81	10.31	<0.001
VIQ	114.14	101.09	99.86	96.69	5.89	<0.001
PIQ	109.63	98.64	100.36	92.75	9.42	<0.001

ASD, autism spectrum disorder; F, F-value; FIQ, full-scale intelligence quotient; VIQ, verbal IQ; PIQ, performance IQ;

supports the conclusion that it is effective in identifying important brain regions associated with target disorders.

D. Clustering Subtypes in Autism Spectrum Disorder

ASD is known for its diverse and heterogeneous nature, with various characteristics associated with ASD-related brain regions, and varying symptom severity levels and comorbidities [59]. As a result, several previous research projects in ASD have aimed to identify distinct behavioral subtypes within ASD populations. In this study, we consider such heterogeneity among ASD groups during ROI selection. To explore this heterogeneity further, subtype analysis is performed using hierarchical clustering with ward linkage, a commonly used method in neuroimaging. The results reveal three subtypes of ASD (Fig. 4), each with unique characteristics and functional connectivity patterns. Further, we investigate demographic information, such as age and gender, of each identified ASD subtype using statistical tests (Table VI). In Table VI, the average age of TD individuals is observed to be 15.5 years. Among the clustered ASD subtypes, ASD1 exhibits an average age of 12.9 years; ASD2, 15.7 years; and ASD3, 15.2 years. In the case of gender distribution, the TD group exhibits a female proportion of 23.1%. For the identified ASD subtypes, ASD1 exhibits a female proportion of 18.2%; ASD2, 7.1%; and ASD3, 12.5%. Further, in terms of Full-Scale Intelligence Quotient (FIQ), Verbal Intelligence Quotient (VIQ), and Performance Intelligence Quotient (PIQ), TD individuals exhibit average scores of 113.2, 114.1, and 109.6, respectively. In comparison, the ASD1 subtype exhibits average scores of 98.0 (FIQ), 101.1 (VIQ), and 98.6 (PIQ), respectively; ASD2, 99.6 (FIQ), 99.9 (VIQ), and 100.4 (PIQ), respectively; and ASD3, 94.8 (FIQ), 96.7 (VIQ), and 92.8 (PIQ), respectively. This analysis provides deep insight into the potential relationships between the identified subtypes and various demographic factors, as well as their associations with psychological and educational assessments.

Based on our analysis of the selected ROIs depicted in Fig. 3(b), they are sorted based on their SRs for each subtype of ASD. In addition, the significant regions associated with ASD subtypes are visualized in Fig. 5(b). Two common regions are observed—the right posterior parahippocampus (r.PHIPpos) and right temporal occipital fusiform cortex (r.TOFus). These are known to be associated with ASD, as depicted in Fig. 5(a) and Fig 5(b). Interestingly, as presented in Fig. 5, we observed different SRs for different ASD subtypes in these regions. Finally, to provide a more intuitive understanding, the SRs corre-

sponding to each ASD subtype are mapped onto 3D brain images and visualized in Fig. 6. In Fig. 6, the brain regions are represented by circular markers, and the color of each marker corresponds to the range of SR values for that specific region. These results provide beneficial insights into ASD subtyping from a neuroscientific viewpoint. The proposed ROI selection network successfully identifies and differentiates specific brain regions associated with different ASD subtypes. We believe that this capability enhances our understanding of the biological basis of the heterogeneity within the ASD population and may have significant implications for the advancement of subtype analysis in autism research.

VI. Conclusion

In this work, we propose a novel explainability-guided ROI selection (EAG-RS) framework that dynamically selects informative features at the ROI-level for brain disease diagnosis. Our EAG-RS framework learns inter-regional relationships using random seed-based network masking to estimate non-linear relationships, representing other neighboring connections to restore masked seed-ROI connections. We also estimated connection-wise relevance scores to explore high-order relationships between FCs using LRP. Finally, we utilized the estimated non-linear high-order FCs to select diagnosis-informative ROIs and diagnose brain disease simultaneously. To demonstrate its validity, ASD diagnosis was performed using the proposed EAG-RS framework on the ABIDE dataset. Furthermore, the cluster subtypes for ASD were identified based on individually selected ROIs. The results demonstrate that our EAG-RS framework provides new neuroscientific insights into ASD subtypes and their biomarkers.

However, this study suffers from the following practical limitations. First, regarding the architectural design, MLPs were used for the encoder-decoder structure in inter-regional relation learning and the ROI selection network. Although MLPs provide flexibility and expressiveness, they involve a high number of tunable parameters. Given the scarcity of neuroimaging data and labels, optimization of such MLP-based architectures may be challenging, requiring strong regularization. To address this limitation, alternative approaches may be explored, *e.g.*, incorporating convolutional neural networks and transformers into each module of the proposed framework. Second, as the sole focus of this study was the use of FC for ASD diagnosis, it did not incorporate other neuroimaging modalities or clinical information. Integrating multiple modalities and clinical data could potentially provide complementary insights and improve the accuracy and robustness of ASD diagnosis. Addressing the abovementioned limitations in the future will contribute to a more comprehensive understanding of the application of FC in ASD diagnosis and enhance the effectiveness and interpretability of the proposed framework.

References

[1] C. Ecker, A. Marquand, J. Mourão-Miranda, P. Johnston, E. M. Daly, M. J. Brammer, S. Maltezos, C. M. Murphy, D. Robertson, S. C. Williams *et al.*, "Describing the brain in autism in five dimensions—magnetic resonance imaging-assisted diagnosis of autism spectrum disorder using a multiparameter classification approach," *Journal of Neuroscience*, vol. 30, no. 32, pp. 10 612–10 623, 2010.

[2] A. V. Buescher, Z. Cidav, M. Knapp, and D. S. Mandell, "Costs of autism spectrum disorders in the united kingdom and the united states," *JAMA pediatrics*, vol. 168, no. 8, pp. 721–728, 2014.

[3] E. Fernell, M. A. Eriksson, and C. Gillberg, "Early diagnosis of autism and impact on prognosis: a narrative review," *Clinical epidemiology*, vol. 5, p. 33, 2013.

[4] L. Zwaigenbaum, M. L. Bauman, W. L. Stone, N. Yirmiya, A. Estes, R. L. Hansen, J. C. McPartland, M. R. Natowicz, R. Choueiri, D. Fein *et al.*, "Early identification of autism spectrum disorder: Recommendations for practice and research," *Pediatrics*, vol. 136, no. Supplement_1, pp. S10–S40, 2015.

[5] A. Abraham, M. P. Milham, A. Di Martino, R. C. Craddock, D. Samaras, B. Thirion, and G. Varoquaux, "Deriving reproducible biomarkers from multi-site resting-state data: An autism-based example," *NeuroImage*, vol. 147, pp. 736–745, 2017.

[6] J. Kim, V. D. Calhoun, E. Shim, and J.-H. Lee, "Deep neural network with weight sparsity control and pre-training extracts hierarchical features and enhances classification performance: Evidence from whole-brain resting-state functional connectivity patterns of schizophrenia," *Neuroimage*, vol. 124, pp. 127–146, 2016.

[7] M. R. Brier, J. B. Thomas, A. Z. Snyder, T. L. Benzinger, D. Zhang, M. E. Raichle, D. M. Holtzman, J. C. Morris, and B. M. Ances, "Loss of intranetwork and internetwork resting state functional connections with alzheimer's disease progression," *Journal of Neuroscience*, vol. 32, no. 26, pp. 8890–8899, 2012.

[8] I. Cribben, R. Haraldsdottir, L. Y. Atlas, T. D. Wager, and M. A. Lindquist, "Dynamic connectivity regression: determining state-related changes in brain connectivity," *Neuroimage*, vol. 61, no. 4, pp. 907–920, 2012.

[9] N. C. Dvornek, P. Ventola, K. A. Pelphrey, and J. S. Duncan, "Identifying autism from resting-state fmri using long short-term memory networks," in *International Workshop on Machine Learning in Medical Imaging*. Springer, 2017, pp. 362–370.

[10] E. Kang and H.-I. Suk, "Probabilistic source separation on resting-state fmri and its use for early mci identification," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2018, pp. 275–283.

[11] E. Jeon, E. Kang, J. Lee, J. Lee, T.-E. Kam, and H.-I. Suk, "Enriched representation learning in resting-state fmri for early mci diagnosis," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2020, pp. 397–406.

[12] X. Chen, H. Zhang, Y. Gao, C.-Y. Wee, G. Li, D. Shen, and A. D. N. Initiative, "High-order resting-state functional connectivity network for mci classification," *Human brain mapping*, vol. 37, no. 9, pp. 3282–3296, 2016.

[13] F. Zhao, Z. Chen, I. Rekik, S.-W. Lee, and D. Shen, "Diagnosis of autism spectrum disorder using central-moment features from low-and high-order dynamic resting-state functional connectivity networks," *Frontiers in neuroscience*, vol. 14, 2020.

[14] B. B. Biswal, "Resting state fmri: a personal history," *Neuroimage*, vol. 62, no. 2, pp. 938–944, 2012.

[15] L. Lee, L. M. Harrison, and A. Mechelli, "A report of the functional connectivity workshop, dusseldorf 2002," *Neuroimage*, vol. 19, no. 2, pp. 457–465, 2003.

[16] M. P. Van Den Heuvel and H. E. H. Pol, "Exploring the brain network: a review on resting-state fmri functional connectivity," *European neuropsychopharmacology*, vol. 20, no. 8, pp. 519–534, 2010.

[17] T. D. Wager and C.-W. Woo, "Imaging biomarkers and biotypes for depression," *Nature medicine*, vol. 23, no. 1, pp. 16–17, 2017.

[18] S. Parisot, S. I. Ktena, E. Ferrante, M. Lee, R. Guerrero, B. Glocker, and D. Rueckert, "Disease prediction using graph

convolutional networks: application to autism spectrum disorder and alzheimer's disease," *Medical image analysis*, vol. 48, pp. 117–130, 2018.

[19] A. S. Heinsfeld, A. R. Franco, R. C. Craddock, A. Buchweitz, and F. Meneguzzi, "Identification of autism spectrum disorder using deep learning and the abide dataset," *NeuroImage: Clinical*, vol. 17, pp. 16–23, 2018.

[20] H.-I. Suk, C.-Y. Wee, S.-W. Lee, and D. Shen, "State-space model with deep learning for functional dynamics estimation in resting-state fmri," *NeuroImage*, vol. 129, pp. 292–307, 2016.

[21] X. Guo, K. C. Dominick, A. A. Minai, H. Li, C. A. Erickson, and L. J. Lu, "Diagnosing autism spectrum disorder from brain resting-state functional connectivity patterns using a deep neural network with a novel feature selection method," *Frontiers in neuroscience*, vol. 11, p. 460, 2017.

[22] T. Eslami, V. Mirjalili, A. Fong, A. R. Laird, and F. Saeed, "Asd-diagnet: a hybrid learning approach for detection of autism spectrum disorder using fmri data," *Frontiers in neuroinformatics*, vol. 13, p. 70, 2019.

[23] C. Wang, Z. Xiao, B. Wang, and J. Wu, "Identification of autism based on svm-rfe and stacked sparse auto-encoder," *Ieee Access*, vol. 7, pp. 118 030–118 036, 2019.

[24] M. Rakić, M. Cabezas, K. Kushibar, A. Oliver, and X. Llado, "Improving the detection of autism spectrum disorder by combining structural and functional mri information," *NeuroImage: Clinical*, vol. 25, p. 102181, 2020.

[25] W. Jung, D.-W. Heo, E. Jeon, J. Lee, and H.-I. Suk, "Inter-regional high-level relation learning from functional connectivity via self-supervision," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2021, pp. 284–293.

[26] C.-Y. Wee, P.-T. Yap, D. Zhang, K. Denny, J. N. Browndyke, G. G. Potter, K. A. Welsh-Bohmer, L. Wang, and D. Shen, "Identification of mci individuals using structural and functional connectivity networks," *Neuroimage*, vol. 59, no. 3, pp. 2045–2056, 2012.

[27] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[28] M. Wang, B. Jie, W. Bian, X. Ding, W. Zhou, Z. Wang, and M. Liu, "Graph-kernel based structured feature selection for brain disease classification using functional connectivity networks," *IEEE Access*, vol. 7, pp. 35 001–35 011, 2019.

[29] Q. Zhu, H. Li, J. Huang, X. Xu, D. Guan, and D. Zhang, "Hybrid functional brain network with first-order and second-order information for computer-aided diagnosis of schizophrenia," *Frontiers in neuroscience*, vol. 13, p. 603, 2019.

[30] H. Zhang, X. Chen, F. Shi, G. Li, M. Kim, P. Giannakopoulos, S. Haller, and D. Shen, "Topographical information-based high-order functional connectivity and its application in abnormality detection for mild cognitive impairment," *Journal of Alzheimer's Disease*, vol. 54, no. 3, pp. 1095–1112, 2016.

[31] F. Zhao, X. Zhang, K.-H. Thung, N. Mao, S.-W. Lee, and D. Shen, "Constructing multi-view high-order functional connectivity networks for diagnosis of autism spectrum disorder," *IEEE Transactions on Biomedical Engineering*, vol. 69, no. 3, pp. 1237–1250, 2021.

[32] Y. Zhang, H. Zhang, X. Chen, S.-W. Lee, and D. Shen, "Hybrid high-order functional connectivity networks using resting-state functional mri for mild cognitive impairment diagnosis," *Scientific reports*, vol. 7, no. 1, p. 6530, 2017.

[33] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PloS one*, vol. 10, no. 7, p. e0130140, 2015.

[34] A. Di Martino, C.-G. Yan, Q. Li, E. Denio, F. X. Castellanos, K. Alaerts, J. S. Anderson, M. Assaf, S. Y. Bookheimer, M. Dapretto *et al.*, "The Autism Brain Imaging Data Exchange: Towards a Large-scale Evaluation of The Intrinsic Brain Architecture in Autism," *Molecular Psychiatry*, vol. 19, no. 6, pp. 659–667, 2014.

[35] J. Liu, Y. Sheng, W. Lan, R. Guo, Y. Wang, and J. Wang, "Improved asd classification using dynamic functional connectivity and multi-task feature selection," *Pattern Recognition Letters*, vol. 138, pp. 82–87, 2020.

[36] W. Yan, S. Plis, V. D. Calhoun, S. Liu, R. Jiang, T.-Z. Jiang, and J. Sui, "Discriminating schizophrenia from normal controls using resting state functional network connectivity: A deep neural network and layer-wise relevance propagation method," in *2017 IEEE 27th international workshop on machine learning for signal processing (MLSP)*. IEEE, 2017, pp. 1–6.

[37] T. Azevedo, A. Campbell, R. Romero-Garcia, L. Passamonti, R. A. Bethlehem, P. Liò, and N. Toschi, "A deep graph neural network architecture for modelling spatio-temporal dynamics in resting-state functional mri data," *Medical Image Analysis*, vol. 79, p. 102471, 2022.

[38] M. Zhao, W. Yan, N. Luo, D. Zhi, Z. Fu, Y. Du, S. Yu, T. Jiang, V. D. Calhoun, and J. Sui, "An attention-based hybrid deep learning framework integrating brain connectivity and activity of resting-state functional mri data," *Medical image analysis*, vol. 78, p. 102413, 2022.

[39] Q.-H. Lin, Y.-W. Niu, J. Sui, W.-D. Zhao, C. Zhuo, and V. D. Calhoun, "Sspnet: An interpretable 3d-cnn for classification of schizophrenia using phase maps of resting-state complex-valued fmri data," *Medical Image Analysis*, vol. 79, p. 102430, 2022.

[40] S. Dang and S. Chaudhury, "Novel relative relevance score for estimating brain connectivity from fmri data using an explainable neural network approach," *Journal of Neuroscience Methods*, vol. 326, p. 108371, 2019.

[41] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16 000–16 009.

[42] S. Gholizadeh and N. Zhou, "Model explainability in deep learning based natural language processing," *arXiv preprint arXiv:2106.07410*, 2021.

[43] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[44] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proceedings of the 25th international conference on Machine learning*, 2008, pp. 1096–1103.

[45] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, P.-A. Manzagol, and L. Bottou, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion." *Journal of machine learning research*, vol. 11, no. 12, 2010.

[46] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2536–2544.

[47] Q. McNemar, "Note on the Sampling Error of the Difference between Correlated Proportions or Percentages," *Psychometrika*, vol. 12, no. 2, pp. 153–157, 1947.

[48] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine Learning*, vol. 46, pp. 389–422, 2002.

[49] J. Kawahara, C. J. Brown, S. P. Miller, B. G. Booth, V. Chau, R. E. Grunau, J. G. Zwicker, and G. Hamarneh, "Brainnetcnn: Convolutional neural networks for brain networks; towards predicting neurodevelopment," *NeuroImage*, vol. 146, pp. 1038–1049, 2017.

[50] X. Li, Y. Zhou, N. Dvornek, M. Zhang, S. Gao, J. Zhuang, D. Scheinost, L. H. Staib, P. Ventola, and J. S. Duncan, "Braingnn: Interpretable brain graph neural network for fmri analysis," *Medical Image Analysis*, vol. 74, p. 102233, 2021.

[51] X. Kan, W. Dai, H. Cui, Z. Zhang, Y. Guo, and C. Yang, "Brain network transformer," *Advances in Neural Information Processing Systems*, vol. 35, pp. 25 586–25 599, 2022.

[52] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, "Greedy layer-wise training of deep networks," *Advances in neural information processing systems*, vol. 19, 2006.

[53] Y. Liu, L. Xu, J. Li, J. Yu, and X. Yu, "Attentional connectivity-based prediction of autism using heterogeneous rs-fmri data from cc200 atlas," *Experimental neurobiology*, vol. 29, no. 1, p. 27, 2020.

[54] G. S. Dichter, J. N. Felder, and J. W. Bodfish, "Autism is characterized by dorsal anterior cingulate hyperactivation during social target detection," *Social cognitive and affective neuroscience*, vol. 4, no. 3, pp. 215–226, 2009.

[55] D. Kim, J. Y. Lee, B. C. Jeong, J.-H. Ahn, J. I. Kim, E. S. Lee, H. Kim, H. J. Lee, and C. E. Han, "Overconnectivity of the right heschl's and inferior temporal gyrus correlates with symptom

severity in preschoolers with autism spectrum disorder," *Autism Research*, vol. 14, no. 11, pp. 2314–2329, 2021.

[56] I. A. van Kooten, S. J. Palmen, P. von Cappeln, H. W. Steinbusch, H. Korr, H. Heinsen, P. R. Hof, H. van Engeland, and C. Schmitz, "Neurons in the fusiform gyrus are fewer and smaller in autism," *Brain*, vol. 131, no. 4, pp. 987–999, 2008.

[57] C. S. Monk, S. J. Peltier, J. L. Wiggins, S.-J. Weng, M. Carrasco, S. Risi, and C. Lord, "Abnormalities of intrinsic functional connectivity in autism spectrum disorders," *Neuroimage*, vol. 47, no. 2, pp. 764–772, 2009.

[58] V. Yuk, C. Urbain, E. W. Pang, E. Anagnostou, D. Buchsbaum, and M. J. Taylor, "Do you know what i'm thinking? temporal and spatial brain activity during a theory-of-mind task in children with autism," *Developmental cognitive neuroscience*, vol. 34, pp. 139–147, 2018.

[59] E. Moradi, B. Khundrakpam, J. D. Lewis, A. C. Evans, and J. Tohka, "Predicting symptom severity in autism spectrum disorder based on cortical thickness measures in agglomerative data," *Neuroimage*, vol. 144, pp. 128–141, 2017.