



Maximum mutual information for feature extraction from graph-structured data: Application to Alzheimer's disease classification

Jiawei Yang¹ · Shaoping Wang^{1,2} · Teresa Wu³

Accepted: 18 March 2022 / Published online: 3 May 2022

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2022

Abstract

A brain network can be constructed from various imaging modalities such as magnetic resonance imaging (MRI), representing the functional or structural connectivity between brain regions. The challenge of brain network analysis is efficient dimensionality reduction while retaining feature interpretability. We propose a new method to extract features from graph-structured data based on maximum mutual information (MMI-GSD). First, we develop a novel equation for the feature extraction from GSD and evaluate the interpretability of the features. We establish a framework to optimize the extracted features using the MMI. We conduct experiments on synthetic networks to validate the effectiveness of the proposed MMI-GSD. Next, we conduct experiments on 119 cognitively normal (CN), 105 mild cognitive impairment (MCI), and 36 Alzheimer's disease (AD) individuals from the Alzheimer's Disease Neuroimaging Initiative. The classification performance of the proposed method is significantly better than using traditional network metrics and existing feature extraction methods. In the clinical interpretation, we discover discriminative brain regions showing significant differences between the MCI and AD groups and identify significant abnormal connections concentrated in the left hemisphere.

Keywords Feature extraction · Mutual information · Neuroimaging · Brain network

1 Introduction

Alzheimer's disease (AD) is the most frequent type of neurodegenerative dementia and a growing health problem. Approximately 5.8 million Americans age 65 and older had AD in 2020, and the figure is expected to grow to 13.8 million by 2050 [1]. Nearly \$244 billion worth of care was provided by family members and caregivers in 2019 [2]. The latest statistics of 2018 indicate that 122,019 deaths

have been attributed to AD, making it the sixth leading cause of death in the United States [3]. There is an urgent need to discover the abnormal changes in the patient's brain as early as possible to implement a timely treatment and slow down the disease progression. Extensive studies have been conducted to develop diagnostic technology with reliable biomarkers. One example is diagnostic imaging, e.g., magnetic resonance imaging (MRI), a noninvasive examination of brain structure, function, and connectivity. The physiological changes in the brain caused by AD can be quantitatively analyzed with MRI to detect and monitor the progression of the disease.

The human brain is an extremely complex structure composed of neurons and connections. Therefore, researchers have studied brain functions from the perspective of brain networks, where nodes represent brain regions, and edges connect the regions. Graph analysis is often used to detect connectivity patterns among brain regions to diagnose AD [4–8]. Two data types represent the brain network in a graph: the graph signal and graph weight. A graph signal refers to the attribute of each node, which is typically the brain region of interest (ROI). The node attributes can

Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

✉ Shaoping Wang
shaopingwang@vip.sina.com

Extended author information available on the last page of the article.

be quantitative measures (e.g., volume, thickness, area) of the brain structures detected using segmentation techniques [9–12] or fused features. Magnin et al. [13] parcellated T1-weighted magnetic resonance (MR) images of the brain into 116 ROIs and used a histogram to determine the proportion of gray matter, white matter, and cerebrospinal fluid in the ROIs. The results were input into a support vector machine (SVM) for AD classification. The graph weight describes the connectivity between each pair of nodes. Khzaee et al. [14] parcellated the resting-state functional MR images of the brain into 90 ROIs. A representative signal of each region was acquired by averaging the time series of voxels in the ROI. Functional connectivity networks were constructed, and the edges were defined using Pearson's correlation coefficients of the signals of all pairs of brain regions. An SVM model was developed using the features derived from the local and global graph measures (e.g., node degree) to differentiate AD individuals from healthy controls.

Using a graph model to describe brain networks is a straightforward approach. However, extracting the features from the graph in a meaningful way for AD diagnosis is challenging. Researchers have used some common graph features, including the node degree, clustering coefficient, and the small world. For example, Wee et al. [15] computed the clustering coefficients (common metrics used in brain networks to measure the connectivity among brain regions) as features for SVMs to differentiate mild cognitive impairment (MCI) individuals from cognitively normal (CN) individuals. Prasad et al. [16] developed two connectivity networks. One quantified the pairwise connectivity strength as the relative proportion of fibers connecting the two brain regions. The other quantified the maximum flow between brain regions by interpreting the diffusion image as a flow graph. An SVM classifier was used to differentiate CN, MCI, and AD individuals, and the performance was evaluated using global efficiency, transitivity, path length, modularity, and small world [16]. These methods are based on prior brain network information using domain knowledge (e.g., using a pre-defined brain atlas). In the analysis of mental disorders, domain knowledge improves the classification performance of small sample tasks and the interpretability of experimental results. Although this approach has been successful, studies have mainly focused on deriving features without considering the stochastic nature of the data. This approach may be problematic as data stochasticity prevails due to patient heterogeneity and variability.

Other studies used a data-driven approach for feature extractions that does not require a priori brain network topology. Principal component analysis (PCA) is a common technique for dimensionality reduction; it transforms the data into principal components with fewer dimensions

[17–19]. Salvatore et al. [20] used PCA to reduce the dimensions of white matter and gray matter density maps. The results were fed into an SVM for classifying patients who were likely or unlikely to develop AD. Data-driven approaches have the advantages of computational efficiency, addressing the curse of dimensionality, and being less dependent on domain knowledge. However, we contend that ignoring the inherent topological properties may be problematic because the extracted features may be difficult to interpret. Here we argue that studies on the connectivity among brain ROIs should focus on improving clinical interpretation and using domain knowledge (e.g., brain atlas).

The challenge of AD diagnosis is efficient feature dimension reduction while retaining feature interpretability. High interpretability of features will promote the use of machine learning methods for clinical diagnosis. However, the latest research in brain network analysis has not addressed this issue adequately. In this paper, we propose a novel feature extraction method for graph-structured data based on maximum mutual information (MMI-GSD). The proposed MMI-GSD can efficiently reduce the dimensionality of GSD while retaining the interpretability of the extracted features. This method has promising application prospects in pathological interpretation. We develop a Gaussian graphical model (GGM) for neuroimaging data. We utilize the scale of attention (SOA) concept, describing the range of connection weights for a specific node in the network. Information entropy is employed to quantify the uncertainty of the variables, and mutual information is used as a decision criterion to reduce the degree of uncertainty with respect to (w.r.t.) the knowledge from other variables [21]. Specifically, an optimization problem on the mutual information is constructed to select the salient features with different SOAs. Since the features are derived from the ROIs defined by domain knowledge, they provide meaningful clinical interpretations, enabling biomarker studies to discover new biomarkers. A synthetic dataset and a real AD dataset are used to validate the proposed method. Our method outperforms traditional network metrics and existing feature extraction methods.

2 Related work

Mutual information is a measure of the statistical dependency between random variables. It has also been used as a key measure to evaluate the effectiveness of feature extraction in some recent studies. Marinoni and Gamba [22] proposed a method for identifying affinity patterns using mutual information maximization and validated the method using remote sensing images. Özdenizci and Erdoğan [23] presented an MMI linear transformation and a

nonlinear transformation framework. The proposed method was applied to a brain–computer interface task and was assessed with electroencephalographic data. However, the above feature extraction methods based on the MMI were not specially designed for graph-structured data, which is the most common form of data in the task of brain network analysis.

In addition, neural networks have also been used for automatic feature extraction [24–27], and graph convolutional networks (GCNs) were recently designed for non-Euclidean data [28, 29]. However, there are two problems when applying GCNs to neuroimaging-related tasks [30–32]. First, the feature extraction processes in GCN are automatic, and spectral graph convolutional layers are typically used. Thus, the internal high-level features of the model are difficult to interpret. Second, the GCN inputs are graph signals (i.e., node features). However, we can only obtain graph weights (i.e., adjacency matrices) after preprocessing the diffusion tensor images (DTI). In previous studies, graph weights have represented the similarity between individuals. Graph signals were constructed from the vectorized adjacency matrices or measures of the brain regions without considering the topological properties of individual brain networks. Table 1 lists the characteristics of traditional methods and our proposed method for an intuitive comparison.

We propose the MMI-GSD, which is inspired by graph convolutional layers and considers mutual information. We extract the features from the adjacency matrices spatially, retaining the flexibility of the graph convolutional layer and enhancing the interpretability of the extracted features. The main contributions of this paper are as follows: (1) We develop a novel feature extraction method for GSD and discuss its interpretability. (2) We develop a framework to optimize the extracted features based on MMI. (3) We carry out experiments to verify the performance of the proposed method. We verify its applicability on a real AD dataset and analyze the interpretability of the features. The remainder of the paper is organized as follows. Section 2

presents the feature extraction method and the optimization framework. Section 3 describes experiments using synthetic and real data. In Section 4, we discuss the physiological meaning of the experimental results. Section 5 concludes the paper.

3 Methodology

The aim of feature extraction is to reduce the dimensionality of the data. In this study, the data dimension is $D \times D$, where D is the number of nodes in the individual brain network. The proposed MMI-GSD derives the M ($M < D$) feature vectors of dimension D . This section introduces the proposed MMI-GSD. As shown in Fig. 1, given a graph describing the brain network, the features are derived from the adjacency matrix and are fit to multivariate distributions. Next, a GGM is developed based on the observations. The entropy is calculated to consider the uncertainty of the variables. The mutual information obtained from the GGM is used as the criterion to evaluate the quality of the extracted features. An optimization problem is constructed to maximize the mutual information. The MMI-GSD identifies the salient network features, which are input into the classifiers.

3.1 Data acquisition and preprocessing

The data used in the experiments were obtained from the Alzheimer’s Disease Neuroimaging Initiative 3 (ADNI3) database (adni.loni.usc.edu) [33, 34]. The ADNI3 began in 2016 and includes scientists at 59 research centers in the United States and Canada. The primary goal of the ADNI is to measure the progression of MCI and early AD. In this study, the selected subjects included three cohorts: CN, MCI, and AD. We used T1-weighted MR images and DTI. We selected imaging scans from the same manufacturer (SIEMENS) to avoid data discrepancy. Images from baseline visit, initial visit, and screening visit of 260 subjects (119 CN, 105 MCI, and 36 AD) were used.

The DTI data processing and white matter network construction were performed using the PANDA toolbox [35]. Fiber Assignment by Continuous Tracking (FACT), a deterministic fiber-tracking algorithm, was used with an angle threshold of 45° and a fractional anisotropy (FA) range of 0.2–1. The brain was segmented into 90 ROIs using the automated anatomical labeling (AAL) atlas [36]. The nodes in the network were defined by the ROIs, and the edges were defined by the number of fibers connecting two ROIs. The construction of the white matter brain network is shown in Fig. 2.

Table 1 Comparison of the Characteristics of Different Methods

	Flexibility	Interpretability	Topology
Graph metric methods	↓	↑	○
PCA methods	↑	↓	×
Existing mutual information methods	↑	↑	×
GCN methods	↑	↓	○
Our proposed MMI-GSD	↑	↑	○

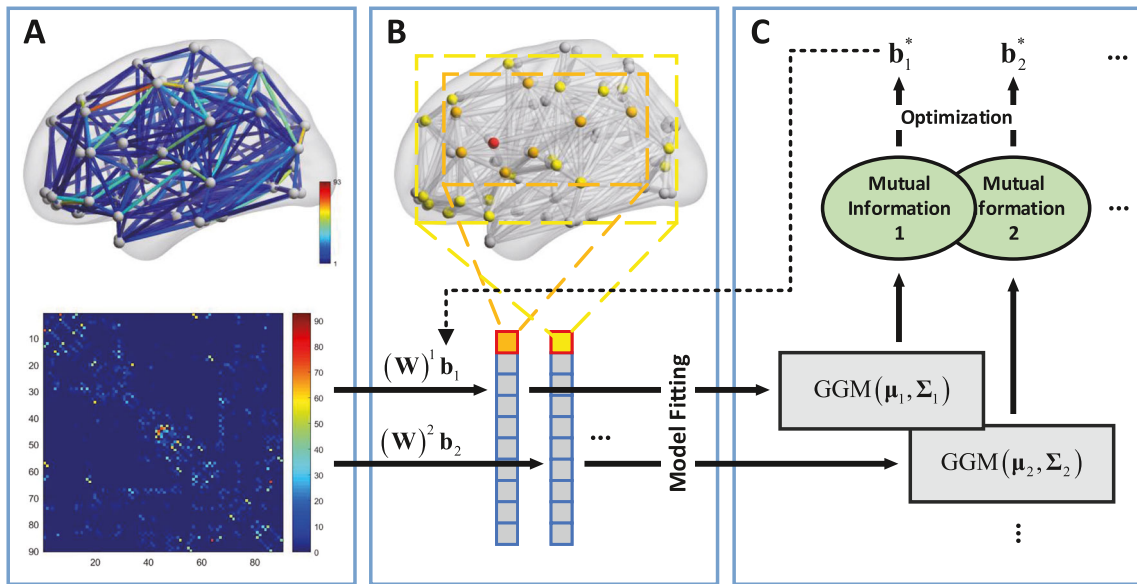


Fig. 1 The framework of the MMI-GSD. (A) A brain network with D nodes (Fig. 1A top), $D \times D$ adjacency matrix of the network (Fig. 1A bottom). (B) The adjacency matrix of different powers is multiplied by a direction vector \mathbf{b}_m to obtain several column vectors

(Fig. 1B bottom). The greater the power, the more connection information is contained in the column vector. The adjusted brain network after the multiplication operation (Fig. 1B top). (C) Fitting the GGM; the mutual information is used to optimize \mathbf{b}_m^*

3.2 Feature extraction from the graph using direction vectors

Given N weighted undirected graphs \mathcal{G}_n ($n = 1, \dots, N$), let $\mathcal{G}_n = (\mathcal{V}_n, \mathcal{E}_n, \mathbf{W}_n)$, where \mathcal{V}_n is a set of nodes (vertices), and $|\mathcal{V}_n| = D$, \mathcal{E}_n is a set of edges. \mathbf{W}_n is the adjacency matrix, and element $W_{n,kl}$ is the connection weight between node k and node l ($k \neq l$). Let $(\mathbf{W}_n)^m$ be the m^{th} power

of \mathbf{W}_n , indicating the connectivity between any two nodes within m hops. With a direction vector \mathbf{b}_m , we obtain:

$$\mathbf{x}_{n,m} = (\mathbf{W}_n)^m \mathbf{b}_m. \tag{1}$$

The k^{th} component in $\mathbf{x}_{n,m}$ represents a linear combination of weights of the m -hop pathways connected to the k^{th} node. According to the concept of ‘‘hop’’ in a network, an

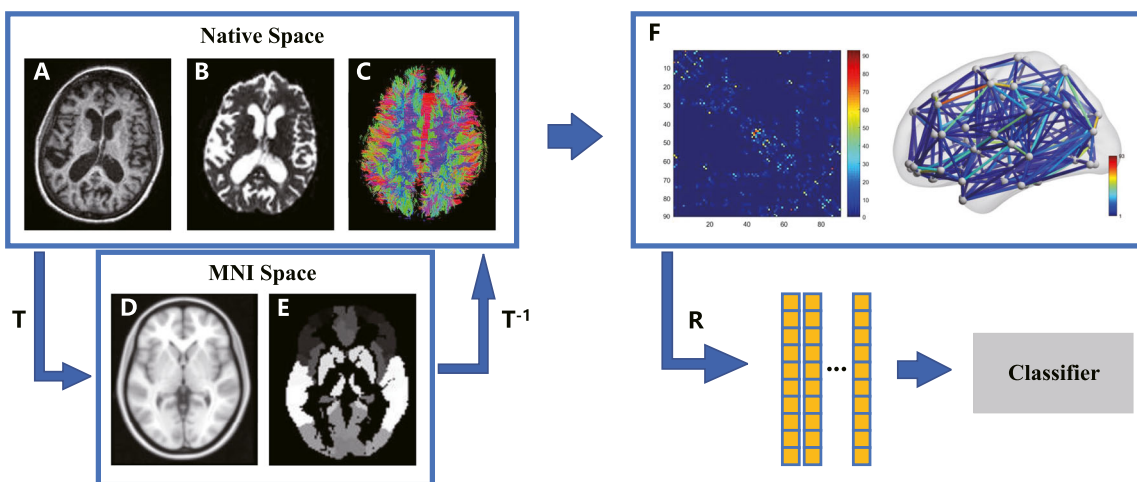


Fig. 2 Construction of the white matter brain network. (1) Registration of DTI image (B) to T1-weighted image (A) in the native space. (2) Deterministic fiber tracking (C). (3) Registration (T) of T1-weighted image in the native space to the ICBM152 T1 template in the Montreal Neurological Institute (MNI) space [37] (D). (4) Inverse

transformation (T^{-1}) to the AAL atlas in the MNI space. (5) The brain connectivity matrix (F) is calculated by counting the fiber numbers between each pair of brain regions defined by the AAL atlas. (6) Data dimensionality reduction (R) is performed on the brain connectivity matrix, and the feature vectors are obtained as the input of the classifier

increase in m indicates that a broader region of the network is being explored. Here we call m the SOA. For example, when m is 1, $\mathbf{b}_1 = [1 \ 1 \ \dots \ 1]^T$,

$$\mathbf{x}_{n,1} = (\mathbf{W}_n)^1 \mathbf{b}_1 = \begin{bmatrix} \sum_{l=1}^D W_{n,1l} \\ \vdots \\ \sum_{l=1}^D W_{n,kl} \\ \vdots \\ \sum_{l=1}^D W_{n,Dl} \end{bmatrix}, \tag{2}$$

where $\mathbf{x}_{n,1}$ is a column vector. The features $\mathbf{x}_{n,m}$ are extracted independently from different SOAs. Next, we create matrix $\mathbf{X}_n = [\mathbf{x}_{n,1} \ \mathbf{x}_{n,2} \ \dots \ \mathbf{x}_{n,M}] \in \mathbb{R}^{D \times M}$ where $M < D$. The matrix \mathbf{X}_n contains the extracted features with reduced dimensionality. Note that M is the number of direction vectors \mathbf{b}_m ; it is related to the scale of the network and is determined empirically.

3.3 Gaussian graphical model fitting with direction vector as a prior

Without loss of generality, we remove subscript m for simplification. With respect to the direction vector \mathbf{b} , we obtain the column feature vector \mathbf{x}_n . We use the GGM to fit the observations $\{\mathbf{x}_n\}$. Specifically, the vector $\mathbf{x} \in \mathbb{R}^D$ with a multivariate Gaussian (MVG) distribution $\mathcal{N}_D(\boldsymbol{\mu}, \Sigma)$ has the following density function:

$$f(\mathbf{x}; \boldsymbol{\mu}, \Sigma) = \frac{1}{(2\pi)^{\frac{D}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}, \tag{3}$$

where $\boldsymbol{\mu}$ is the mean, Σ is the covariance matrix, $|\Sigma|$ is the determinant of Σ , and $\Theta, \Theta = \Sigma^{-1}$ is the precision matrix.

Let $\mathcal{G}_N = (\mathcal{V}, \mathcal{E})$ be an undirected graph where each node represents a component of the vector $\mathbf{x} \in \mathbb{R}^D$ ($|\mathcal{V}| = D$). Vector \mathbf{x} satisfies the (undirected) GGM with graph \mathcal{G}_N if it has an MVG distribution $\mathcal{N}_D(\boldsymbol{\mu}, \Sigma)$ with the following constraint:

$$(\Theta)_{i,j} = \left(\Sigma^{-1}\right)_{i,j} = 0 \quad \text{for all } (i, j) \notin \mathcal{E}. \tag{4}$$

The constraint means that the variables on nodes i and j are conditionally independent given the variables on the other nodes if there is no edge between nodes i and j . If the two variables are independent, the element in the precision matrix Θ is equal to 0.

With observations $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\}$, the likelihood function is defined as:

$$L(\boldsymbol{\mu}, \Sigma; \mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N) = \prod_{n=1}^N f(\mathbf{x}_n; \boldsymbol{\mu}, \Sigma) = \frac{1}{(2\pi)^{\frac{ND}{2}} |\Sigma|^{\frac{N}{2}}} \exp\left\{-\frac{1}{2} \sum_{n=1}^N (\mathbf{x}_n - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x}_n - \boldsymbol{\mu})\right\}. \tag{5}$$

The log-likelihood function of (5) is:

$$l(\boldsymbol{\mu}, \Sigma) = \log L(\boldsymbol{\mu}, \Sigma; \mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N) = -\frac{ND}{2} \log(2\pi) - \frac{N}{2} \log|\Sigma| - \frac{1}{2} \sum_{n=1}^N (\mathbf{x}_n - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x}_n - \boldsymbol{\mu}). \tag{6}$$

We remove the first term, which is a constant, and re-write the log-likelihood function as:

$$l(\boldsymbol{\mu}, \Sigma) = -\frac{N}{2} \log|\Sigma| - \frac{1}{2} \sum_{n=1}^N (\mathbf{x}_n - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x}_n - \boldsymbol{\mu}) = \frac{N}{2} \left(-\log|\Sigma| - \text{Tr}(\Sigma^{-1} \mathbf{S}) - (\boldsymbol{\mu} - \bar{\mathbf{x}})^T \Sigma^{-1} (\boldsymbol{\mu} - \bar{\mathbf{x}})\right), \tag{7}$$

where $\bar{\mathbf{x}}$ and \mathbf{S} are the empirical mean and covariance, respectively [38].

The maximum likelihood estimation (MLE) of GGM is called the covariance selection [39]. It is expressed with constraints as:

$$\begin{aligned} &\text{maximize} \quad \log|\Theta| - \text{Tr}(\Theta \mathbf{S}) - (\boldsymbol{\mu} - \bar{\mathbf{x}})^T \Theta (\boldsymbol{\mu} - \bar{\mathbf{x}}) \\ &\text{subject to} \quad \Theta_{ij} = 0, \text{ if } (i, j) \notin \mathcal{E}, \end{aligned} \tag{8}$$

with the domain $\{(\boldsymbol{\mu}, \Theta) \in \mathbb{R}^N \times \mathbb{R}^{N \times N} | \Theta \succ 0, \Theta = \Theta^T\}$. Since Θ is a positive definite matrix, the third term of the objective function $(\boldsymbol{\mu} - \bar{\mathbf{x}})^T \Theta (\boldsymbol{\mu} - \bar{\mathbf{x}}) > 0$ if and only if $\boldsymbol{\mu} - \bar{\mathbf{x}} \neq 0$. To maximize the log-likelihood function, we have

$$\hat{\boldsymbol{\mu}} = \bar{\mathbf{x}} = \frac{1}{N} \sum_{n=1}^N \mathbf{x}_n. \tag{9}$$

The problem construct can be simplified as:

$$\begin{aligned} &\text{maximize} \quad l(\Theta) = \log|\Theta| - \text{Tr}(\Theta \mathbf{S}) \\ &\text{subject to} \quad \Theta_{ij} = 0, \text{ if } (i, j) \notin \mathcal{E}. \end{aligned} \tag{10}$$

This equality-constrained convex optimization problem can be solved using a modified regression algorithm due to its simplicity and computational efficiency [40]. The outcome of this process is a GGM with a given direction vector \mathbf{b} . Next, we will discuss the use of the entropy and mutual information to assess the quality of the extracted features.

3.4 Mutual information maximization for director vector identification

The selection of the direction vector \mathbf{b} affects the distribution of the observations $\{\mathbf{x}_n\}$. In this research, we utilize information entropy to assess the impact of \mathbf{b} quantitatively. Information entropy describes the uncertainty of a random variable, and mutual information measures the dependence between two random variables.

The entropy is defined as follows for a single discrete random variable:

$$H(X) = - \sum_x p(x) \log p(x), \tag{11}$$

where $p(x)$ is the probability mass function. For continuous variables, the entropy is:

$$H(X) = - \int_x f(x) \log f(x) dx. \tag{12}$$

For a vector of random variables with density $f(x_1, \dots, x_D)$, the joint entropy is:

$$H(X_1, X_2, \dots, X_D) = - \int f(x^D) \log f(x^D) dx^D. \tag{13}$$

The conditional entropy is denoted by $H(X|Y)$, which measures the entropy of a random variable X conditional on the knowledge of the random variable Y . Mutual information is defined as:

$$\begin{aligned} I(X; Y) &= \sum_{x,y} p(x, y) \log \frac{p(x, y)}{p(x)p(y)} \\ &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X). \end{aligned} \tag{14}$$

We consider a classification task where X are the observations and Y are the labels (responses) of the observations. Mutual information can be used to evaluate the features extracted from the observations. If $H(Y)$ is constant, $H(Y|X)$ decreases as $I(X; Y)$ increases. We can extract a set of features to maximize $I(X; Y)$ and minimize the uncertainty Y .

When graph \mathcal{G}_N is a complete graph, we first consider the MLE of GGM (in (10)) without constraints, which is similar to an MVG distribution:

$$\text{maximize } l(\Theta) = \log |\Theta| - \text{Tr}(\Theta \mathbf{S}). \tag{15}$$

The gradient of the objective function is a log-likelihood function:

$$\nabla l(\Theta) = \Theta^{-1} - \mathbf{S}. \tag{16}$$

For MLE, we have $\nabla l(\Theta) = 0$:

$$\hat{\Sigma} = \hat{\Theta}^{-1} = \mathbf{S}. \tag{17}$$

Considering (12) and (13), the entropy of MVG [21] is:

$$\begin{aligned} H(X_1, X_2, \dots, X_D) &= H(\mathcal{N}_D(\boldsymbol{\mu}, \Sigma)) \\ &= \frac{1}{2} \log \left[(2\pi e)^D |\Sigma| \right] \\ &= \frac{1}{2} \log \left[(2\pi e)^D |\Theta|^{-1} \right] \\ &= \frac{D}{2} \log(2\pi e) - \frac{1}{2} \log |\Theta|. \end{aligned} \tag{18}$$

The mutual information between $X^D = (X_1, \dots, X_D)$ and Y is:

$$I(X^D; Y) = H(X^D) - H(X^D|Y), \tag{19}$$

and it can be calculated by

$$\begin{aligned} H(X^D|Y) &= \sum_c P(Y = y_c) H(X^D|Y = y_c) \\ &= \sum_c \frac{N_c}{N} H(\mathcal{N}_D(\boldsymbol{\mu}_c, \Sigma_c)) \\ &= \sum_c \frac{N_c}{N} \left(\frac{D}{2} \log(2\pi e) - \frac{1}{2} \log |\Theta_c| \right) \\ &= \frac{D}{2} \log(2\pi e) - \frac{1}{2} \log \left(\prod_c |\Theta_c|^{\frac{N_c}{N}} \right), \end{aligned} \tag{20}$$

where

$$\begin{aligned} N &= \sum_c N_c \\ \Theta_c &= \Sigma_c^{-1}. \end{aligned} \tag{21}$$

By substituting (18) and (20) into (19), we obtain:

$$\begin{aligned} I(X^D; Y) &= \frac{1}{2} \log \left(\prod_c |\Theta_c|^{\frac{N_c}{N}} \right) - \frac{1}{2} \log |\Theta| \\ &= \frac{1}{2} \log \frac{\prod_c |\Theta_c|^{\frac{N_c}{N}}}{|\Theta|}. \end{aligned} \tag{22}$$

When the graph structure of the GGM is complete,

$$\begin{aligned} I(X^D; Y) &= \frac{1}{2} \log \frac{\prod_c |\hat{\Theta}_c|^{\frac{N_c}{N}}}{|\hat{\Theta}|} \\ &= \frac{1}{2} \log \frac{|\mathbf{S}|}{\prod_c |\mathbf{S}_c|^{\frac{N_c}{N}}}. \end{aligned} \tag{23}$$

Without loss of generality, let the power of \mathbf{W}_n equal one; then

$$\bar{\mathbf{x}} = \frac{1}{N} \sum_{n=1}^N \mathbf{x}_n = \frac{1}{N} \sum_{n=1}^N \mathbf{W}_n \mathbf{b} = \bar{\mathbf{W}} \mathbf{b}. \tag{24}$$

We substitute (24) into the empirical covariance matrix \mathbf{S} and obtain:

$$\begin{aligned}\mathbf{S} &= \frac{1}{N} \sum_{n=1}^N (\mathbf{W}_n \mathbf{b} - \bar{\mathbf{W}} \mathbf{b}) (\mathbf{W}_n \mathbf{b} - \bar{\mathbf{W}} \mathbf{b})^T \\ &= \frac{1}{N} \sum_{n=1}^N (\mathbf{W}_n - \bar{\mathbf{W}}) \mathbf{b} \mathbf{b}^T (\mathbf{W}_n - \bar{\mathbf{W}})^T.\end{aligned}\quad (25)$$

Let $\mathbf{b} = \lambda \mathbf{a}$, and λ be a scalar; we obtain:

$$\begin{aligned}\mathbf{S}^{(b)} &= \frac{\lambda^2}{N} \sum_{n=1}^N (\mathbf{W}_n - \bar{\mathbf{W}}) \mathbf{a} \mathbf{a}^T (\mathbf{W}_n - \bar{\mathbf{W}})^T \\ &= \lambda^2 \mathbf{S}^{(a)},\end{aligned}\quad (26)$$

and

$$\begin{aligned}I(X^D; Y) &= \frac{1}{2} \log \frac{|\mathbf{S}^{(b)}|}{\prod_c |\mathbf{S}^{(b)}|^{\frac{N_c}{N}}} \\ &= \frac{1}{2} \log \frac{|\lambda^2 \mathbf{S}^{(a)}|}{\prod_c |\lambda^2 \mathbf{S}^{(a)}|^{\frac{N_c}{N}}} \\ &= \frac{1}{2} \log \frac{\lambda^{2D} |\mathbf{S}^{(a)}|}{\prod_c \lambda^{2D} |\mathbf{S}^{(a)}|^{\frac{N_c}{N}}} \\ &= \frac{1}{2} \log \frac{|\mathbf{S}^{(a)}|}{\prod_c |\mathbf{S}^{(a)}|^{\frac{N_c}{N}}}.\end{aligned}\quad (27)$$

Equation (27) indicates that the mutual information between X^D and Y is not influenced by the scalar λ . Thus, we can regard the direction vector \mathbf{b} as a unit vector. By restricting the space of \mathbf{b} to a hypersphere, we can reduce the range of the search space. The mutual information is not affected when we scale the adjacency matrix to the range of 0 to 1.

Next, we revisit (10) with constraints. The optimization problem construct is:

$$\begin{aligned}\text{maximize} \quad & I(X^D; Y) = \frac{1}{2} \log \frac{\prod_c |\hat{\Theta}_c(\mathbf{b})|^{\frac{N_c}{N}}}{|\hat{\Theta}(\mathbf{b})|} \\ \text{subject to} \quad & \|\mathbf{b}\|_2 = 1,\end{aligned}\quad (28)$$

where the estimation of the precision matrix (Θ) is a function of the direction vector \mathbf{b} . This is a non-convex optimization problem. We choose a particle swarm optimization (PSO) solver due to its simplicity and high convergence rate [41]. Interested readers are referred to [41] for the technical details on PSO. Algorithm 1 describes the MMI-GSD.

Algorithm 1 MMI-GSD.

Input : W_n : training dataset with N instances
 Y_n : labels for N brain networks
 G : the graph on which the GGM is based
 M : maximum value of SOA
 P : the number of particles

Output: X_n : N feature matrices of dimension $D \times M$, the m -th column is $X_{n,m}$

```

1 for  $m \leftarrow 1$  to  $M$  do
2   Randomly initialize  $P$  direction vectors  $b_m$ ;
3   repeat
4     Calculate  $X_{n,m} = (W_n)^m b_m$  for all
       instances;
5     Calculate the empirical covariance matrix
        $S$  with  $X_{n,m}$  of all instances, and  $S_c$  for
       each class;
6     Estimate GGM with  $S/S_c$  and  $G$ ;
7     Calculate mutual information  $I$  with  $S$ ,  $S_c$ ,
        $Y_n$ ;
8     Update  $b_m$  with PSO;
9   until convergence;
10 end
```

4 Experiments

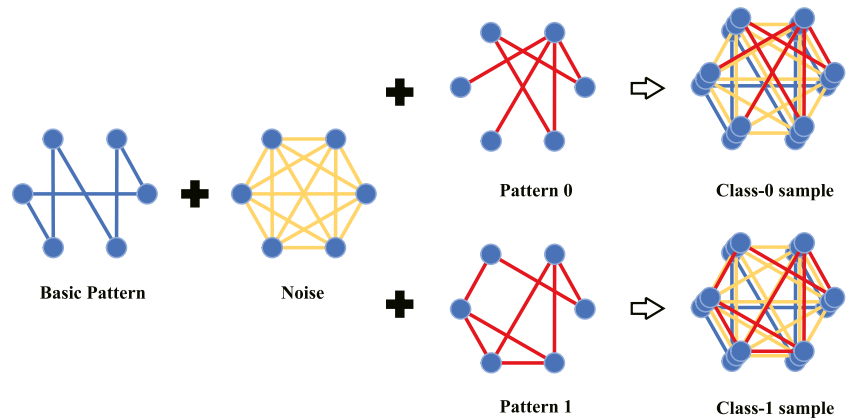
4.1 Experiment on synthetic networks

We first conducted an experiment on synthetic networks to analyze the properties of the MMI-GSD. They included two patterns of networks; a two-class classification was carried out. We evaluated multiple characteristics of the proposed method by adjusting the hyperparameter settings. We also compared the MMI-GSD with the traditional PCA method to evaluate the feature extraction ability of the MMI-GSD for graph data.

4.1.1 Synthetic network generation

We generated small networks with 6 nodes to evaluate the impacts of the parameter settings (see Fig. 3). We assumed that the two classes of networks share basic connection patterns with noise of the same distribution. In addition, the two classes of networks have different network-dependent connection patterns (Pattern 0 vs. Pattern 1). The shared basic patterns, network-dependent patterns, and noise result in two classes of synthetic networks. The network weights are constrained to non-negative integers to represent the number of fiber connections between two brain regions (note this research focuses on neuroimaging and AD). The noise added to the weights follows the Poisson distribution, and the weights have a

Fig. 3 Generating the two patterns of the networks



Gaussian distribution. As a result, two groups of networks were generated: 2000 Class-0 samples and 2000 Class-1 samples. The synthetic network generation is described in Algorithm 2.

Algorithm 2 Synthetic Network Generation.

Input : N : the number of synthetic networks to be generated

Output: G_0, G_1 : synthetic networks

- 1 Generate Watts–Strogatz small-world graphs $G, G_0, G_{00}, G_{000}, G_1, G_{11}, G_{111}$ with Python package (NetworkX);
- 2 Initial $\mu, \sigma, \mu_0, \sigma_0, \mu_{00}, \sigma_{00}, \mu_{000}, \sigma_{000}, \mu_1, \sigma_1, \mu_{11}, \sigma_{11}, \mu_{111}, \sigma_{111}, \lambda$;
- 3 **repeat** N **times**
- 4 Randomly initialize graph weights of $G_0, G_{00}, G_{000}, G_1, G_{11}, G_{111}$, following Gaussian distribution $N(\mu_0, \sigma_0), N(\mu_{00}, \sigma_{00}), N(\mu_{000}, \sigma_{000}), N(\mu_1, \sigma_1), N(\mu_{11}, \sigma_{11}), N(\mu_{111}, \sigma_{111})$;
- 5 Randomly initialize graph weights of G , following Gaussian distribution $N(\mu, \sigma)$;
- 6 Add noise to G , following Poisson distribution $P(\lambda)$;
- 7 $G_0 \leftarrow G + G_0 + G_{00} + G_{000}$;
- 8 $G_1 \leftarrow G + G_1 + G_{11} + G_{111}$;
- 9 **end**

4.1.2 Gaussian graphical model

The GGM was derived based on the number of synthetic networks. Since the connection weights in the networks contain noise, we first focused on the precision matrix and used sparse constraints to identify zero elements. This approach reduces the number of GGM parameters to be estimated, which is crucial for studies with limited data. We performed a two-sample t-test on each connection weight in a set of network observations. A higher p-value cutoff indicates more connections and a higher sparsity

rate (SR), which is the ratio of the number of existing edges to the number of all possible edges. The edge with a predetermined SR (see Section 4.1.3 describing the experiments on different SRs) was removed from the graph. Since the GGM estimation requires a connected graph structure, we ensured that each node had at least one edge connected to other nodes.

4.1.3 Feature extraction based on maximum mutual information

Among the 4000 synthetic networks, 3600 samples were used as the training set, and 400 samples were used as the testing set. First, we performed feature extraction on the training set. A PSO was employed as the solver for the non-convex optimization problem (Section 3.4.). We used different SRs (SR=0.2, 0.4, 0.6, 0.8, 1) during preprocessing. The best performance was achieved for SR=0.2. We report the results of SR=0.2 in the following and summarize the overall performance at the end of the section.

As shown in Fig. 4, we used 20 particles for searching the extrema and tracked the mutual information convergence when m in \mathbf{W}^m is equal to 1. The dotted red line represents the MMI value of the particle group. The MMI reaches 99% (the global optimum obtained from PSO) at the 27th iteration (black star). Although the focus of this research is not PSO, it is noteworthy that the mutual information trajectory of each particle fluctuates substantially at the beginning, and most particles converge after 700 iterations. Figure 5 shows the convergence of the mutual information for different numbers of particles (5, 10, 20, and 30). The four curves converge to the maximum after 25 iterations. The convergence for 5 and 10 particles occurs at a lower MMI value than for 20 and 30 particles, indicating that the number of particles is insufficient for this search space. In addition, the optimal solution is slightly better for 30 particles than for 20 particles, but a larger number of particles results in greater computation and learning time.

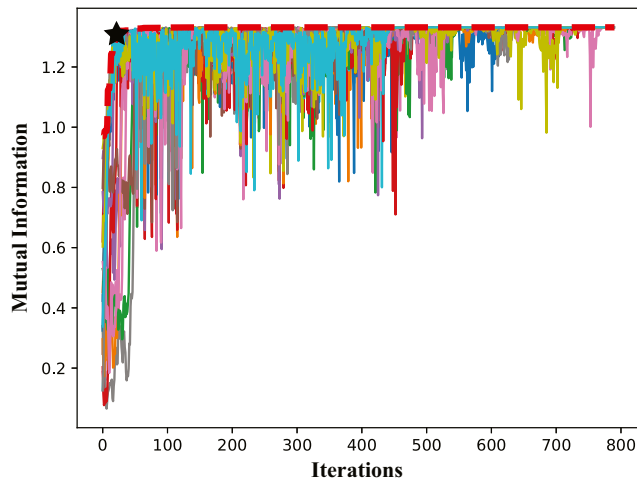


Fig. 4 Mutual information convergence (SR=0.2). The mutual information value of a group of 20 particles increases during the iterations of the particle swarm optimization algorithm. The dotted red line shows the mutual information corresponding to the globally optimal particle

The PSO algorithm balances the number of particles and the learning performance.

We performed feature extraction on the network weights using different SOA values, i.e., we changed the m in \mathbf{W}^m from 1 to 6. The bigger the value of m , the higher the information level of the feature extraction is. Figure 6 illustrates the mutual information convergence for different values of m . The curves converge after about 50 iterations. The mutual information value is the highest value for an SOA of 6 (mutual information: 1.77), followed by 3 (mutual information: 1.49), 2 (mutual information: 1.39), 1 (mutual information: 1.32), 4 (mutual information: 1.27), and 5 (mutual information: 1.26). These results indicate that features with different SOAs describe the different

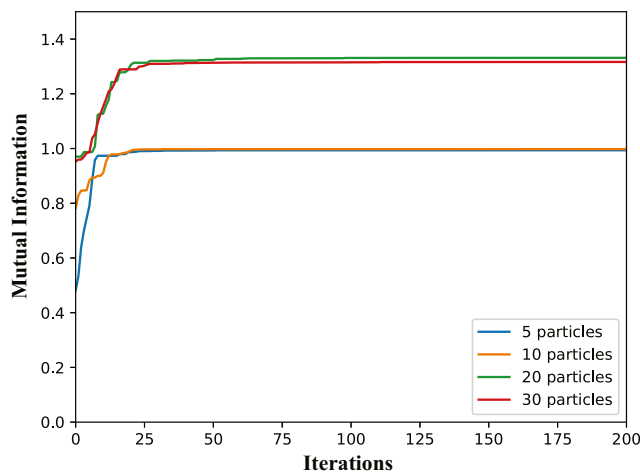


Fig. 5 Mutual information convergence with different numbers of particles

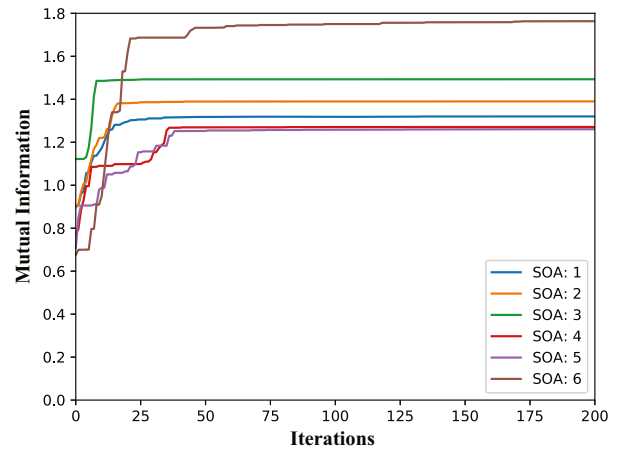


Fig. 6 Mutual information convergence with different scale of attention (SOA) values (SR=0.2)

characteristics of the network. For example, some features may describe the local graph features (the SOA m is small), whereas some features may describe the connectivity at a larger range (the SOA m is large). This result demonstrates the need for optimization to identify the salient features at different scales.

We visualized the extracted features for $m=1$ to 6 using 400 samples from the testing set (Fig. 7). We used the t-distributed stochastic neighbor embedding (t-SNE) method to map the high-dimensional samples into a two-dimensional plane. Figure 7(A) – (F) are the visualization maps for the six SOAs. Figure 7(G) shows all extracted features from the six SOAs, and Fig. 7(H) depicts the sample after PCA transformation. It is observed that the features from different SOAs have different levels of discriminative power to differentiate the two groups of networks.

4.1.4 Classification of synthetic networks

We used the extracted features and transformed each original brain network \mathbf{W}_n into a feature matrix \mathbf{X}_n composed of multiple feature column vectors $(\mathbf{W}_n)^m \mathbf{b}_m$. Each column vector corresponds to a different SOA. Fisher’s scoring, a widely used supervised feature selection method, was applied for feature selection. For a set of labeled scalar observations $\mathbf{O} = \{O_1, O_2, \dots, O_N\}$, the mean and variance of each class is μ_c and σ_c . The Fisher’s score of this set of observations is defined as:

$$F(\mathbf{O}) = \frac{\sum_c N_c (\mu_c - \mu)^2}{\sum_c N_c (\sigma_c)^2}, \tag{29}$$

where μ is the mean of all observations.

We obtain 36 features from the 6-node synthetic networks with 6 SOAs. The Fisher’s scores were sorted from large to small (Fig. 8). We observe that the features with a larger Fisher’s score are distributed in different

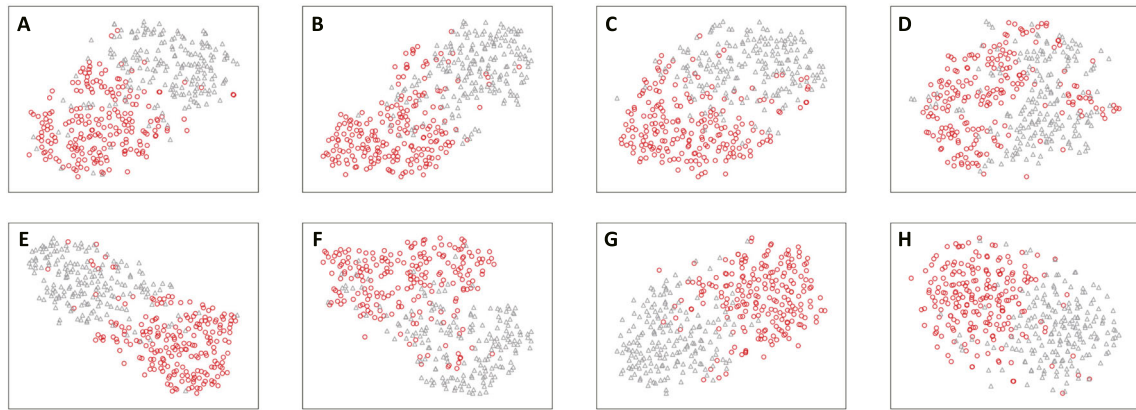


Fig. 7 Visualizations of 400 samples in the testing set (SR=0.2). (A)-(F) correspond to different SOAs from 1 to 6. (G) contains all extracted features of different SOAs. (H) contains the features transformed by PCA

SOAs, confirming our hypothesis that the features from different SOAs contribute to the classification. We used Fisher’s score to select the top 10 features as inputs to a classifier.

We adopted an SVM with a Gaussian radial basis function kernel for classification. The SVM is a nonlinear classifier that performs spatial partitioning of data with high-dimensional complex features. The classification accuracies for different SRs are listed in Table 2. The highest accuracy (95.0%) is obtained for an SR of 0.2. As the SR increases (the network becomes more connected), the classification accuracy decreases. The accuracy is 92.0% for a fully connected graph (SR=1). For comparison, we implemented PCA. We chose to extract the same number of features as 10, the final classification accuracy rate is 92.8%.

4.2 Alzheimer’s disease classification experiments

4.2.1 Brain network preprocessing to reduce dimensionality

A brain network with 90 ROIs would be too computationally expensive and may result in overfitting for our small dataset. Similar to the experiment on the synthetic network, we extracted a smaller brain network. Specifically, we first performed a t-test for each edge weight in the network and retained the edges with significant connectivity. We divided the dataset into 10 parts to ensure the robustness of the derived network. In each run, 9 parts were chosen to identify which edges should be retained. 10 runs were conducted. The nodes connected by the retained edges in all 10 runs were used in the smaller brain network.

4.2.2 Feature extraction based on maximum mutual information and graph metrics

Three classification experiments were conducted: CN vs. AD, CN vs. MCI, and MCI vs. AD. After preprocessing the brain network, the original 90-node network was reduced to a network with fewer nodes (16 for all three comparisons). The MMI-GSD was implemented with different SOAs. Based on our preliminary experiments, we

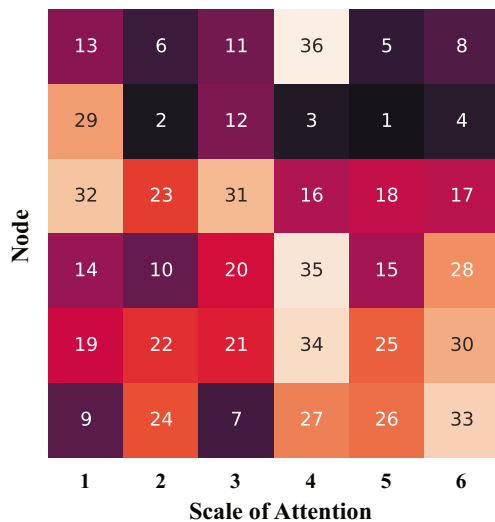


Fig. 8 P-value ranks of the extracted features (SR=0.2)

Table 2 Classification accuracy for Different Sparsity Rates (SRs)

Sparsity Rate (SR)	Accuracy (%)
SR=0.2	95.0
SR=0.4	94.5
SR=0.6	93.2
SR=0.8	92.8
SR=1	92.0
PCA	92.8

used SOAs ranging from 1 to 4. Four direction vectors were obtained in each classification task after mutual information optimization, as shown in Fig. 9. Each element in a direction vector has values ranging from -1 to 1, indicating the impact of the node on maximizing mutual information.

Similar to the synthetic network experiment, we compared our method with the PCA and other state-of-the-art methods. In addition, we selected the most commonly used graph metrics to extract the network features: degree, average neighbor degree, and clustering coefficient. The metrics are indicators of nodal centrality, network resilience, and functional segregation [42]. The weighted degree of node k is the sum of the edge weights for edges adjacent to that node; it is defined as:

$$Deg(v_k) = \sum_{v_l \in \mathcal{V}_k} W_{kl}, \tag{30}$$

where \mathcal{V}_k is the set of neighbors of node k .

The average neighbor degree of node k is the average degree of the neighbors of that node:

$$AvgDeg(v_k) = \frac{1}{|\mathcal{V}_k|} \sum_{v_l \in \mathcal{V}_k} Deg(v_l). \tag{31}$$

For unweighted graphs, the clustering coefficient of node k is the fraction of the triangles passing through that node to all possible triangles; it is defined as:

$$Cluster(v_k) = \frac{2T(v_k)}{Deg(v_k)(Deg(v_k) - 1)}, \tag{32}$$

where $T(v_k)$ is the number of triangles passing through node k . After calculating the graph metrics, we conducted feature selection based on Fisher's scoring to reduce the dimensions of the learning model inputs.

We implemented the t-SNE method to visualize the discriminative power from the features selected from the proposed MMI-GSD, commonly used graph metrics and PCA (see Fig. 10). The degree of separation indicates the difference between the two classes of the samples. The visualization results are not necessarily consistent with the classification results since the samples may be discriminated with higher dimensional features (see Section 4.2.3).

4.2.3 Classification results

Fisher's scoring was used for feature selection. The selected features were fed into classifiers for three tasks: CN vs. AD, CN vs. MCI, MCI vs. AD. Ten-fold cross-validation was implemented to prevent overfitting. Three performance metrics were calculated: accuracy, sensitivity, and specificity. We selected some samples from one group to match with the other group with fewer samples to avoid sample imbalance. Since the samples came from multiple clinical sites, we gave priority to the subjects from the same sites when matching samples, and we aimed for a similar male to female ratio. We obtained 38/36 (CN vs. AD), 119/105 (CN vs. MCI), and 39/36 (MCI vs. AD) samples in the three experiments.

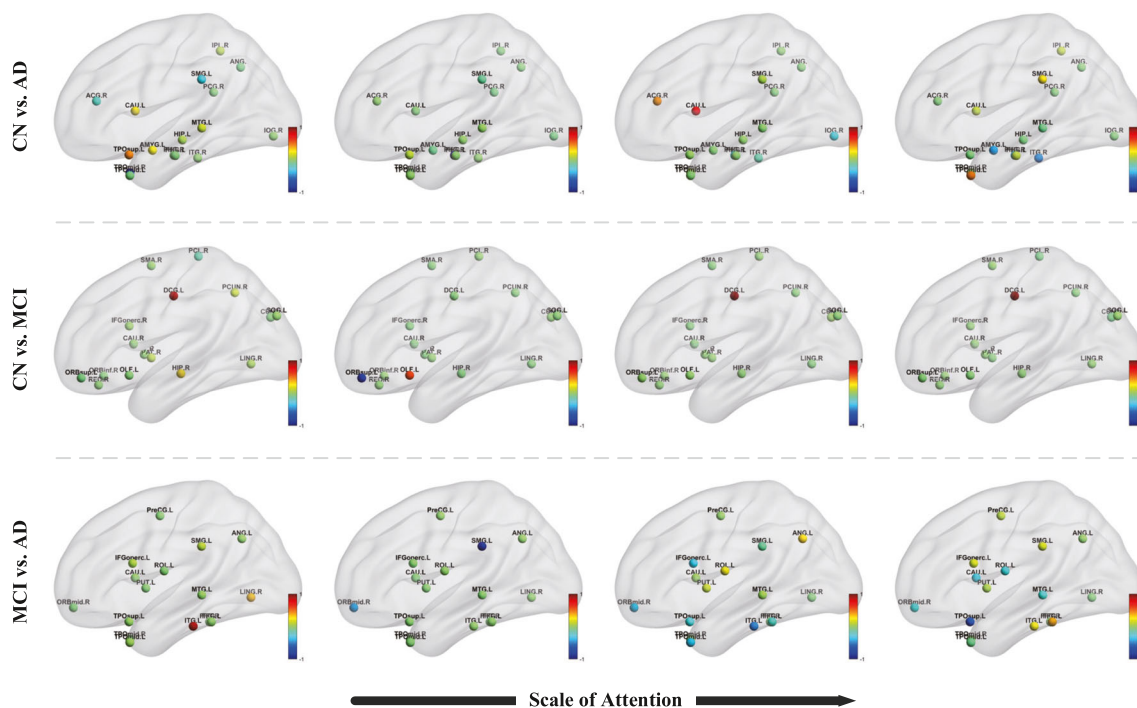


Fig. 9 Visualization of the direction vectors. Four direction vectors are optimized in each classification task. The four pictures from left to right correspond to SOAs from 1 to 4. Red indicates a value of 1, and blue indicates a value of -1 of the direction vector

Fig. 10 Visualization of the samples using three different feature extraction methods in three classification tasks

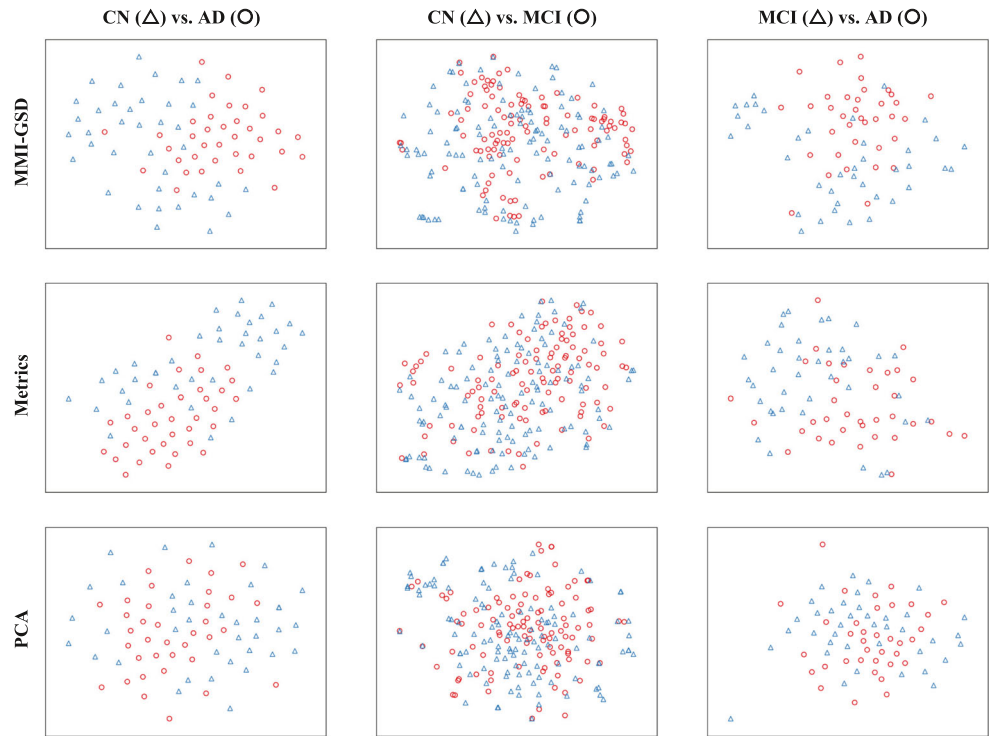
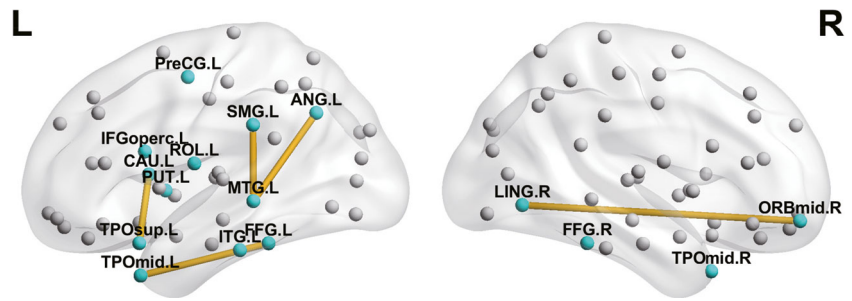


Table 3 Classification of CN vs. MCI vs. AD individuals

Group	Method	Acc. (%)	Sens. (%)	Spec. (%)
CN vs. AD (38/36)	<u>MMI-GSD</u>	77.03	72.22	81.58
	DRP-PCA	72.97	69.44	76.32
	PCA	62.16	58.33	65.79
	DRP-Metrics	71.62	63.89	78.95
	Metrics	63.51	63.89	63.16
	GCN [43]	66.22	69.44	63.16
	MMI-LinT [23]	72.97	80.56	65.79
	MMI-NonLinT [23]	70.27	61.11	78.95
CN vs. MCI (119/105)	<u>MMI-GSD</u>	63.39	56.19	69.75
	DRP-PCA	62.50	52.38	71.43
	PCA	54.46	46.67	61.34
	DRP-Metrics	60.71	49.52	70.59
	Metrics	48.21	36.19	58.82
	GCN	55.36	48.57	61.34
	MMI-LinT	62.95	54.29	70.59
	MMI-NonLinT	57.59	53.33	61.34
MCI vs. AD (39/36)	<u>MMI-GSD</u>	76.00	69.44	82.05
	DRP-PCA	70.67	72.22	69.23
	PCA	61.33	41.67	79.49
	DRP-Metrics	72.00	69.44	74.36
	Metrics	66.67	63.89	69.23
	GCN	62.67	61.11	64.10
	MMI-LinT	69.33	69.44	69.23
	MMI-NonLinT	65.33	52.78	76.92

Fig. 11 Discriminative connections between MCI and AD. The blue balls represent discriminative brain regions selected by a two-sample t-test, and the orange lines represent discriminative connections detected by the MMI-GSD



We compared the performances of the MMI-GSD and other feature extraction methods. As the most commonly used methods in brain network analysis, PCA and graph metrics are selected for comparison; PCA and graph metrics with dimensionality reduction preprocessing (DRP) are selected to validate the effectiveness of the proposed DRP. A GCN [43] was selected to compare the classification performance of a deep learning network and the proposed MMI-GSD. The MMI-LinT and MMI-NonLinT proposed in [23], two mutual information-based methods, were selected for comparison. As shown in Table 3, the MMI-GSD achieved accuracies of 77.03%, 63.39%, and 76.00% for the three classification tasks. We believe that the MMI-GSD achieved the highest classification performance because the edge weights in the graph have an inherent connection pattern, which is not considered in other methods.

In summary, our proposed MMI-GSD can be used to extract features from neuroimaging GSD and classify CN, MCI, and AD. The classification results show that the MMI-GSD considers the inherent network connections between

the data, resulting in better classification performance than comparable methods.

5 Discussion

We focus on the discussion of distinguishing AD from MCI because early detection is crucial for AD. We conducted dimensionality reduction using a t-test to reduce the size of the brain network. We obtained the brain regions with the most significant differences between the MCI and AD groups, as shown in Table 4. These abnormal brain regions are consistent with the results of previous studies, including inferior and middle frontal gyri [44–46], Rolandic operculum [47], parieto-temporal cortex [48], caudate nucleus [49], putamen [50], and parts of the temporal lobe [51].

The MMI-GSD with the smaller network identified important connectivity-related features to distinguish different groups of AD patients (see Table 5). In a previous study [52], the local nodal attributes in the left temporal lobe were significantly different between amnesic MCI converters and amnesic MCI non-converters. Our results agree with these findings, i.e., discriminative connections exist for the direction vector with an SOA of 1 (Fig. 11). For an SOA of 2, we found that the connection weights between the left supramarginal gyrus and the left fusiform gyrus were significantly different ($p = 0.0473 < 0.05$) between the two groups. The connection weight between two nodes with a distance of two hops is defined as the sum of the weights of all possible two-hop pathways. The

Table 4 Discriminative Brain Regions Between MCI and AD

Labels	Regions	Abbr.
1	Precentral gyrus	PreCG.L
10	Middle frontal gyrus, orbital part	ORBmid.R
11	Inferior frontal gyrus, opercular part	IFGoperc.L
17	Rolandic operculum	ROL.L
48	Lingual gyrus	LING.R
55	Fusiform gyrus	FFG.L
56	Fusiform gyrus	FFG.R
63	Supramarginal gyrus	SMG.L
65	Angular gyrus	ANG.L
71	Caudate nucleus	CAU.L
73	Putamen	PUT.L
83	Temporal pole: superior temporal gyrus	TPOsup.L
85	Middle temporal gyrus	MTG.L
87	Temporal pole: middle temporal gyrus	TPOmid.L
88	Temporal pole: middle temporal gyrus	TPOmid.R
89	Inferior temporal gyrus	ITG.L

Table 5 Discriminative Connections Between MCI and AD

	Regions 1	Regions 2	p-value
SOA=1	ORBmid.R	LING.R	0.0079
	TPOsup.L	CAU.L	0.0037
	MTG.L	SMG.L	0.0003
	MTG.L	ANG.L	0.0076
SOA=2	TPOmid.L	FFG.L	0.0024
	SMG.L	FFG.L	0.0473

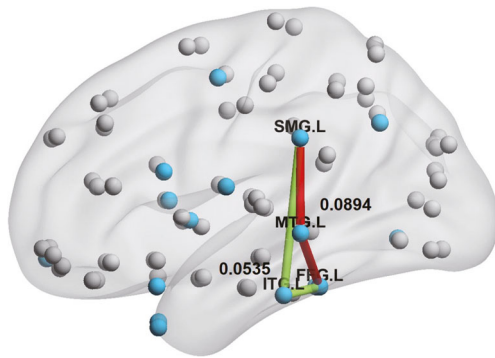


Fig. 12 Two white matter connection pathways with significant changes in AD group

weight of a two-hop pathway is defined as the product of its two components. Within all possible two-hop pathways between the left supramarginal gyrus and the left fusiform gyrus, we found two white matter connection pathways with significant differences, as shown in Fig. 12. One pathway is “SMG.L-MTG.L-FFG.L”, and the other one is “SMG.L-ITG.L-FFG.L”. The p-value of pathway “SMG.L-MTG.L-FFG.L” is 0.0894, which has a certain trend toward significance, and the p-value of pathway “SMG.L-ITG.L-FFG.L” is 0.0535, which is close to being statistically significant. We found that the connectivity of these two pathways was substantially worse in the AD individuals than the MCI individuals. The left supra-marginal gyrus is crucial for writing [53], and the left fusiform gyrus is required for visual word recognition [54]. The middle temporal gyrus and inferior temporal gyrus are involved in semantic memory processing. The impairment of these two pathways may lead to dysfunction in reading and writing in AD patients. Changes in these pathways also have the potential for biomarkers to diagnose AD patients.

6 Conclusion

The challenge in AD diagnosis is efficient feature extraction while preserving feature interpretability. We proposed a novel feature extraction method and an optimization framework based on mutual information to address this problem. The result of experiments with synthetic networks and AD classification showed that the proposed method achieved higher classification accuracy than comparable methods. The other advantage of our method is the high interpretability of the extracted features. The AD patients exhibited abnormal connections in the left hemisphere, especially in the left temporal lobe. Two white matter connection pathways had lower connectivity in the AD group, indicating reading and writing dysfunction in AD patients.

In future works, we intend to analyze whether the brain white matter changes found in this study can be used as a reliable biomarker for AD diagnosis. In addition, investigating the form of the proposed method in the nonlinear case is a way to improve performance.

Acknowledgements This work was supported by Beijing Advanced Innovation Center for Big Data-based Precision Medicine. Data collection and sharing for this project was funded by the Alzheimer’s Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer’s Association; Alzheimer’s Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer’s Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

References

1. Hebert LE, Weuve J, Scherr PA, Evans DA (2013) Alzheimer disease in the United States (2010–2050) estimated using the 2010 census. *Neurol.* 80(19):1778–1783. <https://doi.org/10.1212/WNL.0b013e31828726f5>
2. Alzheimer’s Association (2020) 2020 Alzheimer’s disease facts and figures. *Alzheimer’s & Dementia* 16(3):391–460. <https://doi.org/10.1002/alz.12068>
3. U.S. Department of Health and Human Services Centers for Disease Control and Prevention & National Center for Health Statistics (2020) CDC WONDER online database: About Underlying Cause of Death, 1999–2018. <https://wonder.cdc.gov/ucd-icd10.html>
4. Zhang Q et al (2018) Integrated proteomics and network analysis identifies protein hubs and network alterations in Alzheimer’s disease. *Acta Neuropathologica Communications* 6:19. <https://doi.org/10.1186/s40478-018-0524-2>
5. Dai Z et al (2015) Identifying and mapping connectivity patterns of brain network hubs in Alzheimer’s disease. *Cereb Cortex* 25:3723–3742. <https://doi.org/10.1093/cercor/bhu246>
6. Coninck JCP et al (2020) Network properties of healthy and Alzheimer brains. *Physica A Stat Mech Appl* 547:124475. <https://doi.org/10.1016/j.physa.2020.124475>
7. Bi X, Zhao X, Huang H, Chen D, Ma Y (2020) Functional brain network classification for Alzheimer’s disease detection with deep features and extreme learning machine. *Cogn Comput* 12:513–527. <https://doi.org/10.1007/s12559-019-09688-2>

8. Mheich A, Wendling F, Hassan M (2020) Brain network similarity: methods and applications. *Netw Neurosci* 4:507–527. <https://doi.org/10.1162/netn.a.00133>
9. Huang B et al (2021) Deep learning network for medical volume data segmentation based on multi axial plane fusion. *Comput Methods Prog Biomed* 212:106480. <https://doi.org/10.1016/j.cmpb.2021.106480>
10. Yamanakkanavar N, Choi JY, Lee B (2020) MRI segmentation and classification of human brain using deep learning for diagnosis of alzheimer's disease: a survey. *Sensors* 20:3243. <https://doi.org/10.3390/s20113243>
11. Huo Y et al (2019) 3D whole brain segmentation using spatially localized atlas network tiles. *NeuroImage* 194:105–119. <https://doi.org/10.1016/j.neuroimage.2019.03.041>
12. Li Y, Li H, Fan Y (2021) ACENet: Anatomical context-encoding network for neuroanatomy segmentation. *Med Image Anal* 70:101991. <https://doi.org/10.1016/j.media.2021.101991>
13. Magnin B et al (2009) Support vector machine-based classification of Alzheimer's disease from whole-brain anatomical MRI. *Neuroradiology* 51:73–83. <https://doi.org/10.1007/s00234-008-0463-x>
14. Khazaei A, Ebrahimpzadeh A, Babajani-Feremi A (2015) Identifying patients with Alzheimer's disease using resting-state fMRI and graph theory. *Clin Neurophysiol* 126:2132–2141. <https://doi.org/10.1016/j.clinph.2015.02.060>
15. Wee C-Y et al (2011) Enriched white matter connectivity networks for accurate identification of MCI patients. *NeuroImage* 54:1812–1822. <https://doi.org/10.1016/j.neuroimage.2010.10.026>
16. Prasad G, Joshi SH, Nir TM, Toga AW, Thompson PM (2015) Brain connectivity and novel network measures for Alzheimer's disease classification. *Neurobiol Aging* 36:S121–S131. <https://doi.org/10.1016/j.neurobiolaging.2014.04.037>
17. Hu C, He S, Wang Y (2021) A classification method to detect faults in a rotating machinery based on kernelled support tensor machine and multilinear principal component analysis. *Appl Intell* 51:2609–2621. <https://doi.org/10.1007/s10489-020-02011-9>
18. Li Z, Fan J, Ren Y, Tang L (2020) A novel feature extraction approach based on neighborhood rough set and PCA for migraine rs-fMRI. *J Intell Fuzz Syst* 38:5731–5741. <https://doi.org/10.3233/JIFS-179661>
19. Bilgen I, Guvercin G, Rekek I (2020) Machine learning methods for brain network classification: Application to autism diagnosis using cortical morphological networks. *J Neurosci Methods* 343:108799. <https://doi.org/10.1016/j.jneumeth.2020.108799>
20. Salvatore C et al (2015) Magnetic resonance imaging biomarkers for the early diagnosis of Alzheimer's disease: a machine learning approach. *Frontiers in Neuroscience* 9. <https://doi.org/10.3389/fnins.2015.00307>
21. Cover TM (1999) *Elements of information theory*. Wiley, New Jersey
22. Marinoni A, Gamba P (2017) Unsupervised data driven feature extraction by means of mutual information maximization. *IEEE Trans Comput Imaging* 3:243–253. <https://doi.org/10.1109/TCI.2017.2669731>
23. Özdenizci O, Erdoğan D (2020) Information theoretic feature transformation learning for brain interfaces. *IEEE Trans Biomed Eng* 67:69–78. <https://doi.org/10.1109/TBME.2019.2908099>
24. Hu C, Wang Y, Gu J (2020) Cross-domain intelligent fault classification of bearings based on tensor-aligned invariant subspace learning and two-dimensional convolutional neural networks. *Knowl-Based Syst* 209:106214. <https://doi.org/10.1016/j.knosys.2020.106214>
25. Liu M et al (2020) A multi-model deep convolutional neural network for automatic hippocampus segmentation and classification in Alzheimer's disease. *NeuroImage* 208:116459. <https://doi.org/10.1016/j.neuroimage.2019.116459>
26. Mehmood A, Maqsood M, Bashir M, Shuyuan Y (2020) A deep siamese convolution neural network for multi-Class classification of alzheimer disease. *Brain Sci* 10:84. <https://doi.org/10.3390/brainsci10020084>
27. Janghel RR, Rathore YK (2021) Deep convolution neural network based system for early diagnosis of alzheimer's disease. *IRBM* 42:258–267. <https://doi.org/10.1016/j.irbm.2020.06.006>
28. Bruna J, Zaremba W, Szlam A, LeCun Y (2014) Spectral networks and Locally connected networks on graphs. *International Conference on Learning Representations (ICLR2014)*. Banff, Canada
29. Defferrard M, Bresson X, Vandergheynst P (2016) Convolutional neural networks on graphs with fast localized spectral filtering. In: Lee DD, Sugiyama M, Luxburg UV, Guyon I, Garnett R (eds) *Advances in neural information processing systems*, vol 29. Curran Associates Inc, pp 3844–3852
30. Song X, Elazab A, Zhang Y (2020) Classification of mild cognitive impairment based on a combined high-Order network and graph convolutional network. *IEEE Access* 8:42816–42827. <https://doi.org/10.1109/ACCESS.2020.2974997>
31. Parisot S et al (2018) Disease prediction using graph convolutional networks: Application to Autism Spectrum Disorder and Alzheimer's disease. *Med Image Anal* 48:117–130. <https://doi.org/10.1016/j.media.2018.06.001>
32. Liu J et al (2021) MMHGE: detecting mild cognitive impairment based on multi-atlas multi-view hybrid graph convolutional networks and ensemble learning. *Clust Comput* 24:103–113. <https://doi.org/10.1007/s10586-020-03199-8>
33. Jack CR et al (2008) The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods. *J Magn Reson Imaging* 27:685–691. <https://doi.org/10.1002/jmri.21049>
34. Weiner MW et al (2017) The alzheimer's disease neuroimaging initiative 3: Continued innovation for clinical trial improvement. *Alzheimer's & Dementia* 13:561–571. <https://doi.org/10.1016/j.jalz.2016.10.006>
35. Cui Z, Zhong S, Xu P, Gong G, He Y (2013) PANDA: A pipeline toolbox for analyzing brain diffusion images. *Frontiers in Human Neuroscience* 7. <https://doi.org/10.3389/fnhum.2013.00042>
36. Tzourio-Mazoyer N et al (2002) Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-Subject brain. *NeuroImage* 15:273–289. <https://doi.org/10.1006/nimg.2001.0978>
37. Di C, Tm PNP, Ac E (1994) Automatic 3D intersubject registration of MR volumetric data in standardized Talairach space. *J Comput Assist Tomogr* 18:192–205
38. Dahl J, Vandenberghe L, Roychowdhury V (2008) Covariance selection for nonchordal graphs via chordal embedding. *Optim Methods Softw* 23:501–520. <https://doi.org/10.1080/10556780802102693>
39. Dempster AP (1972) Covariance selection. *Biometrics* 28:157–175. <https://doi.org/10.2307/2528966>
40. Hastie T, Tibshirani R, Friedman J (2009) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd edn. Springer Series in Statistics. Springer, New York
41. Eberhart R, Kennedy J (1995) A new optimizer using particle swarm theory 39–43
42. Rubinov M, Sporns O (2010) Complex network measures of brain connectivity: uses and interpretations. *Neuroimage* 52:1059–1069
43. Yang J et al (2020) Transfer learning from grid-structured data to graph-structured data: Application to diagnosis of depression, *Proceedings of the 30th European Safety and Reliability Conference and the 15th Probabilistic Safety Assessment and Management Conference*, 1373–1378 (Research Publishing, Singapore, Venice, Italy)

44. Bakkour A, Morris JC, Wolk DA, Dickerson BC (2013) The effects of aging and Alzheimer's disease on cerebral cortical anatomy: Specificity and differential relationships with cognition. *NeuroImage* 76:332–344. <https://doi.org/10.1016/j.neuroimage.2013.02.059>
45. Fjell AM et al (2009) High consistency of regional cortical thinning in aging across multiple samples. *Cereb Cortex* 19:2001–2012. <https://doi.org/10.1093/cercor/bhn232>
46. Cajanus A et al (2019) The Association Between Distinct Frontal Brain Volumes and Behavioral Symptoms in Mild Cognitive Impairment, Alzheimer's Disease, and Frontotemporal Dementia. *Frontiers in Neurology* 10. <https://doi.org/10.3389/fneur.2019.01059>
47. Zhang T et al (2019) Classification of Early and Late Mild Cognitive Impairment Using Functional Brain Network of Resting-State fMRI. *Frontiers in Psychiatry* 10. <https://doi.org/10.3389/fpsy.2019.00572>
48. Yang H et al (2019) Study of brain morphology change in Alzheimer's disease and amnesic mild cognitive impairment compared with normal controls. *Gen Psychiatr* 32:e100005. <https://doi.org/10.1136/gpsych-2018-100005>
49. Persson K et al (2018) Finding of increased caudate nucleus in patients with Alzheimer's disease. *Acta Neurol Scand* 137:224–232. <https://doi.org/10.1111/ane.12800>
50. Hamasaki H et al (2019) Tauopathy in basal ganglia involvement is exacerbated in a subset of patients with Alzheimer's disease: The Hisayama study. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring* 11:415–423. <https://doi.org/10.1016/j.dadm.2019.04.008>
51. Berron D, van Westen D, Ossenkoppele R, Strandberg O, Hansson O (2020) Medial temporal lobe connectivity and its associations with cognition in early Alzheimer's disease. *Brain* 143:1233–1248. <https://doi.org/10.1093/brain/awaa068>
52. Sun Y et al (2019) Prediction of Conversion From Amnesic Mild Cognitive Impairment to Alzheimer's Disease Based on the Brain Structural Connectome. *Frontiers in Neurology* 9. <https://doi.org/10.3389/fneur.2018.01178>
53. Penniello M.-J. et al (1995) A PET study of the functional neuroanatomy of writing impairment in Alzheimer's disease The role of the left supramarginal and left angular gyri. *Brain* 118:697–706. <https://doi.org/10.1093/brain/118.3.697>
54. Binder JR, Medler DA, Westbury CF, Liebenthal E, Buchanan L (2006) Tuning of the human left fusiform gyrus to sublexical orthographic structure. *NeuroImage* 33:739–748. <https://doi.org/10.1016/j.neuroimage.2006.06.053>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Jiawei Yang received the B.Eng. degree in electrical engineering and automation from Nanjing University of Aeronautics and Astronautics, China, in 2017. He is currently working toward the Ph.D. degree in the School of Automation Science and Electrical Engineering, Beihang University. His research interests include neuroimaging and brain network.



Shaoping Wang received the Ph.D., M.Eng. and B.Eng. degrees in mechatronics engineering from Beihang University, China, in 1994, 1991 and 1988. She has been with the School of Automation Science and Electrical Engineering at Beihang University since 1994 and promoted to the rank of professor in 2000. She was honored as a Changjiang Scholar Professor by the Ministry of Education of China in 2013. Her research interests include big data precision

medicine, pattern recognition, complex system reliability, and health management.



Teresa Wu is a Professor from the School of Computing, Informatics, Decision Systems Engineering at Arizona State University. She received her Ph.D. in Industrial Engineering from the University of Iowa in 2001. Her current research interests include: swarm intelligence, distributed decision support, health informatics. She is an IISE Fellow.

Affiliations

Jiawei Yang¹ · Shaoping Wang^{1,2} · Teresa Wu³

Jiawei Yang
yjwbuaa@buaa.edu.cn

Teresa Wu
Teresa.Wu@asu.edu

- ¹ School of Automation Science and Electrical Engineering, Beihang University, Beijing 100191, China
- ² Beijing Advanced Innovation Center for Big Data-Based Precision Medicine, Beihang University, Beijing 100191, China
- ³ School of Computing, Informatics, and Decision Systems Engineering, Arizona State University and ASU-Mayo Center for Innovative Imaging, Tempe AZ 85287, USA