Diagnosis-Guided Deep Subspace Clustering Association Study for Pathogenetic Markers Identification of Alzheimer's Disease Based on Comparative Atlases

Cui-Na Jiao, *Graduate Student Member, IEEE*, Junliang Shang, *Member, IEEE*, Feng Li, Xinchun Cui, Yan-Li Wang, Ying-Lian Gao, and Jin-Xing Liu, *Member, IEEE*

***Abstract*—The roles of brain region activities and genotypic functions in the pathogenesis of Alzheimer's disease (AD) remain unclear. Meanwhile, current imaging genetics methods are difficult to identify potential pathogenetic markers by correlation analysis between brain network and genetic variation. To discover disease-related brain connectome from the specific brain structure and the fine-grained level, based on the Automated Anatomical Labeling (AAL) and human Brainnetome atlases, the functional brain network is first constructed for each subject. Specifically, the upper triangle elements of the functional connectivity matrix are extracted as connectivity features. The clustering coefficient and the average weighted node degree are developed to assess the significance of every brain area. Since the constructed brain network and genetic data are characterized by non-linearity, high-dimensionality, and few subjects, the deep subspace clustering algorithm is proposed to reconstruct the original data. Our multilayer neural network helps capture the non-linear manifolds, and subspace clustering learns pairwise affinities between samples. Moreover, most approaches in neuroimaging genetics are unsupervised learning, neglecting the diagnostic information related to diseases. We presented a label constraint with diagnostic status to instruct the imaging genetics correlation analysis. To this end, a diagnosis-guided deep subspace clustering association (DDSCA) method is developed to discover brain connectome and risk genetic factors by integrating genotypes with functional network phenotypes. Extensive experiments prove that DDSCA achieves superior performance to most association methods and effectively selects disease-relevant genetic markers and brain connectome at the coarse-grained and fine-grained levels.**

***Index Terms*—Brain imaging genetics, deep subspace clustering, functional connectivity network, diagnosis information, sparse canonical correlation analysis.**

I. Introduction

ALZHEIMER'S disease (AD) is considered the main cause of the cognitive decline, and the incidence of AD in the elderly worldwide lies at a high level [1]. Before presenting obvious symptoms, such as memory decline, further causing speech and sleep disorders, and thought confusion, it requires several years of slow degeneration of the brain, and this degeneration is irreversible. Therefore, how to diagnose this disease at an early stage and alleviate the progression of AD through timely treatment is a hot research topic.

As a brain disease, AD mainly manifests in two aspects. One is the degeneration of the brain structures and functions in the macrolevel and the other is the potential heritability in the microlevel. Brain imaging genetics is a rising field of brain research in recent years, which integrated analysis of macroscopic brain imaging data and microscopic genetic data [2]. Noting that single nucleotide polymorphism (SNP) as genetic data is the genomic variant at a single base position in the deoxyribonucleic acid (DNA). Meanwhile, brain imaging quantitative trait (QT) is the measurable phenotype of the human brain. Our work aimed to jointly analyze SNPs and brain QTs and further study associations between them, which could provide new insights into the phenotypic, genetic, and molecular characteristics of the brain. Furthermore, distinct brain atlases result in distinct partitions in the aspect of the number of regions and the size and location of these regions in the brain. To verify the association performance with distinct atlases and identify brain markers

Manuscript received 25 February 2023; revised 3 December 2023 and 24 January 2024; accepted 24 February 2024. Date of publication 4 March 2024; date of current version 7 May 2024. This work was supported by the National Natural Science Foundation of China under Grant 62172254. *(Corresponding author: Jin-Xing Liu.)*

Cui-Na Jiao is with the School of Engineering, Qufu Normal University, Rizhao 276826, China, and also with the School of Health and Life Sciences, University of Health and Rehabilitation Sciences, Qingdao 266113, China (e-mail: jiaocuina123@163.com).

Junliang Shang and Feng Li are with the School of Computer Science, Qufu Normal University, Rizhao 276826, China (e-mail: shangjunliang110@163.com; lifeng_10_28@163.com).

Xinchun Cui is with the School of Foundational Education, University of Health and Rehabilitation Sciences, Qingdao 266113, China (e-mail: cuixinchun@uor.edu.cn).

Yan-Li Wang is with the Qingdao Central Hospital, University of Health and Rehabilitation Sciences, Qingdao 266113, China (e-mail: wangyanli1105@163.com).

Ying-Lian Gao is with the Library of Qufu Normal University, Rizhao 276826, China (e-mail: yinliangao@126.com).

Jin-Xing Liu is with the School of Health and Life Sciences, University of Health and Rehabilitation Sciences, Qingdao 266113, China (e-mail: sdcavell@126.com).

Digital Object Identifier 10.1109/JBHI.2024.3372294

at the coarse-grained and fine-grained levels, we focused on developing a novel imaging genetics method to discover risk SNPs and brain connectome by integrating genetics data and functional brain network constructed based on two different atlases.

For neuroimaging genetics studies, early works mainly concentrated on pairwise univariate analysis for exploring associations between SNPs and QTs. The univariate approaches regard every SNP or region of interest (ROI) as an independent variable, which neglected the complex relations between them [3]. Later, the bi-multivariate analysis is developed to explore complex correlations between multiple SNPs and multiple QTs. For instance, Witten et al. developed the sparse canonical correlation analysis (SCCA) [4] with the lasso penalty to obtain sparse association results in imaging genetics. After that, SCCA and its variants gained increasing attention because of their simplicity and efficiency. In 2014, a knowledge-guided SCCA (KG-SCCA) was introduced for utilizing prior structures to capture weight similarity between group features [5]. The discriminative SCCA (DSCCA) [6] was then presented by Yan et al., which introduced a penalty term with the diagnostic information of data to explore imaging-proteomics associations. To further reflect the non-linear relationships of the original data, kernel CCA (KCCA) [7] as a non-linear extension of CCA was proposed to learn meaningful representations of data. Shortly after, the deep CCA (DCCA) [8] was developed to achieve the non-linear transformations of two views of data.

Currently, to further comprehend the interaction mechanism among SNPs and QTs, several scholars used genes and brain regions as nodes to construct networks for exploring the functional correlations between pathogenetic factors. For instance, literature [9] reviewed the corresponding studies on association analysis between genes and brain functional connections. Wang et al. adopted a multi-modality regression framework [10] to explore network features related to a risk SNP and disease status by using network prior knowledge. Bi et al. also designed a fusion network to integrate brain regions with genes, and presented a genetic evolutionary random forest (GERF) approach [11] to solve problems of small sample size and high-dimensional features. Moreover, in 2022, they further developed a graph convolution model to enrich node features of the network by feature aggregation [12]. Though the above constructed networks are associated with genotypic and brain disease, they only attached importance to identifying disease-related brain imaging markers of the network at the coarse-grained level.

As each brain ROI may be involved in many anatomical regions at the same time, to further improve the anatomical interpretability, atlas-based partition feature extraction methods with a predefined brain atlas are used to extract brain features. Presently, the automated anatomical labeling (AAL) atlas is the most popular brain atlas, which has been widely applied to explore a variety of brain diseases in recent years [6], [13]. Since AAL brain atlas covers mainly specific structures and lacks fine-grained parcellations, to further provide functionally important connectivity information of the brain, a connectivity-based Brainnetome (BN) atlas that identified subdivisions of the entire human brain is developed to reveal the in vivo connectivity architecture. Distinct brain atlases provide distinct partitions based on the number of regions and the size and locations of these regions in the brain. Up to now, few studies verified the association performance using different brain atlases, and no study has employed the BN atlas to explore correlations between genotypes and brain network phenotypes. Here, we aimed to employ the two mentioned brain atlases to perform association analysis between brain network QTs and SNPs and extract biomarkers for AD diagnosis.

In our work, to further explore connections among brain areas, the brain functional connectivity network is constructed based on resting-state functional magnetic resonance imaging (fMRI) data. Nodes of the brain network denote ROIs and edges denote the connectivity among brain areas [14]. Specifically, to further identify the disease-related connectome in brain networks from the coarse-grained and fine-grained levels, the fMRI data are parcellated by the AAL atlas and human BN atlas, respectively. Then, average the time series of every brain region and construct a functional connectivity (FC) matrix for each subject. The node importance is calculated using the graph theory indicators. Meanwhile, edge features are extracted from the FC matrix to reflect the connection strength between brain areas. Based on the brain network constructed from the coarse-grained and fine-grained levels, this paper aims to comprehensively select the brain connectome (including node and edge features) and risk SNPs by correlation analysis between SNPs and fMRI imaging data.

Since the AD Neuroimaging Initiative (ADNI) data involve high-dimensional features and a limited number of subjects distributed on non-linear manifolds, existing methods are hard to identify effective correlations. Moreover, most of the association studies are unsupervised bi-multivariate methods, and its identified SNPs or ROIs could be disease-irrelevant. To solve the problems above, a novel diagnosis-guided deep subspace clustering association (DDSCA) model is proposed for identifying the risk SNPs and brain network features. The main innovations of the work are given below.

1) DDSCA aims to explore the AD-related connectome and SNPs by association analysis between genotypes and network phenotypes from coarse-grained and fine-grained levels. Thus, based on AAL and human BN atlases, we constructed brain functional networks for each subject and further integrated genetic data with brain networks.
2) Deep subspace clustering is proposed to reconstruct the original data. Our neural network captures the non-linear structure of data and subspace clustering learns pairwise affinities between samples. For utilizing the diagnosis information to select the disease-relevant features, a label constraint term is presented to instruct the imaging genetic association learning.
3) Numerous experiments prove the superiority of DDSCA on distinct kinds of phenotype data, including brain network, node features, and edge features. DDSCA also finds some disease-relevant SNPs and brain connectome that involves the specific brain structure and fine-grained subregions.

TABLE I
CHARACTERISTICS OF ALL SUBJECTS

Variables	NC	SMC	EMCI	LMCI	AD
Subject Number	37	14	41	31	26
Gender (male/female)	16/21	5/9	17/24	20/11	11/15
Age (mean±std)	75.96±7.04	73.46±5.52	71.57±5.84	71.76±7.62	72.65±7.45

The remainder of this paper is organized as follows. The data description and our proposed method are shown in Section II. Experiments and biological analysis are given in Section III. Discussion and limitations are revealed in Section IV. At last, conclusions are given in Section V.

II. MATERIALS AND METHODS

A. Brain Imaging Genetics Data

The fMRI neuroimaging and genetic data utilized in the paper are from the ADNI database. In 2003, Principal Investigator Michael W. Weiner, MD developed the ADNI. As a longitudinal multicenter study, it aims to obtain clinical, imaging, genetic, and biochemical biomarkers for the early detection and tracking of AD. In this work, by aligning SNP and fMRI data, and then removing samples with missing values, we can obtain 149 valid samples, including 26 AD, 37 normal control (NC), 14 significant memory concern (SMC), 41 early mild cognitive impairment (EMCI), and 31 late MCI (LMCI). Using the regression weights derived from the NC participants, fMRI and SNP data were pre-adjusted for removing the effects of the baseline age and gender factors. Detailed contents of data have been listed in Table I. The multi-modal data in this study are obtained from the ADNI (https://adni.loni.usc.edu/). Our work is available at https://github.com/JiaoCuina/DDSCA.

Moreover, to reduce the influence of noise caused by data acquisition and physiology, all fMRI data must be preprocessed. Standard preprocessing steps for fMRI data are implemented in DPARSF v5.4 toolbox (http://rfmri.org/DPARSF). The specific processes included transforming formats, removing the first 10 volumes, slice timing, adjusting head motion, normalizing, smoothing with a 4mm full width at half maximum (FWHM) Gaussian kernel, removing linear trends, filtering with cut-off frequency of [0.01 0.1] Hz, and eliminating the covariates. Finally, the AAL atlas and BN atlas (http://www.nitrc.org) are used to segment the functional image into 90 ROIs and 246 ROIs, respectively. Each subject contains the time signal sequences of all ROIs. For the genetic data, the quality control (QC) for SNPs is performed by using the PLINK v1.9 software. Through the deletion rate QC, Hardy-Weinberg test with $P < 1 \times 10^{-6}$, and minor allele frequency less than 0.05, 85 SNPs extracted from risk genes [15] are applied in the research.

B. Brain Networks Construction

1) Brain Connectivity Features: As mentioned earlier, alteration in brain functional connectivity between brain areas is expected to provide potential biomarkers for classifying or predicting brain diseases. Suppose that each subject consists of k ROIs from fMRI data, then the time series of every ROI for every subject is calculated by averaging the fMRI time series across all voxels in each of k ROIs. Then, the Pearson correlation coefficient (PCC) is utilized to calculate the functional connectivity between the ROI pairs.

In this functional brain network, each ROI can be seen as a node, and the correlation coefficient denotes the edge weight. The functional connectivity matrix is normalized to the Z score through the Fisher's r-to-z transformation, constructing a $k \times k$ symmetric FC matrix for every subject. Furthermore, to obtain significant network features, all negative correlations are neglected. Lastly, after removing k diagonal elements, the upper triangle components of the FC matrix are extracted as connectivity features for the brain network. That is to say, the functional connectivity features of each subject are $(k \times (k-1))/2$, which contain 4005 and 30135 features for AAL and human BN atlases, respectively.

2) Brain Region Features: The study [16] has indicated that the topological property and the connection property can comprehensively measure the significance of a brain region. Here, the notion of the clustering coefficient is applied to explore the local topological property of a node, and the average weighted node degree is used to reflect its connection property.

Thus, after constructing the mentioned FC matrix for each subject based on AAL and BN atlases, we can obtain an adjacency matrix and a weighted matrix, which reflect the degree and weight of each ROI. Then, the clustering coefficient is calculated as:

$$ac_i = \frac{2e_i}{d_i \times (d_i - 1)}, \tag{1}$$

where e_i denotes the number of triangles that are composed of a node i with any of its two neighbors and d_i denotes the degree of the node i. Compared with the node degree, the clustering coefficient can reflect the local clustering property with its neighborhood, but cannot obtain the average connectivity strength to its neighborhood. Here, to further calculate the average connectivity strength of the node to its neighborhood, the average weighted degree of the node is utilized in this study, which is given as:

$$wd_i = \frac{\sum_{j \in N(i)} \delta_{ij}}{d_i}, \tag{2}$$

where $N(i)$ denotes the collection of nodes that link to the node i. In the above formula, if node i and j are directly connected, δ_{ij} is the functional connectivity value between them; otherwise, it takes 0.

Hence, not only to make use of the information of its neighborhood but also to consider the node clustering property, we

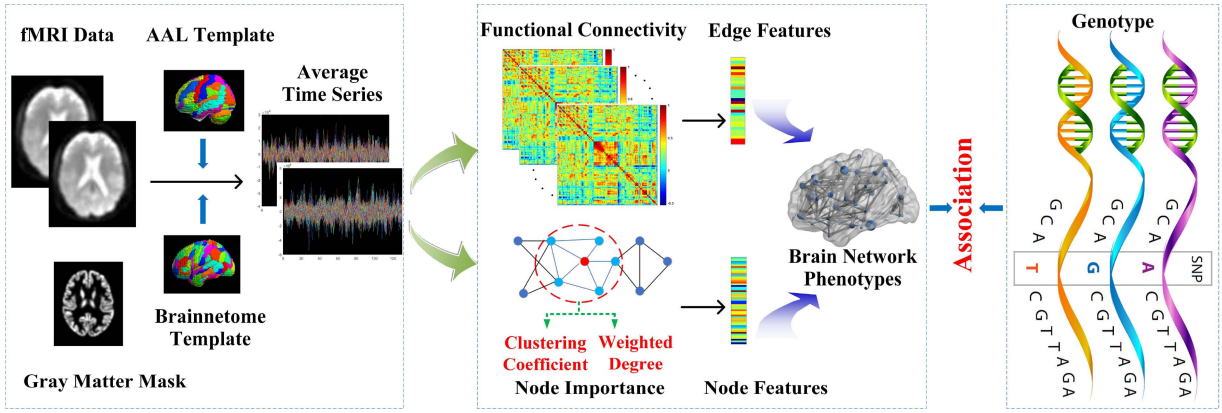

Fig. 1. Visual framework of the proposed method.

can define a novel node evaluation metric:

$$N_i = \text{Mean}\left(ac_i, wd_i\right). \tag{3}$$

At last, the node significance vectors are calculated as brain region features in brain networks. Based on AAL atlas, the human brain is segmented into 90 ROIs, so the dimension of node features is 90 for each sample. Meanwhile, the brain is divided into 246 ROIs via BN atlas, so the dimension of node features is 246 for each sample.

Moreover, Fig. 1 reveals the flowchart of our proposed framework. In detail, edge features and node features of the brain network are constructed based on the fMRI data that are preprocessed by the BN atlas and AAL atlas, respectively. Finally, a novel DDSCA model is proposed to discover SNP-QT associations between genetic factors and brain network features.

C. Sparse Canonical Correlation Analysis

SCCA explores the bi-multivariate associations between SNPs and brain imaging QTs. Given n subjects with p SNPs and q brain imaging QTs from m diagnostic groups, $\mathbf{X} = [x_1, \ldots, x_n]^T \in \mathbb{R}^{n\times p}$ represents the related genetic data and $\mathbf{Y} = [y_1, \ldots, y_n]^T \in \mathbb{R}^{n\times q}$ represents the related brain network data. Then, $u \in \mathbb{R}^{p\times 1}$ and $v \in \mathbb{R}^{q\times 1}$ are canonical weights corresponding to $\mathbf{X}$ and $\mathbf{Y}$, respectively. From this, SCCA model is defined as:

$$\begin{aligned} &\max_{u,v} u^T\mathbf{X}^T\mathbf{Y}v, \\ &\text{s.t. } u^T\mathbf{X}^T\mathbf{X}u \le 1, v^T\mathbf{Y}^T\mathbf{Y}v \le 1, \Omega(u) \le b_1, \Omega(v) \le b_2, \end{aligned} \tag{4}$$

where $\Omega(u)$ and $\Omega(v)$ are sparse penalties to control the sparsity and select significant SNPs and ROIs. Moreover, the first two terms are utilized to enhance the covariance structure of data.

D. Deep Subspace Clustering Learning

When identifying associations between SNPs and brain networks, there exist some challenges in handling the high-dimensional original data distributed on non-linear manifolds. In this work, we employ the multilayer feed-forward neural network to project the original data into the non-linear spaces, comprehensively preserving the non-linear structures of data. Further, the output at the top layer of the deep neural network is clustered by the subspace clustering algorithm to learn the pairwise affinities between samples. Through the deep subspace clustering network, we can cluster samples with the non-linear distribution and further reconstruct the original imaging genetics data.

In detail, to introduce this module concisely, the input layer of our feed-forward neural network is denoted as $t_l^0 = x \in \mathbb{R}^p$ (or $t_l^0 = y \in \mathbb{R}^q$) and $l = 1, 2, \ldots, n$. Then, the output of the k-th layers is given as:

$$t_l^{(k)} = G\left(\mathbf{W}^{(k)} t_l^{(k-1)} + c^{(k)}\right) \in \mathbb{R}^{d_k}, \tag{5}$$

where $k = 1, 2, \ldots, K$ is the layer number. $\mathbf{W}^{(k)} \in \mathbb{R}^{d_k \times d_{k-1}}$ and $c^{(k)} \in \mathbb{R}^{d_k}$ represent the weight and bias matrix of the k-th layer. The d_k is the output dimension of the k-th layer and $G(\cdot)$ is a sigmoid activation function. For the input data $\mathbf{X}$ (or $\mathbf{Y}$) with n samples, the output $\mathbf{T}^{(K)}$ at the top layer is defined below:

$$\mathbf{T}^{(K)} = \left[t_1^{(K)}, t_2^{(K)}, \ldots, t_n^{(K)}\right]. \tag{6}$$

Hence, the input data are mapped into nonlinear spaces to get $\mathbf{T}^{(K)}$ by utilizing the deep feed-forward neural network.

Further, to explore the complex multi-subspace structures of the original data, our method utilizes the self-expressive property to reconstruct data by conducting subspace clustering at the top layer iteratively. Subspace clustering can preserve the local data structure and learn the subject-subject similarity relations [17]. Specifically, self-expressive considers that the output $\mathbf{T}^{(K)}$ is a collection of n subjects from multiple linear subspaces. Thus, $t_l^{(K)}$ in specific subspace can be expressed by other subjects within the same subspace.

$$\min_{s_l, \left\{W^{(k)}, c^{(k)}\right\}_{k=1}^K} \sum_{l=1}^{n} \frac{1}{2} \left\| t_l^{(K)} - \mathbf{T}^{(K)} s_l \right\|_F^2 + \alpha \|s_l\|_1, s_{ll} = 0, \tag{7}$$

where α is a regularized parameter to improve sparsity. Once get the optimal solution of (7), a $n \times n$ self-expressed matrix $\mathbf{A} = |\mathbf{S}| + |\mathbf{S}|^T$ is obtained to reflect the intrinsic geometric

Algorithm 1: DDSCA Algorithm.

Input: SNP data $\mathbf{X} = [x_1, \ldots, x_n] \in \mathbb{R}^{n \times p}$, $z \in \mathbb{R}^{n \times 1}$;
Brain network data $\mathbf{Y} = [y_1, \ldots, y_n] \in \mathbb{R}^{n \times q}$;
Subjects with labels (NC, SMC, EMCI, LMCI, and AD).
Output: Correlation Weights u, v
Optimization:
Initialize $\mathbf{W}^{(k)}, c^{(k)}, \mathbf{T}^{(K)}and\mathbf{S}$;
while not converging do
for $i = 1, 2, \ldots, n$ **do**
Choose a subject x_i randomly and set $t_i^0 = x_i$;
Update $\mathbf{W}^{(k)}$ and $c^{(k)}$ by (10);
Compute $\mathbf{T}^{(K)}$ by (5);
Update $\mathbf{S}$ by (13);
end
end while
Initialize u and v;
while not converging **do**
Compute diagonal matrix $\mathbf{D}_1$ and update u by (16);
Scale u so that $\|\Lambda u\|_2^2 = 1$;
Compute diagonal matrix $\mathbf{D}_2$ and update v by (18);
Scale v so that $\|\Omega v\|_2^2 = 1$;
end while

properties of data. Ideally, $\mathbf{A}_{ij} \neq 0$ only if the corresponding data points t_i and t_j are drawn from the same subspace. In (7), the first term is applied to reconstruct the input data and ℓ_1-norm is utilized to prevent overfitting and ensure sparsity. The constraint $s_{ll} = 0$ is enforced to avoid trivial solutions.

At last, to improve the interpretability of the identified biomarkers, the projection neural activities of the SNP matrix $\mathbf{X}$ and brain network matrix $\mathbf{Y}$ can be expressed by $f(\mathbf{X}) = \mathbf{A}_x\mathbf{X}$ and $g(\mathbf{Y}) = \mathbf{A}_y\mathbf{Y}$, respectively.

E. Diagnosis-Guided Deep Subspace Clustering Association Model

Through deep subspace clustering, we project each subject to a nonlinear space, and then cluster the subspaces at the top $K+1$ layer by the self-expressive property. Thus, we can obtain a self-expressed matrix that ensures sparsity by the subspace clustering and achieves nonlinearity by the neural network, which further reconstructs the imaging genetics data.

Moreover, to further effectively find the brain connectome and SNPs by association analysis between genetic data and brain network, SCCA is usually utilized to learn canonical weight vectors for SNPs and network features [4]. However, SCCA is unsupervised and overlooks the diagnosis information, resulting in the inability to select some disease-related SNP-QT associations. To overcome the above problem, a label constraint with diagnosis information is proposed to instruct the imaging genetics association analysis, ensuring a supervised framework. The objective function is defined as:

$$\min_{u,v} L_{SCCA}(u, v) + L_R(v),$$

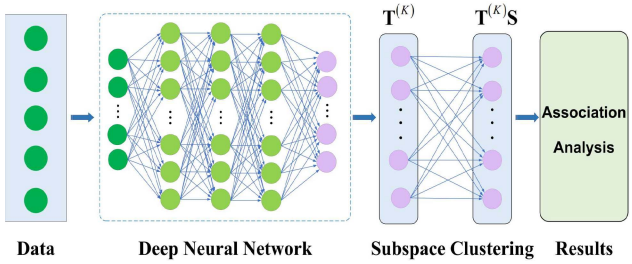

Fig. 2. Embedding deep subspace clustering network before the association analysis.

$$\text{s.t. } u^T\mathbf{X}^T\mathbf{X}u \leq 1, v^T\mathbf{Y}^T\mathbf{Y}v \leq 1, \Omega(u) \leq b_1, \Omega(v) \leq b_2. \tag{8}$$

In this model, L_R contributes to selecting the discriminate imaging QTs by utilizing the diagnosis information, and L_{SCCA} explores the bi-multivariate correlations between SNPs and brain network QTs. Thus, the final objective function of DDSCA is given as:

$$\begin{aligned} &\min_{u,v,\theta_f,\theta_g} \|z - g(\mathbf{Y})v\|_2^2 - u^T f^T(\mathbf{X}) g(\mathbf{Y}) v, \\ &\text{s.t. } u^T f^T(\mathbf{X}) f(\mathbf{X}) u \leq 1, \|u\|_1 \leq b_1, \\ &\quad v^T g^T(\mathbf{Y}) g(\mathbf{Y}) v \leq 1, \|v\|_1 \leq b_2. \end{aligned} \tag{9}$$

In detail, let $z \in R^{n \times 1}$ be a column vector related to the label elements of n samples. The canonical weights u and v represent the importance of every feature, such as SNPs or ROIs. The above elements will boost each other, ensuring a more appropriate feature selection capability compared to the existing association studies.

To sum up, DDSCA uses the regression term to achieve the label constraint on brain imaging data, while applying the SCCA to both neuroimaging and genetic data. Before this, DDSCA mainly utilizes the deep subspace clustering network to preprocess the real imaging genetics data, which can reconstruct the original data by the non-linear subspace clustering. Moreover, Fig. 2 visualizes the deep subspace clustering process of the original data before the association analysis.

F. Optimization Algorithm

To address the objective function in (9), a proximal alternating optimization algorithm is used to get the optimal solution. The specific iterative update formulas of $\mathbf{W}^{(k)}$, $c^{(k)}$, $\mathbf{S}$, u, and v are given here:

Update $\mathbf{W}^{(k)}$ and $c^{(k)}$: Let $\mathbf{S}$ and $\mathbf{T}^{(K)}$ be fixed, then the update formulas of corresponding $\mathbf{W}^{(k)}$ and $c^{(k)}$ are learned by the sub-gradient descent algorithm:

$$\begin{aligned} &J\left(\left\{\mathbf{W}^{(k)}, c^{(k)}\right\}_{k=1}^{K}\right) \\ &= \min_{\{\mathbf{W}^{(k)}, c^{(k)}\}_{k=1}^{K}} \frac{1}{2}\sum_{l=1}^{n}\left\|t_l^{(K)} - \mathbf{T}^{(K)}s_l\right\|_F^2. \end{aligned} \tag{10}$$

This problem can be solved by the sub-gradient descent algorithm. Take the derivatives of parameters $\mathbf{W}^{(k)}$ and $c^{(k)}$ in (10), then set them to zero. For saving space, here we can update this neural network as follows:

$$\begin{cases} \mathbf{W}^{(k)} = \mathbf{W}^{(k)} - \varphi \frac{\partial J}{\partial \mathbf{W}^{(k)}} \\ c^{(k)} = c^{(k)} - \varphi \frac{\partial J}{\partial c^{(k)}} \end{cases}, \tag{11}$$

where φ denotes a small positive constant, representing the step size, which is set as $\varphi = 10^{-4}$ in this study.

Update $\mathbf{S}$*:* When $\mathbf{W}^{(k)}$ and $c^{(k)}$are fixed, we can get the update formula of $\mathbf{S}$ via:

$$\min_{\mathbf{S}} \frac{1}{2} \left\| \mathbf{T}^{(K)} - \mathbf{T}^{(K)} \mathbf{S} \right\|_F^2 + \alpha \|\mathbf{S}\|_1. \tag{12}$$

Then, take the derivative of (12) related with $\mathbf{S}$ and set it to zero, so the update formula is given as:

$$\left(\mathbf{T}^{(K)}\right)^T \left(\mathbf{T}^{(K)}\right) \mathbf{S} = \left(\mathbf{T}^{(K)}\right)^T \left(\mathbf{T}^{(K)}\right) - \alpha. \tag{13}$$

The $\mathbf{W}^{(k)}$, $c^{(k)}$, and $\mathbf{S}$ are iteratively updated until the error function in (10) meets the convergence situation.

Update u *and* v*:* For updating u or v, the network parameters $\mathbf{W}^{(k)}$, $c^{(k)}$, and $\mathbf{S}$ stay in an optimum state. Next, the objective function of the proposed DDSCA in (9) is transformed into the following form by using the Lagrange multiplier and then dropping the constant. Finally, the transformed formula can be given as follows:

$$\begin{aligned} &\min_{u,v} \|f(\mathbf{X})\, u - g(\mathbf{Y})\, v\|_2^2 + \|z - g(\mathbf{Y})\, v\|_2^2 \\ &\quad + \lambda_1 \|f(\mathbf{X})\, u\|_2^2 + \lambda_2 \|g(\mathbf{Y})\, v\|_2^2 + \gamma_1 \|u\| + + \gamma_2 \|v\|, \end{aligned} \tag{14}$$

where z is a column vector, denoting the diagnosis status of n samples. λ_1, λ_2, γ_1, and γ_2 are parameters to ensure the global sparsity. We have proven that the results are insensitive to λ_1 and λ_2 settings through the experimental evaluation. Also, according to some related studies [13], [18], these works have studied a similar problem using a different approach. Here, they are set to the fixed values $\lambda_1 = 1$ and $\lambda_2 = 1$ for simplicity.

Finally, take the derivative of Lagrange function in (14) for u by fixing v, and let it be zero, then we can obtain:

$$(1 + \lambda_1) f(\mathbf{X})^T f(\mathbf{X})\, u - f(\mathbf{X})^T g(\mathbf{Y})\, v + \gamma_1 \mathbf{D}_1 u = 0. \tag{15}$$

The related updating formula for u is obtained as follows:

$$u = \left(\gamma_1 \mathbf{D}_1 + (1 + \lambda_1) f(\mathbf{X})^T f(\mathbf{X})\right)^{-1} f(\mathbf{X})^T g(\mathbf{Y})\, v. \tag{16}$$

Similarly, we can obtain the derivative of v by fixing u, and set it to zero, then we can get:

$$(2 + \lambda_2)\, g(\mathbf{Y})^T g(\mathbf{Y})\, v - g(\mathbf{Y})^T (f(\mathbf{X})\, u + z) + \gamma_2 \mathbf{D}_2 v = 0. \tag{17}$$

At last, the updating formula for v is given as follows:

$$v = \left(\gamma_2 \mathbf{D}_2 + (2 + \lambda_2)\, g(\mathbf{Y})^T g(\mathbf{Y})\right)^{-1} g(\mathbf{Y})^T (f(\mathbf{X})\, u + z). \tag{18}$$

In (18), $\mathbf{D}_1$ and $\mathbf{D}_2$ are diagonal matrices, in which the r_1-th element is $1/2\|u^{r_1}\|_1 (r_1 \in [1, p])$ and the r_2-th element is $1/2\|v^{r_2}\|_1 (r_2 \in [1, q])$, which cannot be calculated when $|u| = 0$ and $|v| = 0$. Hence, $1/2\|u^{r_1}\|_1$ and $1/2\|v^{r_2}\|_1$ are rewritten as $1/2\sqrt{u^{r_1^2} + \varsigma}$ and $1/2\sqrt{v^{r_2^2} + \varsigma}$, where ς is a real number with a small value.

Obviously, $\mathbf{D}_1$ and $\mathbf{D}_2$ are dependent on u and v, which are unknown. So, in the iterative algorithm, first, the initial values of u and v are given, and then the diagonal matrices of them are calculated. This process is stopped until the predefined stopping criterion is satisfied. Algorithm 1 gives the overall optimization process of the proposed DDSCA method.

G. Convergence Analysis of DDSCA

DDSCA is biconvex and its local optimum can be attained finally by running Algorithm 1. DDSCA mainly consists of two parts: deep subspace clustering and the final association model. To clearly state the stop conditions of DDSCA, we analyze the convergence of the deep subspace clustering and the association model, respectively. We record the error values of the above loss functions in each iteration. The initial iteration is set to 400. Meanwhile, the stopping criterion is to set a threshold 10^{-3} for deep subspace clustering and 10^{-5} for the association model. If the relative changes are lower than the threshold, the optimization process will terminate. Convergence performance of two parts in DDSCA is visually shown in Fig. 3. As revealed in Fig. 3(a), for deep subspace clustering, we discovered that error values decreased within 200 initial iterations, so we empirically set the number of iterations as 200 in our experiment. For the association model in Fig. 3(b), the error values decreased rapidly within 15 initial iterations. This indicates that the fast convergence is included in the optimization algorithm of the association model. Accordingly, we empirically set the number of iterations as 50 for the association analysis experiment.

III. Experimental Results

A. Experiment Settings

In our imaging genetics study, for assessing the association ability, five-fold cross validation (5-CV) is conducted in the experiment. Meanwhile, the grid search method is applied to choose the optimum parameters (γ_1 and γ_2) automatically in a range of $\{10^{-3}, 10^{-2}, 10^{-1}, 1, 10^1, 10^2, 10^3\}$ by a nested 5-CV on the training set. Furthermore, in the comparison process, the correlation coefficient (CC) [19] is utilized to verify the association performance between SNPs and brain network features. The higher the CC is, the better the method's performance is. During the experiment, for the diagnosis labels control, we employed the random over sampling strategy to address the sample imbalance issue. This strategy performs over sampling on imbalanced data by replicating samples from minority classes and combining them with the original data, which mitigates the potential bias introduced by uneven class distribution.

To identify the disease-related brain connectome from the specific brain structure and the fine-grained level, based on the AAL atlas and the human BN atlas, we preprocessed fMRI imaging

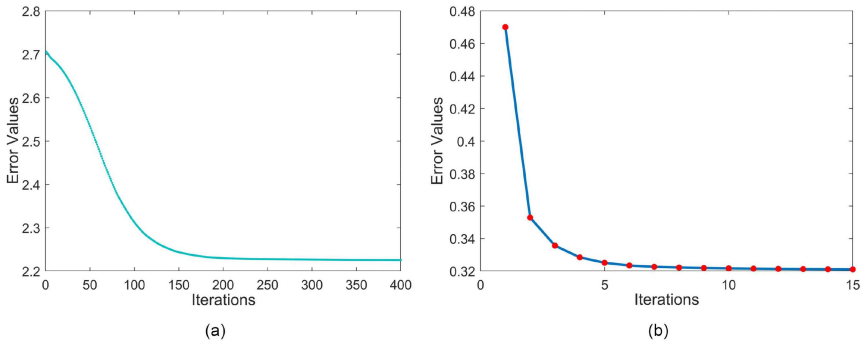

Fig. 3. Convergence analysis of the DDSCA algorithm. (a) Convergence of deep subspace clustering; (b) convergence of the association model.

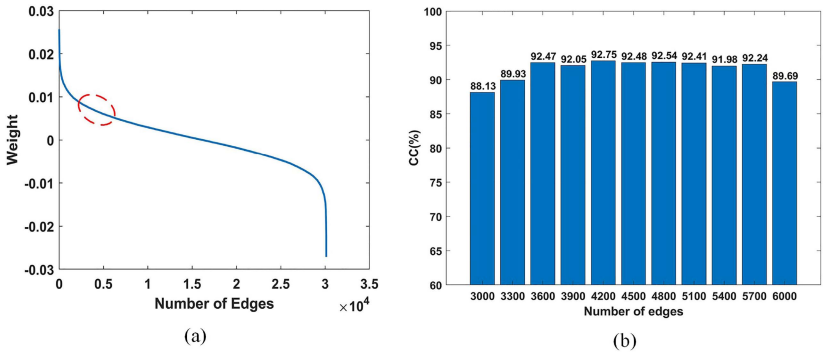

Fig. 4. (a) The weight distribution of all edges on BN atlas; (b) the optimal number of edges is selected by the ReliefF algorithm.

data and constructed brain networks. In addition, for discovering associations between SNPs and brain network features, edge features are collected from the constructed FC matrix. Three different neuroimaging node features are also applied in this experiment, including (1) VBM (voxel-based morphometry): non-network node features from the structural imaging data in ADNI; (2) NIP: network node significance calculated by the page-rank algorithm; (3) NCD: network node significance learned from the clustering coefficient and the average weighted node degree. Network is a brain network combined with edge features and NCD node features.

For node features, VBM is obtained from structural MRI imaging data, so 90 and 246 ROIs level measurements of mean gray matter densities are extracted as node features based on AAL and BN atlases. Also, NIP node features are calculated using a popular page-rank algorithm in the study [2], which is defined as $s_i = \sum_{j \in N(i)} s_j / p_j$, where i is a node and N_i are nodes connecting with node i. s_j denotes the score of a node j and p_j is the number of links of node j. The bigger the value of s_j, the higher the significance of the node.

However, the number of edges of the brain network processed by the human BN atlas is 30135, which is time-consuming to run on the developed DDSCA model. Thus, it is crucial to utilize feature selection to reduce redundancy. Here, the ReliefF algorithm [20] is used to select informative features, which is performed as an independent data preprocessing step before inputting the reduced-dimensional data into our framework. Specifically, each feature of raw data is allocated a weight statistically according to its relevance to the diagnostic information. From Fig. 4(a), we can obtain weights of all edges in descending order, and then select top δ weighted features, which can remove redundant features. To get the optimal value of δ, the area near the inflection point in Fig. 4(a) is taken as the parameter selection area of the δ, which is selected from 3000 to 6000 with a 300 interval. The parameter selection process is shown in Fig. 4(b). It is easy to observe that a higher CC result is obtained when the δ value is 4200.

Moreover, the traditional SCCA (CCA with lasso) [21] is utilized as the benchmark algorithm. At the same time, DSCCA [6], DS-SCCA [22], sCCAR [23], KSCCA [7], and DCCA [8] are comparison approaches in the experiment. The hyper-parameters of all methods are adjusted according to the related papers and the best results are obtained. For comparison methods based on deep neural networks, the related hyper-parameters are shown here. The maximum epoch is set as 300, the learning rate is 0.01, and the momentum is 0.99. Three hidden layers are used here and the number of nodes of each layer is set as 512 according to the related literature [8], [22].

B. Associations Between SNPs and Brain Network Phenotype

The association analysis experiment of brain imaging genetics is mainly to compare CC results obtained from different approaches. To prove the robustness of DDSCA, Table II lists the average testing CC results from seven methods on five QTs properties of AAL atlas and human BN atlas, respectively. The **bold** values denote best results. From these results, we can

TABLE II
RESULTS WITH 5-CV ON DIFFERENT QTS PROPERTIES OF AAL AND BN ATLASES

Atlas	Method	CC (mean%±std)				
		VBM	NIP	NCD	Edge	Network
AAL	SCCA	15.40±0.027	8.03±0.081	16.06±0.049	10.55±0.011	9.54±0.083
	sCCAR	20.10±0.014	12.03±0.013	18.21±0.023	8.44±0.087	8.53±0.088
	DSCCA	20.82±0.034	12.46±0.023	17.00±0.035	17.93±0.074	14.88±0.047
	DS-SCCA	61.37±0.017	66.33±0.008	68.03±0.017	72.26±0.011	77.33±0.006
	KSCCA	30.30±0.061	32.02±0.018	37.17±0.031	40.69±0.025	43.36±0.041
	DCCA	59.33±0.018	62.32±0.010	63.85±0.006	66.55±0.015	74.56±0.002
	DDSCA	**67.02±0.011**	**73.90±0.009**	**78.28±0.011**	**81.26±0.007**	**87.02±0.002**
BN	SCCA	13.29±0.024	11.92±0.041	14.44±0.037	14.18±0.052	12.45±0.056
	sCCAR	27.75±0.069	10.64±0.041	15.42±0.031	20.92±0.033	23.61±0.037
	DSCCA	18.76±0.011	12.70±0.028	19.61±0.042	15.14±0.018	14.95±0.031
	DS-SCCA	56.77±0.027	58.33±0.015	68.76±0.015	75.05±0.003	79.62±0.002
	KSCCA	28.63±0.047	33.68±0.022	36.45±0.024	39.91±0.004	35.73±0.017
	DCCA	52.93±0.022	60.28±0.017	59.31±0.016	71.64±0.003	78.06±0.003
	DDSCA	**68.44±0.007**	**70.51±0.008**	**75.30±0.008**	**83.53±0.003**	**87.88±0.002**

observe that all mentioned methods have stable performance across five different QTs properties. Though the above methods can get acceptable results, we find that sCCAR performs generally better than SCCA, indicating the effective of the label constraint. Moreover, it is observed that DS-SCCA, KSCCA, DCCA, and DDSCA outperform other SCCA-based approaches significantly. These results prove that the deep extension methods can identify non-linear relations between features and capture the potential correlations between SNPs and brain network phenotypes. More importantly, for different QTs properties, our proposed DDSCA method obtains the best CC testing results in both AAL and BN atlases. The reasons are as follows. First, the multilayer neural network captures the nonlinear manifold of data and the subspace clustering learns the pairwise affinities between samples. Second, the regression term with label constraint achieves supervised learning, which can discover disease-relevant associations.

Among the five QT properties, the better performance is obtained on the Network property that considers both node features and connectivity information. Also, node features (NIP and NCD) and edge features of the brain network are also well correlated with SNPs, revealing the relations between SNPs and brain connectome. Noting that NCD node property performs better than NIP and VBM, which can identify more meaningful biomarkers. Thus, calculating the node significance by the clustering coefficient and the average weighted node degree is practical for mining the information of brain areas.

To save space and give a clear description, the heatmaps of canonical loadings at the AAL atlas obtained from DDSCA on five QTs properties are shown in Fig. 5. Each column represents the different brain QTs properties. The estimated weights of u on the left represent the SNP weights. Also, the estimated v denoting QTs weights is shown on the right. From Fig. 5, we find that our proposed DDSCA can identify consistent and clear patterns (rs429358-C) across all QTs properties. Studies have found that the *APOE* SNP rs429358 is the most significant genetic risk factor for late onset AD [24]. DDSCA also identifies some consistent SNPs with a much clear pattern on the Network and NCD properties, including SNPs rs429358 (*APOE*), rs1081105 (*APOE*), and rs283814 (*NECTIN2*), with further research being warranted for rs188535946. At the same time, some ROI markers are also detected by the DDSCA. Specifically, a number of consistent ROIs related to AD are found on both NCD and Network QTs, including *Thalamus* [25], *Temporal* [26], *Cingulum* [27], and *ParaHippocampal* [28]. The meaningful ROIs, such as *Caudate* [25], *Amygdala* [29], and *Hippocampus* [30], are also identified in previous studies. In short, DDSCA can successfully explore some valuable SNPs and ROIs closely related to AD by correlation analysis between SNPs and brain network.

TABLE III
TOP 10 ROIS FOR BRAIN NETWORK QTS ON AAL ATLAS (L.=LEFT; R.=RIGHT)

ID	ROIs	Abbreviation	Studies
40	R. ParaHippocampal gyrus	R.PHG	[28]
37	L. Hippocampus	L.HIP	[30]
78	R. Thalamus	R.THA	[25]
89	L. Inferior Temporal gyrus	L.ITG	[26]
36	R. Posterior Cingulate gyrus	R.PCG	[27]
5	L. Superior Frontal gyrus (orbital)	L.ORBsup	[33]
41	L. Amygdala	L.AMYG	[29]
88	R. Temporal pole: middle temporal	R.TPOmid	[26]
72	R. Caudate nucleus	R.CAU	[25]
52	R. Middle Occipital gyrus	R.MOG	[34]

C. The Most Related ROI Markers

To further detect the crucial brain ROIs based on AAL and BN atlases, we average the canonical loadings obtained on the Network QTs across five different folds to choose the significant ROIs. The top 10 maximum weight ROIs based on AAL and BN atlases are listed in Tables III and IV, respectively. Moreover, through the ROI drawing function in the BrainNet Viewer v1.7 toolbox [31], the corresponding brain regions extracted from AAL and BN atlases are visualized in Fig. 6(a) and (b), respectively.

From the above tables and figures, the significant brain regions and fine-grained subregions related to AD are extracted based on AAL atlas and BN atlas, respectively. They are not only consistent with the previous works [11], [32] but also supported

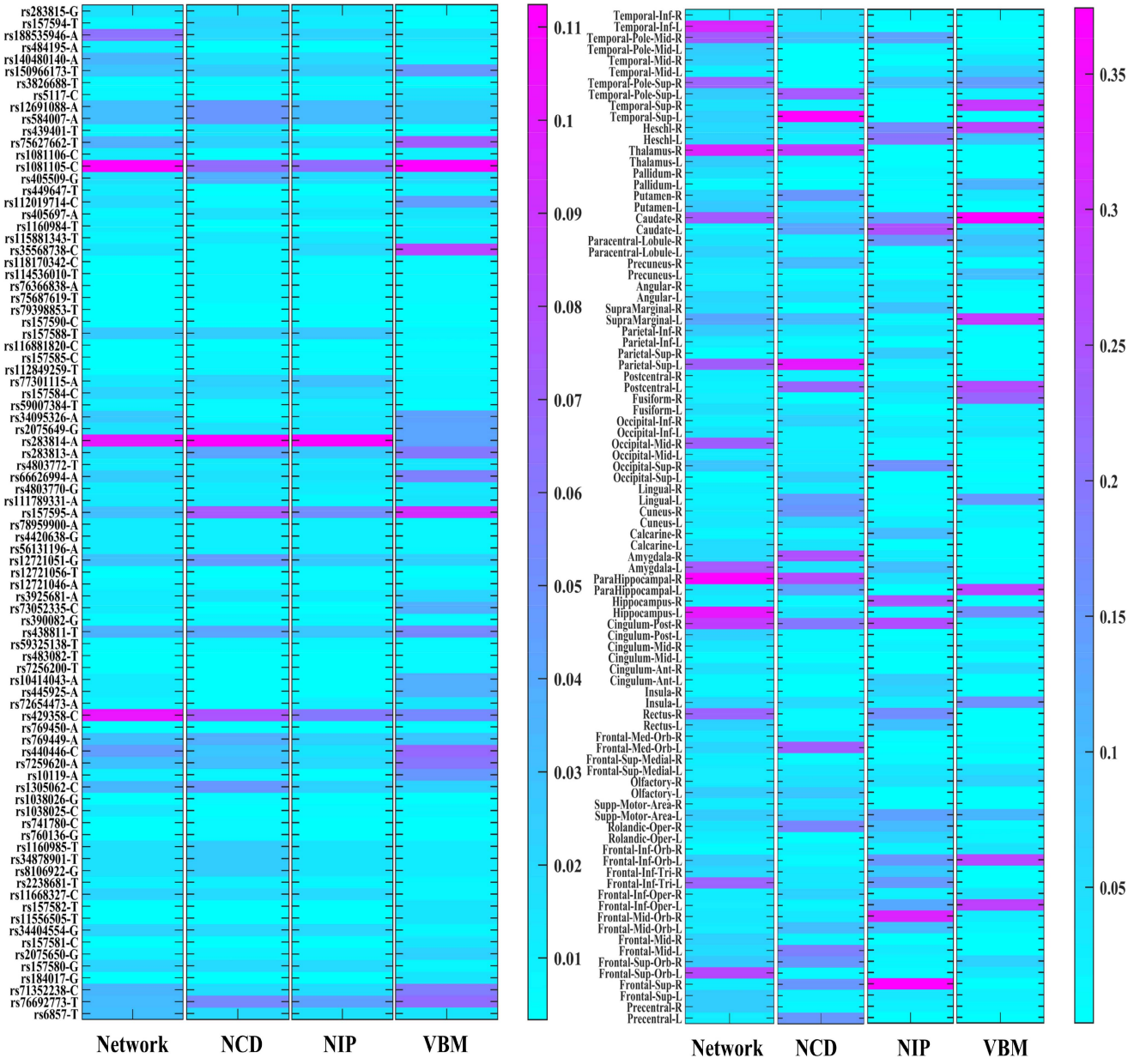

Fig. 5. Average weight maps of SNPs (left panel) and ROIs (right panel) with five-fold cross-validation on different QTs properties data.

by existing studies. For instance, the significant ROIs identified on the AAL atlas include *Hippocampus* and *Amygdala* [29], [30], which are closely associated with memory functions of AD. Meanwhile, the pathological abnormality will first appear in *ParaHippocampal* gyrus in patients with AD [28]. *Temporal* gyrus is related to visual, memory, and emotional control [26]. Further, the *Superior Frontal* gyrus has been correlated with mild cognitive impairment [33]. *Occipital* lobe is the center of the visual cortex. The damage to the *Occipital* gyrus [34] will result in visual impairment and memory disorder. More importantly, we also identified ROIs at a fine-grained level based on the BN atlas, such as the *Caudal Hippocampus*, *Rostral Hippocampus*, and *Lateral Amygdala* [30], [35]. Some clinical examples have also verified that significant AD individuals suffer from movement and perception problems. The atypical symptoms, such as stiff hands and feet, or curl and incontinence, are appeared in certain cases. For instance, literature [36] has proven that the *Middle Occipital* gyrus identified on AAL is related to perception and vision in AD. Based on the BN atlas, we have explored its subregions including *mOccG*, *IsOccG*, and *cLinG* [34]. Moreover, the *Thalamus* gyrus identified on AAL [25] may contribute to the cognitive and emotional control since patients with AD usually have greater emotional changes. Here, its subregions *Otha*, *mPFtha*, and *cTtha* have also been detected on the BN atlas. In addition to verifying previous findings, the proposed approach also identifies some potential risk ROIs including *Basal Ganglia* [37] and *Caudal Lingual* gyrus [27], which deserve further investigation.

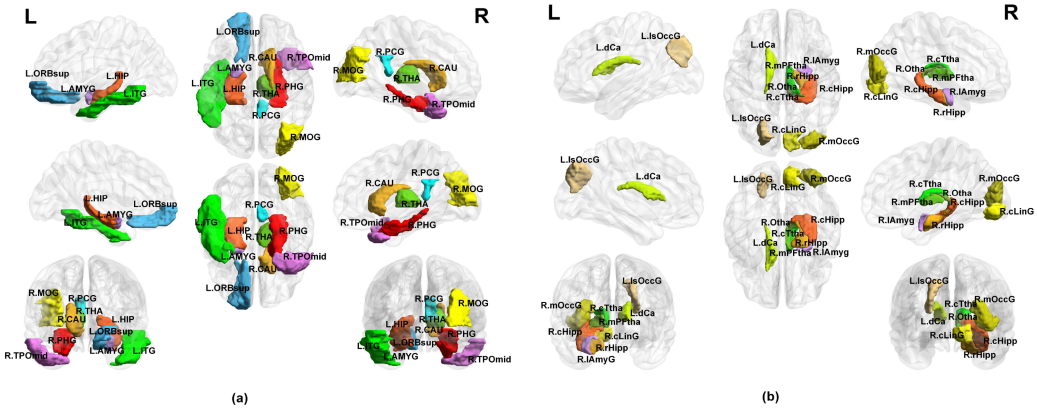

Fig. 6. (a) The extracted top 10 ROIs for brain network QTs property on AAL atlas; (b) the extracted top 10 ROIs for brain network QTs property on BN atlas. Different color denotes different brain regions.

TABLE IV
TOP 10 ROIS FOR BRAIN NETWORK QTS ON BN ATLAS (L.=LEFT; R.=RIGHT)

ID	Gyrus	ROIs	Studies
218	Hippocampus	R. cHipp	[35]
200	Lateral Occipital Cortex	R. mOccG	[34]
227	Basal Ganglia	L. dCa	[37]
214	Amygdala	R. lAmyg	[29]
242	Thalamus	R. Otha	[25]
216	Hippocampus	R. rHipp	[35]
209	Lateral Occipital Cortex	L. lsOccG	[34]
232	Thalamus	R. mPFtha	[25]
190	MedioVentral Occipital Cortex	R. cLinG	[35]
244	Thalamus	R. cTtha	[25]

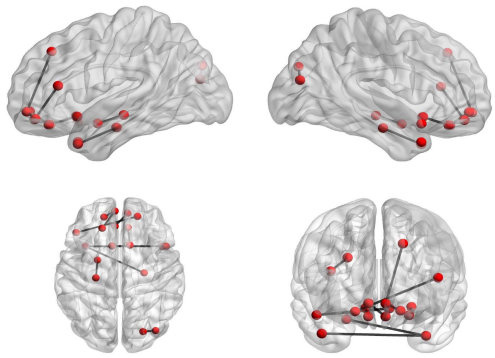

Fig. 7. Visualization of the edges selected by DDSCA based on the AAL atlas.

D. Brain Connectivity Analysis

Brain network connections of different subjects reflect the comprehensive characteristics of different brain systems. In a brain network, nodes represent the different brain areas. Edges represent the connections between brain areas. Connectivity weights denote the connection strength of edges after feature selection. To analyze the AD-related brain connectome from different aspects, according to the acquired average sparse coefficients by 5-CV, our DDSCA selects the top 10 edges with maximum weight in the brain network constructed based on AAL and BN atlases. They are also visualized using the draw edge function in the BrainNet Viewer v1.7 toolbox.

Edges detected on AAL atlas are shown in Table V and Fig. 7. We discover that brain region connections between *Temporal* gyrus and other brain areas (*ParaHippocampal* and *Olfactory* gyrus) [38] are closely associated with AD. The *Olfactory* decline in elderly patients is related to volume reductions in the *ParaHippocampal*. Moreover, the selected ROI such as *Superior Frontal Gyrus* [39] is related to the *Inferior Frontal Gyrus (Triangular)* and *Gyrus Rectus* in the brain, which reflects the depression state of AD patients. In [33], this study shows that the *Inferior Frontal Gyrus (Triangular)* area is most commonly related to mild cognitive impairment. The reduction of the *Middle Occipital* gray matter and *Superior Occipital* gray matter has been correlated with the ability of perception and vision of AD patients [34]. In addition to verifying the previous studies, DDSCA also uncovers the disease-relevant brain connectome like the memory control areas *Hippocampus* and *Amygdala*. *Olfactory* network includes *Gyrus Rectus*, *Lateral temporal lobe*, and so on [40]. These potential relations among genotypes, phenotypes, and neuropsychiatric symptoms can be researched further.

To enhance the visualization clarity, the overall correlation weights distribution of 24 brain gyri based on the human BN atlas is also shown in Fig. 8, in which the bigger dots denote higher connectivity weights. Specifically, Fig. 8 shows the important functional connectivity of different brain gyrus. The study [41] has shown that there exist structural associations between *Thalamus* (*Tha*) and *Caudate* gyri, as well as between *Hippocampus* (*Hipp*) and *Caudate* gyri in MCI $\varepsilon 4$ carriers. Additionally, FCs that reveal more contributing power mainly focused between *ParaHippocampal* (*PhG*) and *Thalamus* (*Tha*) gyrus [42], and damaging this pathway would influence the allocentric spatial memory. In addition to the default mode network (DMN) related regions, literature [37] also found that there exists abnormal functional connectivity between *BG* and *Amygdala*, and both of them are associated with cognitive impairment. In early AD patients, gray matter volume was lower in the *Frontal*, *Temporal*, *Occipital*, and *Parietal* lobes, which agrees with the identified significant ROIs.

TABLE V
EDGES SELECTED VIA DDSCA BASED ON AAL ATLAS (L.=LEFT; R.=RIGHT)

Number	Node 1	Node 2
1	R. Middle occipital gyrus	R. Superior occipital gyrus
2	L. Middle temporal gyrus	R. Middle temporal gyrus
3	L. Superior frontal gyrus (orbital)	L. Superior frontal gyrus (dorsolateral)
4	L. Inferior frontal gyrus (triangular)	R. Superior frontal gyrus (orbital)
5	L. Hippocampus	L. Amygdala
6	R. Superior temporal gyrus	L. Olfactory cortex
7	L. Superior frontal gyrus (medial orbital)	L. Superior frontal gyrus (orbital)
8	L. Middle temporal gyrus	R. Parahippocampal gyrus
9	R. Gyrus rectus	R. Olfactory cortex
10	L. Gyrus rectus	R. Superior frontal gyrus (medial orbital)

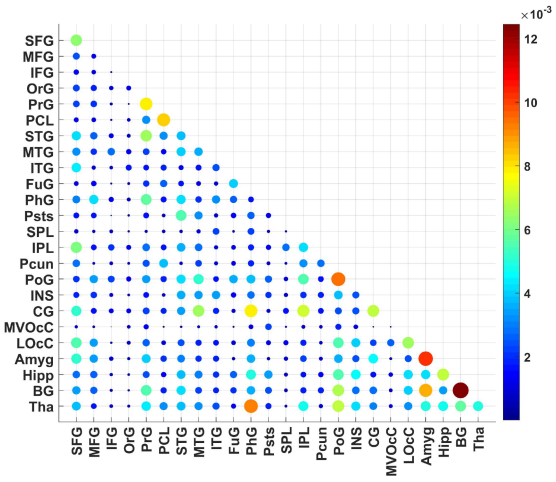

Fig. 8. Visualization of the distribution on mean weights of whole-brain edges based on human BN atlas.

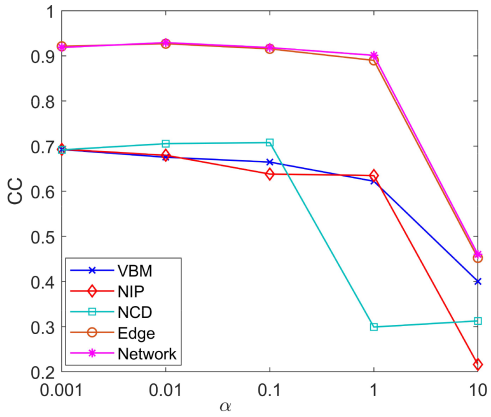

Fig. 9. Experimental results with different α on different QTs properties based on AAL atlas.

IV. DISCUSSION

A. Effect of Parameters

To ensure the sparsity of the self-expressed matrix, the parameter α is introduced in (7), which is selected in the range of $\{0.001, 0.01, 0.1, 1, 10\}$ in the experiment. Furthermore, we utilized the evaluation indicator CC to assess the performance of DDSCA. Due to page limitation, the effect of the parameter α for correlation analysis between SNPs and five QTs properties on AAL atlas is shown in Fig. 9.

As revealed in Fig. 9, we find that association performance is sensitive to the parameter α. In detail, DDSCA performs better when α in the range of $\{0.01, 0.1\}$, indicating that the self-expressed matrix in (7) can learn the similarity structure of data. Obviously, the CC values decrease directly when the α value is bigger than 0.1. That is, the influence of the constraint term in (7) tends to be weakened when α value is increasing, resulting in some non-zero elements in the reconstructive weight vector, which may ignore the local structure of data.

B. Ablation Experiment of DDSCA

DDSCA is composed of three crucial parts, containing the regression term with diagnosis labels, the deep neural network, and the subspace learning. Here, to verify the effectiveness of the three important factors in DDSCA, three unique variants are summarized below. *First*, to demonstrate the clustering performance at the top $K+1$ layer of this network by the self-expressive property, we removed the function of subspace clustering of DDSCA and employed the simple Mean Squared Error (MSE) loss as a substitute, represented by D1. *Second*, to prove the non-linear projection ability of the deep neural network, DDSCA without the function of the deep neural network is called D2. *Third*, if eliminate the regression term with the diagnosis status, we can obtain the degraded model D3. To save space, the corresponding association results between SNPs and the fMRI network processed by BN atlas are shown in Fig. 10. It is observed that our DDSCA is superior to the three degraded models (without one of the crucial parts), revealing the significance of the three crucial parts in DDSCA.

C. Limitations and Future Works

The above experiments prove the robustness and superiority of the presented DDSCA, but there also exist some drawbacks that would be further addressed. *To begin*, to further improve the generalization of our proposed method, we will utilize a wider variety of brain disease data to better explore new biomarkers related to other brain disorders. *Second*, DDSCA as a deep subspace clustering association study can identify

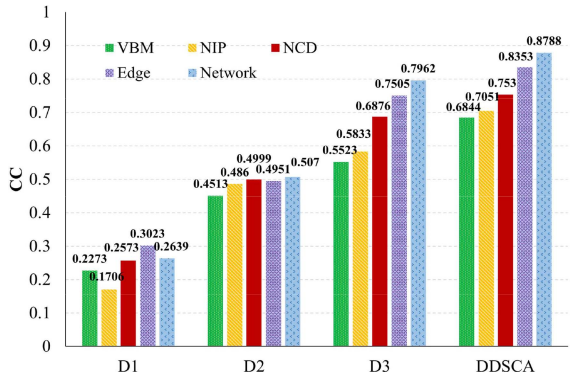

Fig. 10. Ablation experimental results with 5-CV on different QTs properties of the human BN atlas.

biomarkers at the coarse-grained and fine-grained levels. We will later employ more brain atlases to verify the robustness of DDSCA. *Third*, the optimization and training processes of our method are time-consuming, that is, the higher the number of feature dimensions, the higher the time complexity. Hence, future work will form an end-to-end deep learning method to improve performance. *Lastly*, the number of subjects utilized in this study is relatively small. In the future, we will collect and include more data samples and more modalities of imaging genetics data for AD diagnosis.

V. Conclusion

In the work, the edge-level and network-level brain connection properties are identified from the coarse-grained and fine-grained levels by performing a comprehensive brain imaging genetics study. Meanwhile, a novel DDSCA method is developed for identifying associations between genetic variation and brain network features, which is different from previous works that only focused on the brain regions. It should be noted that the traditional association model is unsupervised and the original data are non-linear with high-dimensional features and fewer subjects. Based on the above consideration, the label constraint is incorporated into the DDSCA method, which makes full use of the diagnosis information. In addition, the deep subspace clustering is well-suited for reconstructing the real data collected from the ADNI database. More importantly, this study first attempts to use the AAL atlas and the human BN atlas to parcellate the fMRI imaging data, which can identify brain connectome from the specific brain structure and the fine-grained level. The efficiency of the DDSCA approach is also proven by extensive experimental findings.

References

[1] M. P. Mattson, "Pathways towards and away from Alzheimer's disease," *Nature*, vol. 430, no. 7000, pp. 631–639, 2004.
[2] L. Shen and P. M. Thompson, "Brain imaging genomics: Integrated analysis and machine learning," *Proc. IEEE*, vol. 108, no. 1, pp. 125–162, Jan. 2020.
[3] L. Shen et al., "Whole genome association study of brain-wide imaging phenotypes for identifying quantitative trait loci in MCI and AD: A study of the ADNI cohort," *Neuroimage*, vol. 53, no. 3, pp. 1051–1063, 2010.
[4] D. M. Witten and R. J. Tibshirani, "Extensions of sparse canonical correlation analysis with applications to genomic data," *Stat. Appl. Genet. Mol. Biol.*, vol. 8, no. 1, pp. 1–27, 2009.
[5] J. Yan et al., "Transcriptome-guided amyloid imaging genetic analysis via a novel structured sparse learning algorithm," *Bioinformatics*, vol. 30, no. 17, pp. i564–i571, 2014.
[6] J. Yan, S. L. Risacher, K. Nho, A. J. Saykin, and L. Shen, "Identification of discriminative imaging proteomics associations in Alzheimer's disease via a novel sparse correlation model," in *Proc. Pacific Symp. Biocomput.*, 2017, pp. 94–104.
[7] T. Melzer, M. Reiter, and H. Bischof, "Appearance models based on kernel canonical correlation analysis," *Pattern Recognit.*, vol. 36, no. 9, pp. 1961–1971, 2003.
[8] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 1247–1255.
[9] P. M. Thompson, T. Ge, D. C. Glahn, N. Jahanshad, and T. E. Nichols, "Genetics of the connectome," *Neuroimage*, vol. 80, pp. 475–488, 2013.
[10] M. Wang, X. Hao, J. Huang, W. Shao, and D. Zhang, "Discovering network phenotype between genetic risk factors and disease status via diagnosis-aligned multi-modality regression method in Alzheimer's disease," *Bioinformatics*, vol. 35, no. 11, pp. 1948–1957, 2019.
[11] X.-A. Bi, W. Zhou, L. Li, and Z. Xing, "Detecting risk gene and pathogenic brain region in EMCI using a novel GERF algorithm based on brain imaging and genetic data," *IEEE J. Biomed. Health Inform.*, vol. 25, no. 8, pp. 3019–3028, Aug. 2021.
[12] X.-A. Bi et al., "Feature aggregation graph convolutional network based on imaging genetic data for diagnosis and pathogeny identification of Alzheimer's disease," *Brief. Bioinf.*, vol. 23, no. 3, 2022, Art. no. bbac137.
[13] L. Du et al., "Identifying diagnosis-specific genotype–phenotype associations via joint multitask sparse canonical correlation analysis and classification," *Bioinformatics*, vol. 36, no. Supplement_1, pp. i371–i379, 2020.
[14] A. Dragomir, A. G. Vrahatis, and A. Bezerianos, "A network-based perspective in Alzheimer's disease: Current state and an integrative framework," *IEEE J. Biomed. Health Inform.*, vol. 23, no. 1, pp. 14–25, Jan. 2019.
[15] Y. Yamazaki, N. Zhao, T. R. Caulfield, C.-C. Liu, and G. Bu, "Apolipoprotein E and Alzheimer disease: Pathobiology and targeting strategies," *Nature Rev. Neurol.*, vol. 15, no. 9, pp. 501–518, 2019.
[16] J. Wang, M. Li, H. Wang, and Y. Pan, "Identification of essential proteins based on edge clustering coefficient," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 9, no. 4, pp. 1070–1080, Jul./Aug. 2012.
[17] E. Elhamifar and R. Vidal, "Sparse subspace clustering: Algorithm, theory, and applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 11, pp. 2765–2781, Nov. 2013.
[18] L. Du et al., "Detecting genetic associations with brain imaging phenotypes in Alzheimer's disease via a novel structured SCCA approach," *Med. Image Anal.*, vol. 61, 2020, Art. no. 101656.
[19] A. G. Asuero, A. Sayago, and A. Gonzalez, "The correlation coefficient: An overview," *Crit. Rev. Anal. Chem.*, vol. 36, no. 1, pp. 41–59, 2006.
[20] Y. Sun, "Iterative RELIEF for feature weighting: Algorithms, theories, and applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 6, pp. 1035–1051, Jun. 2007.
[21] E. C. Chi, G. I. Allen, H. Zhou, O. Kohannim, K. Lange, and P. M. Thompson, "Imaging genetics via sparse canonical correlation analysis," in *Proc. IEEE 10th Int. Symp. Biomed. Imag.*, 2013, pp. 740–743.
[22] M. Wang, W. Shao, X. Hao, S. Huang, and D. Zhang, "Identify connectome between genotypes and brain network phenotypes via deep self-reconstruction sparse canonical correlation analysis," *Bioinformatics*, vol. 38, no. 8, pp. 2323–2332, 2022.
[23] L. Du et al., "Diagnosis status guided brain imaging genetics via integrated regression and sparse canonical correlation analysis," in *Proc. IEEE 16th Int. Symp. Biomed. Imag.*, 2019, pp. 356–359.
[24] J.-C. Lambert et al., "Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease," *Nature Genet.*, vol. 45, no. 12, pp. 1452–1458, 2013.
[25] N. S. Ryan et al., "Magnetic resonance imaging evidence for presymptomatic change in thalamus and caudate in familial Alzheimer's disease," *Brain*, vol. 136, no. 5, pp. 1399–1414, 2013.
[26] R. Duara et al., "Medial temporal lobe atrophy on MRI scans and the diagnosis of Alzheimer disease," *Neurology*, vol. 71, no. 24, pp. 1986–1992, 2008.
[27] X. Liu et al., "Decreased functional connectivity between the dorsal anterior cingulate cortex and lingual gyrus in Alzheimer's disease patients with depression," *Behav. Brain Res.*, vol. 326, pp. 132–138, 2017.

[28] G. W. van Hoesen, J. C. Augustinack, J. Dierking, S. J. Redman, and R. Thangavel, "The parahippocampal gyrus in Alzheimer's disease: Clinical and preclinical neuroanatomical correlates," *Ann. New York Acad. Sci.*, vol. 911, no. 1, pp. 254–274, 2000.
[29] D. Horínek, A. Varjassyová, and J. Hort, "Magnetic resonance analysis of amygdalar volume in Alzheimer's disease," *Curr. Opin. Psychiatry*, vol. 20, no. 3, pp. 273–277, 2007.
[30] M. De Leon et al., "The hippocampus in aging and Alzheimer's disease," *Neuroimaging Clin. North Amer.*, vol. 5, no. 1, pp. 1–17, 1995.
[31] M. Xia, J. Wang, and Y. He, "BrainNet viewer: A network visualization tool for human brain connectomics," *PLoS One*, vol. 8, no. 7, 2013, Art. no. e68910.
[32] M. Wang, W. Shao, X. Hao, and D. Zhang, "Identify complex imaging genetic patterns via fusion self-expressive network analysis," *IEEE Trans. Med. Imag.*, vol. 40, no. 6, pp. 1673–1686, Jun. 2021.
[33] Y. Shigihara, H. Hoshi, J. Poza, V. Rodríguez-González, C. Gómez, and T. Kanzawa, "Predicting the outcome of non-pharmacological treatment for patients with dementia-related mild cognitive impairment," *Aging (Albany NY)*, vol. 12, no. 23, 2020, Art. no. 24101.
[34] S. Holroyd, M. L. Shepherd, and J. H. Downs III, "Occipital atrophy is associated with visual hallucinations in Alzheimer's disease," *J. Neuropsychiatry Clin. Neurosciences*, vol. 12, no. 1, pp. 25–28, 2000.
[35] D. Pan et al., "Adaptive 3DCNN-based interpretable ensemble model for early diagnosis of Alzheimer's disease," *IEEE Trans. Comput. Social Syst.*, vol. 11, no. 1, pp. 247–266, Feb. 2024, doi: 10.1109/TCSS.2022.3223999.
[36] K. Kantarci et al., "Regional metabolic patterns in mild cognitive impairment and Alzheimer's disease: A 1H MRS study," *Neurology*, vol. 55, no. 2, pp. 210–217, 2000.
[37] Y. Xiong et al., "Altered functional connectivity of basal ganglia in mild cognitive impairment and Alzheimer's disease," *Brain Sci.*, vol. 12, no. 11, 2022, Art. no. 1555.
[38] M. M. Vasavada et al., "Olfactory cortex degeneration in Alzheimer's disease and mild cognitive impairment," *J. Alzheimer's Dis.*, vol. 45, no. 3, pp. 947–958, 2015.
[39] M. D. Ikonomovic, E. E. Abrahamson, B. A. Isanski, J. Wuu, E. J. Mufson, and S. T. DeKosky, "Superior frontal cortex cholinergic axon density in mild cognitive impairment and early Alzheimer disease," *Arch. Neurol.*, vol. 64, no. 9, pp. 1312–1317, 2007.
[40] A. Bahar-Fuchs et al., "Olfactory deficits and amyloid-β burden in Alzheimer's disease, mild cognitive impairment, and healthy aging: A PiB PET study," *J. Alzheimer's Dis.*, vol. 22, no. 4, pp. 1081–1087, 2010.
[41] F. Novellino, M. E. López, M. G. Vaccaro, Y. Miguel, M. L. Delgado, and F. Maestu, "Association between hippocampus, thalamus, and caudate in mild cognitive impairment APOEε4 carriers: A structural covariance MRI study," *Front. Neurol.*, vol. 10, 2019, Art. no. 1303.
[42] R. G. Mair, J. A. Burk, and M. C. Porter, "Impairment of radial maze delayed nonmatching after lesions of anterior thalamus and parahippocampal cortex," *Behav. Neurosci.*, vol. 117, no. 3, 2003, Art. no. 596.