Contents lists available at ScienceDirect

Information Fusion

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

Interpretable learning based Dynamic Graph Convolutional Networks for Alzheimer's Disease analysis

Yonghua Zhu [a,b,1], Junbo Ma [a,c,1], Changan Yuan [a,*], Xiaofeng Zhu [a,b,*]

[a] *Guangxi Academy of Sciences, Nanning, Gungxi, China*
[b] *Center for Future Media and School of Computer Science and Technology, University of Electronic Science and Technology of China, Chengdu, China*
[c] *Department of Psychiatry, University of North Carolina at Chapel Hill, Chapel Hill, USA*

ARTICLE INFO

Keywords:
Dynamic graph learning
Graph convolutional networks
Interpretable learning
Alzheimer's disease diagnosis

ABSTRACT

Graph Convolutional Networks (GCNs) are widely applied in classification tasks by aggregating the neighborhood information of each sample to output robust node embedding. However, conventional GCN methods do not update the graph during the training process so that their effectiveness is always influenced by the quality of the input graph. Moreover, previous GCN methods lack the interpretability to limit their real applications. In this paper, a novel personalized diagnosis technique is proposed for early Alzheimer's Disease (AD) diagnosis via coupling interpretable feature learning with dynamic graph learning into the GCN architecture. Specifically, the module of interpretable feature learning selects informative features to provide interpretability for disease diagnosis and abandons redundant features to capture inherent correlation of data points. The module of dynamic graph learning adjusts the neighborhood relationship of every data point to output robust node embedding as well as the correlations of all data points to refine the classifier. The GCN module outputs diagnosis results based on the learned inherent graph structure. All three modules are jointly optimized to perform reliable disease diagnosis at an individual level. Experiments demonstrate that our method outputs competitive diagnosis performance as well as provide interpretability for personalized disease diagnosis.

1. Introduction

Alzheimer's Disease (AD) is an irreversible neurodegenerative disease that significantly affects elders' daily activities in memory, recognition, behaviors, or even lives [1,2]. Millions of people worldwide suffer from AD, which has been one of the leading causes of death among elderly population [3,4]. Unfortunately, there is no cure for AD, so an early intervention with existing therapeutics will help delaying its deterioration [5,6]. Neuroimaging techniques like Magnetic Resonance Imaging (MRI) provide effective methods to monitor AD progression, and machine learning techniques are developed to facilitate the process of disease diagnosis on neuroimaging data [4,7,8].

Personalized diagnosis technique is an effective strategy to improve diagnosis performance under limited training data and heterogeneous testing data [3,9]. In medical fields, data collection is limited by many factors, such as subjects' privacy and costs of data collection, which make hard to collect a large amount of data to train a robust classifier for disease diagnosis [10,11]. Moreover, the heterogeneousness is common in AD pathology. This is because subjects are collected from different places, devices, or doctors, and all these factors make to-be-diagnosed data presents apparent heterogeneity [12–14], and thus conventional classification techniques easily output suboptimal diagnosis results for these data points [2,7]. Therefore, personalized diagnosis methods are designed to improve performance of disease diagnosis in aforementioned complex scenarios. The key to conducting personalized diagnosis is to refine classifiers via considering the distribution information of both diagnosed and undiagnosed patients [8,15].

In recent years, deep learning draws wide attention in personalized disease diagnosis, generally outputting better diagnosis performance than conventional machine learning methods. For example, Farooq et al. demonstrated that deep learning techniques can output highly discriminative feature representation for AD classification [16]. However, most deep learning methods do not consider the structure information of the data which has been proved effective to improve classification performance in conventional machine learning [17–19]. Graph Convolutional Networks (GCNs) use the distribution information of data points to conduct semi-supervised classification and have more discriminative ability than other deep learning methods for disease diagnosis [3]. However, current GCN models need to address the following issues. First, the performance of GCN highly depends on

[*] Corresponding authors.
E-mail addresses: yuanchangan@126.com (C. Yuan), seanzhuxf@gmail.com (X. Zhu).
[1] Yonghua Zhu and Junbo Ma contributed equally to this work.

https://doi.org/10.1016/j.inffus.2021.07.013
Received 29 October 2020; Received in revised form 1 May 2021; Accepted 25 July 2021
Available online 31 July 2021

the quality of the graph which keeps unchanged during the training process. When a low-quality graph is input, classification performance gets affected. To address this issue, Zhu et al. designed a GCN model to exploit the importance of interactions between neighbor nodes [20]. Second, using a fixed graph cannot dynamically consider the data distribution of the training data and new unseen testing data, and thus fails to flexibly refine classifiers for specific testing data. As a result, the fixed graph in GCN easily outputs suboptimal performance in personalized diagnosis. Both two factors could result in suboptimal or biased diagnosis results on heterogeneous testing data. Third, similar to other deep learning methods, GCN cannot output interpretable diagnosis results [21–23].

Graph learning is a widely used strategy to address the first two issues by exploiting the distribution of the data [2,24]. On one hand, original data usually contains noise and redundancy so that outputting the low-quality graph [25]. On the other hand, in most cases, although a graph is well predefined, graph methods could not output reasonable results, because this graph is constructed independently from the classification tasks [11,26]. Therefore, how to dynamically adjust graph structure is crucial to improve performance of disease diagnosis. Gan et al. proposed an effective personalized diagnosis method by simultaneously using dynamic graph learning and semi-supervised learning [3], where dynamic graph structure can capture effective correlation of brain regions. Compared with a fixed graph, dynamic graph learning can adjust connectivity or edges weights in a learning manner [27], thus being able to output more flexible and effective graphs for specific learning tasks. However, previous personalized diagnosis methods (*e.g.,* [3,8,27]) use separated classifiers to implement disease diagnosis instead of using an end-to-end manner, thus having limited improvement.

In real medical applications, interpretable diagnosis results are important in pathology research and case study by providing reasonable explanations for disease diagnosis [28]. Bron et al. proposed using a feature selection method before SVM to improve AD classification performance as well as search for Region-Of-Interests (ROIs) to provide more biological interpretability [29]. To conduct interpretability for AD, three principal approaches including voxel-level methods, patch-level methods, and region-level methods, are employed to extract features from medical data like MRI [30]. Comparing with other two approaches, region-level methods get more focus because it generates a lower-dimensional and higher-level representation for brain regions. For example, Liu et al. proposed using a feature selection method to select the most discriminative brain regions for exploiting correlations between brain regions [31]. However, most of these methods (*e.g.,* [31–33]) recognize important features in a shallow linear regression manner, and thus have limited discriminative ability [30]. In addition, these conventional methods separate selecting important features from performing classification. As a result, the feature selection aims at keeping the information as much as possible, rather than achieving minimal classification errors, and thus eventually outputting suboptimal diagnosis results.

In this paper, an Interpretable Dynamic Graph Convolutional Networks (IDGCN) is proposed to improve the performance of personalized diagnosis for AD and output interpretable results. To do this, interpretable feature leaning and dynamic graph leaning are embedded into a GCN architecture. More specifically, the interpretable feature learning provides interpretability for diagnosis results and a pre-classification makes selected features be classification-oriented. In addition, the dynamic graph learning dynamically updates the graph structure for GCN to output superior diagnosis results by adjusting similar and dissimilar correlation of all objects. Thus, by jointly optimizing feature learning, graph learning, and the GCN, the proposed disease diagnosis method cannot only produce reliable personalized diagnosis but also provide interpretability for diagnosis results. In our experiments, we have employed six data sets of ADNI to validate effectiveness of the proposed personalized diagnosis method. The experimental results show that our method can output competitive diagnosis performance with comparison to state-of-the-art classification methods, as well as interpret AD diagnosis results from aspects of brain regions.

Compared to previous methods, our proposed method integrates the graph learning with the interpretability in the GCN framework, the contributions of our proposed method can be summarized as follows.

- It is very popular for coupling the graph learning with the learning tasks in a framework for conventional machine learning methods [11,25]. A few literature of deep learning models focused on this. For example, Jiang et al. proposed simultaneously conducting graph learning and representation learning in a framework [34]. All of them ignore the interpretability, which is very useful for real applications, especially for medical image analysis.
- A number of conventional machine learning methods (*e.g.,* [3, 32]) focused on separately conducting feature selection and classification to achieve the interpretability, while many deep learning models are difficult to achieve the interpretability. On the contrary, our method simultaneously considers feature selection and the graph learning in the GCN model.

2. Method

Throughout the paper, we use boldface uppercase letters, boldface lowercase letters, and normal italic letters, respectively, to denote matrices, vectors, and scalars. Specifically, $\mathbf{X} \in \mathbb{R}^{n\times d}$ denotes a data matrix containing n d-dimensional data points. $\mathbf{x}^i$ and $\mathbf{x}_j$ respectively stands for the ith row and the jth column of a matrix $\mathbf{X}$. Besides, a graph structure can be represented as $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ and $\mathcal{E}$, respectively, stand for the set of vertexes and edges of a graph, which usually can be represented by an adjacency matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$.

In this section, we provide a detailed introduction to our Interpretable Dynamic Graph Convolutional Networks (IDGCN) for AD diagnosis, which involves three modules, *i.e.,* the module of interpretable feature learning, the module of dynamic graph learning, and the GCN module, shown in Fig. 1.

2.1. Dynamic GCN

In traditional machine learning, classifiers trained with the training data are used to predict labels of the testing data, under the assumption that both the training data and the testing data have the same distribution. However, this assumption is not always hold in real applications. First, the training data is often heterogeneous to the testing data in medical data analysis. Second, conventional classifiers are constructed for all testing data while ignores their individual diversity, *i.e.,* different subjects generally have different information or characteristics from others, and thus personalized classifier is preferred [32], where a classifier is constructed for each testing data by considering the consistency between training set and the testing set. In the literature, the transductive semi-supervised learning, *e.g.,* GCN, which can construct the classifier based on both the training data and the testing data for the individual subject, is becoming a popular method for personalized diagnosis [35].

Graph structure is widely used to represent the correlation among objects [4,21]. With computer-aid disease diagnosis methods, graph structure is widely used to analyze the correlation between patients or symptoms. For example, brain functional connectivity analysis on fMRI can help understand neurological disorders in brain region level [22]. By considering the correlation between diagnosed patients and undiagnosed patients, classification models can usually output more prominent diagnosis performance, as similar patients usually present similar properties. In this paper, we employ GCN to conduct disease diagnosis, which involves two steps, *i.e.,* the graph construction and the GCN.

Constructing a graph matrix that can represent the correlation of objects usually can be formulated as: $a_{i,j} = exp(\frac{\|\mathbf{x}^i - \mathbf{x}^j\|^2}{-2\theta^2})$, where θ is

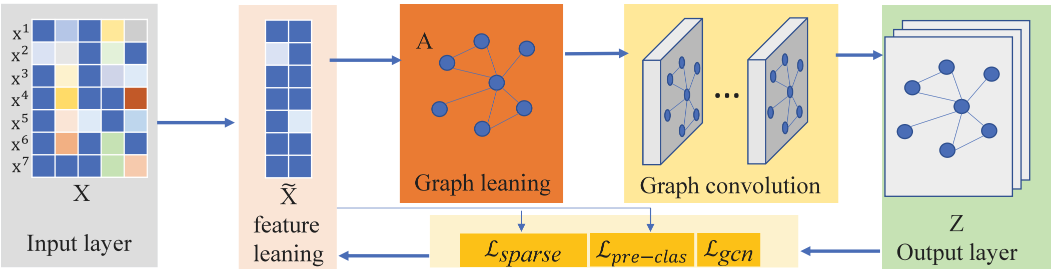


Fig. 1. The architecture of the proposed IDGCN model. In specific, the module of interpretable feature learning aims at finding the most important features to provide interpretability and relieving the impact of redundant features to capture the intrinsic correlation of data points. The module of dynamic graph learning automatically learns the graph structure for training a robust GCN model by adjusting the correlation of the training data and the testing data. The GCN module uses the learned graph structure to output personalized diagnosis. All these three modules are jointly optimized to output personalized diagnosis in an end-to-end learning manner.

length scale for the input space and the adjacency matrix $\mathbf{A} = \{a_{i,j}\}_{i,j=1}^{n}$ is the matrix form of graph structure. Given with an initial graph $\mathbf{A}$, the problem of disease diagnosis under the GCN framework [36] can be converted to the problem of node classification, *i.e.,*

$$\mathbf{H}^{(l+1)} = \sigma(\widetilde{\mathbf{D}}^{-\frac{1}{2}}\widetilde{\mathbf{A}}\widetilde{\mathbf{D}}^{-\frac{1}{2}}\mathbf{H}^{(l)}\mathbf{W}_{g}^{(l)}) \tag{1}$$

where $\mathbf{H}^{(l)} \in \mathbb{R}^{n\times d_l}$ is a d_l-dimensional feature representation in the lth layer and $\widetilde{\mathbf{A}}$ and $\widetilde{\mathbf{D}}$ are respectively the adjacency matrix and its degree matrix.

The objective function of the GCN is minimizing the following loss function:

$$\mathcal{L}_{gcn} = -\sum_{i\in L}\sum_{j=1}^{c} y_{i,j} ln z_{i,j} \tag{2}$$

where $\mathbf{Z} \in \mathbb{R}^{n\times c}$ is the output of the GCN.

GCN is widely used in semi-supervised classification on graph-structured data, aiming at predicting labels of unlabeled nodes of the graph via aggregating information of neighbors. However, it has at least two problems to be addressed for conducting the disease diagnosis of AD, *i.e.,* the inflexibility of a fixed graph and the lack of interpretability.

The graph structure is a critical input of the GCN, and its quality directly links to classification performance [34]. In most cases, this graph is constructed artificially based on prior knowledge, or using the popular k-nearest-neighbor (kNN) strategies. However, two drawbacks possibly result in a suboptimal graph for the GCN. First, this graph is constructed by primitive data usually containing redundancy. Even though two nodes are close to each other in terms of distance, they could be not inherently similar to each other, because distribution of the data can be easily distorted by redundant features [25]. Second, this graph is predefined independently of learning tasks, which easily output suboptimal results. To solve these two problems, the dynamic graph learning is often used for jointing the interpretable feature leaning module and the GCN module. Specifically, a regularization of graph learning is added to the GCN, *i.e.,* given the initial graph $\mathbf{A} \in \mathbb{R}^{n\times n}$, the regularization term $\mathcal{L}_{GL}$ is defined as:

$$\mathcal{L}_{gl} : \min_{\mathbf{A}} \sum_{i,j=1}^{n} \|\mathbf{x}^i - \mathbf{x}^j\|_F^2 a_{ij} + \lambda_1\|\mathbf{A}\|_F^2 \tag{3}$$
$$s.t., \forall i, \mathbf{A}_i\mathbf{1} = 1, a_{ij} \geq 0.$$

where $\mathbf{A} = \{a_{ij}\}_{i,j=1}^{n} \in \mathbb{R}^{n\times n}$ is the graph and $\|\cdot\|_F$ indicates the Frobenius norm. To reduce the complexity, a weight vector $\alpha \in \mathbb{R}^{1\times d}$ is introduced to dynamically polish this graph structure in Eq. (3) [34]:

$$\hat{a}_{ij} = \frac{a_{ij}exp(ReLU(|\mathbf{x}^i\mathbf{W} - \mathbf{x}^j\mathbf{W}|\alpha^T))}{\sum_{j=1}^{n} a_{ij}exp(ReLU(\mathbf{x}^i\mathbf{W} - \mathbf{x}^j\mathbf{W}|\alpha^T))} \tag{4}$$

where $\hat{a}_{ij}$ is the prediction of a_{ij}. Finally, the objective function of the dynamic GCN is:

$$\mathcal{L}_{dgcn} = \mathcal{L}_{gcn} + \mathcal{L}_{gl}. \tag{5}$$

Both graph $\mathbf{A}$ and the new representation $\mathbf{Z}$ are updated iteratively by minimizing Eq. (5). As a result, if $\mathbf{A}$ is low-quality, the dynamic GCN still outputs robust classifier.

2.2. *Interpretability*

Through processing MRI or Positron Emission Tomography (PET) images via region-level methods, multiple Region-Of-Interest (ROI) can be divided, and consequently, some important information like gray matter tissue volumes can be extracted for each ROI. Thus, a feature representation can be obtained for each subject, in which each feature element corresponds to a specific brain region. To provide interpretability for AD diagnosis, we can select the most informative brain region to discriminate patients. Therefore, this problem of interpretable learning can be converted to the problem of feature selection [8]. The key to conduct feature selection is to learn a weight vector of features. Hence, important features, *i.e.,* brain regions, can be preserved to provide interpretability for disease diagnosis [37,38].

In real applications, original data usually is high-dimensional in which most features are redundant or irrelevant to learning tasks. Therefore, it is difficult to capture inherent correlation of the data for a specific task. To solve this problem, the interpretable feature learning is first utilized to conduct feature selection by preserving the most informative and discriminative features, which is a simple but effective way to explore the inherent correlation as well as provide interpretability for classifiers.

Generally, sparse learning conducts feature selection via enforcing weights corresponding to irrelevant features to be zero, so that important features can be recognized [39]. The unsupervised sparse learning based feature selection method is to minimize the following loss function as:

$$\mathcal{L}_{sparse} = \|\mathbf{X} - \mathbf{X}\mathbf{W}\|_F^2 + \lambda\|\mathbf{W}\|_0 \tag{6}$$

where λ is a tuning parameter to balance the magnitude of two terms, *i.e.,* data reconstruction and sparse learning. $\mathbf{W} \in \mathbb{R}^{d\times d}$ is a trainable weight matrix, and $\mathbf{X} \in \mathbb{R}^{n\times d}$ is a matrix representation of data points. The $\|\mathbf{W}\|_0$ is the ℓ_0-norm regularization of the weight matrix $\mathbf{W}$, being used to obtain sparsity. Specifically, if the ith feature is a redundant feature, the corresponding weight of the feature (*i.e.,* $\mathbf{w}_i$) is zero. Under this sparse constraint, a sparse weight matrix (*i.e.,* $\mathbf{W}$) is outputted, which enables to store as much information as original data and removes redundant features, *i.e.,* through feature selection. It is worth noting that the ℓ_0-norm constraint is hard to be optimized [40]. To address this issue, in this paper, we follow the literature [40] to convert the optimization of the ℓ_0-norm regularization to its approximate version, *i.e.,* the optimization of the $\ell_{2,1}$-norm regularization. As a result, we have:

$$\mathcal{L}_{sparse} = \|\mathbf{X} - \mathbf{X}\mathbf{W}\|_F^2 + \lambda\|\mathbf{W}\|_{2,1} \tag{7}$$

where $\|\mathbf{W}\|_{2,1} = \sum_{i=1}\sqrt{\sum_{j=1}|w_{ij}^2|}$, which leads to the row sparsity in $\mathbf{W}$. Moreover, we employ the forward-background splitting method [41] to optimize Eq. (7), *i.e.,*

$$\mathbf{W} \leftarrow Prox21_{\lambda}(\mathbf{W} - \eta\nabla_{\mathbf{W}}\mathcal{L}_{sparse}) \tag{8}$$

Algorithm 1: The pseudo of our IDGCN method.

```
Input: $\mathbf{X} \in \mathbb{R}^{n\times d}$, label information$\mathbf{Y}_L$, and hyper-parameters $\lambda$, $\lambda_1$ and $\lambda_2$;
Output: $\mathbf{W} \in \mathbb{R}^{d\times d}$, predictions of the unlabeled data;
1. Constructing graph matrix $\mathbf{A}$ of $\mathbf{X}$;
2. Initialize model weights $\mathbf{W}$, $\mathbf{W}_g^l$ and $\alpha$;
3. while $epoch < 5000$ do
4.     $\widetilde{\mathbf{X}} \leftarrow \{\mathbf{X}, \mathbf{W}\}$ by Eq. (7);
5.     $\mathbf{A} \leftarrow \{\widetilde{\mathbf{X}}, \alpha\}$ by Eq. (4);
6.     $\mathbf{Z} \leftarrow \{\mathbf{X}, \mathbf{A}, \mathbf{W}_g^l\}$ by Eq. (1);
7.     $\mathcal{L} \leftarrow \{\mathcal{L}_{dgcn}, \mathcal{L}_{sparse}, \mathcal{L}_{pre-clas}\}$ by Eq. (12);
8.     Back-propagate $\mathcal{L}$ to update model weights;
9.     Epoch + = 1;
end while
```

where η is the learning ratio in Eq. (7). Moreover, we have the following equation:

$$Prox21_{\lambda} = \arg\min_{\mathbf{P}} \frac{1}{2}\|\mathbf{P}-\mathbf{O}\|_F^2 + \lambda\|\mathbf{P}\|_{2,1} \tag{9}$$

Based on the literature [40], we have the closed solution of Eq. (9) as follows:

$$\mathbf{p}^{i*} = \begin{cases} \frac{\|\mathbf{o}^i\|_2-\lambda}{\|\mathbf{o}^i\|_2}\mathbf{o}^i, if\ \lambda < \|\mathbf{o}^i\|_2 \\ 0, \quad otherwise \end{cases} \tag{10}$$

where $\mathbf{p}^{i*}$ and $\mathbf{o}^i$, respectively, denote the ith row of the matrix $\mathbf{P}$ and $\mathbf{O}$.

2.3. Loss function

Our model first learns a weight matrix to evaluate the importance of each feature. Being forced with a sparse constraint, the weight matrix enables us to use partial important features to preserve the information as much as possible. However, preserved features possibly do not cater for classification tasks, and thus result in suboptimal diagnosis results so that being unable to provide interpretability. Therefore, classification-oriented features are actually needed. Therefore, a learning task is vital to evaluate new feature representation, *i.e.,* by adding a pre-classification. It is noteworthy that we did not employed the GCN to implement pre-classification because the graph structure has not been updated and the complexity of the GCN is high. Alternatively, a single fully connected layer with the activation function of $softmax(\cdot)$ is used to pre-predict labels on the representation (*i.e.,* $\widetilde{\mathbf{X}} = \mathbf{X}\mathbf{W}$). To do this, we evaluate the cross-entropy error of pre-classification by minimizing the following loss function:

$$\mathcal{L}_{pre-clas} = -\sum_{i\in L}\sum_{j=1}^{c} y_{i,j} ln z_{i,j}^{p} \tag{11}$$

where $\mathbf{Z}^p = softmax(\widetilde{\mathbf{X}}\mathbf{W}^p)$ and $\mathbf{W}^p$ is a trainable weight matrix in the single-layer of pre-classification, $\mathbf{Y}$ is the ground truth, L is the set of labeled nodes, and c is the number of classes. By introducing such a supervision in the training process, it can more accurately recognize important features especially for classification tasks. Finally, the proposed interpretable dynamic GCN is to minimize the following loss function:

$$\mathcal{L} = \mathcal{L}_{dgcn} + \lambda_1\mathcal{L}_{sparse} + \lambda_2\mathcal{L}_{pre-clas} \tag{12}$$

where λ_1 and λ_2 are two non-negative hyper-parameters to make a trade-off for three terms. Moreover, the pseudo of our proposed method is summarized in Algorithm 1.

Table 1
The brief information of six AD data sets.

Data sets	#(Samples)	#(Features)	#(Classes)
AD-NC	51:52	90	2
AD-MCI	51:99	90	2
NC-MCI	52:99	90	2
MCIn-MCIp	44:56	90	2
ADNI-3cla	51:52:99	90	3
ADNI-4cla	51:52:44:56	90	4

3. Experiment

3.1. Experimental setting

3.1.1. Data sets

Raw digital imaging data downloaded from ADNI database,[2] had 202 subjects, which included 51 AD patients, 99 Mild Cognitive Impairment (MCI) patients, and 52 normal controls (NC). Moreover, 99 Mild Cognitive Impairment (MCI) patients included 43 MCI converters (MCIp) and 56 MCI non-converters (MCIn).

We followed the literature [2] to process these raw MRI images. Specifically, we sequentially applied spatial distortion correction, skull-stripping, and cerebellum removal on these images and then segmented the MRI images into gray matter, white matter, and cerebrospinal fluid, followed by warping them into the Automated Anatomical Labeling (AAL) template [42] to obtain 90 regions. Therefore, we extracted 90 ROIs for each subject and each ROI was represented by one feature. We finally obtained 90 gray matter volumes from an MRI image and used them as the original feature matrix $\mathbf{X}$. We further combined these subjects to form six subsets which include four binary data sets (*i.e.,* AD vs. NC, AD vs. MCI, NC vs. MCI, MCIn vs. MCIp), one three-class data set (*i.e.,* AD vs. NC vs. MCI) and one four-class data set (*i.e.,* AD vs. NC vs. MCIn vs. MCIp). The basic information of all data sets is summarized in Table 1.

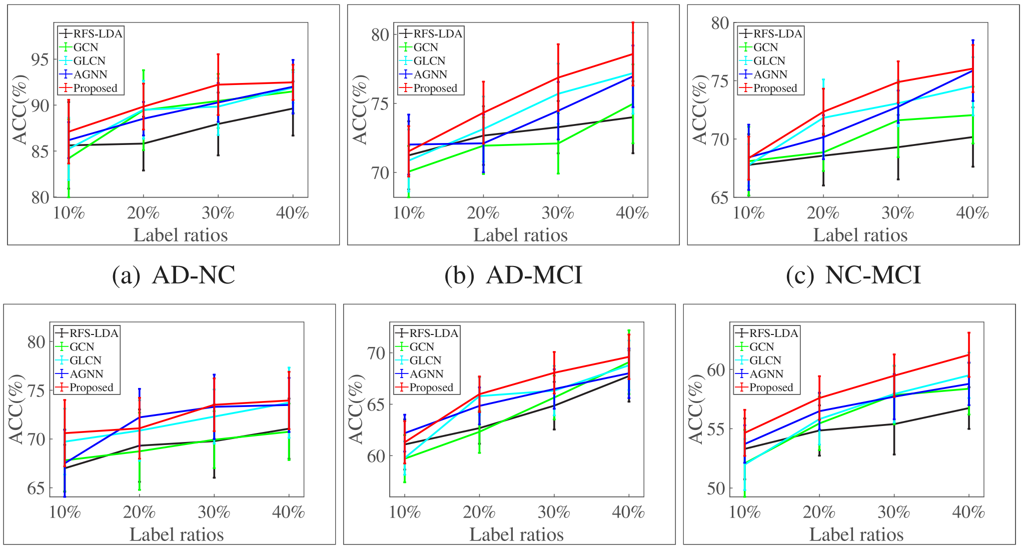
3.1.2. Comparison methods

The comparison methods included four semi-supervised learning methods and we listed their details as follows.

- Robust Feature-Sample Linear Discriminant Analysis (**RFS-LDA**) [32] uses both the labeled training data and the unlabeled testing data to detect the sample outliers and feature noise.
- Graph Convolutional Networks (**GCN**) [36] encodes the graph structure and the node representation to conduct the layer-wise propagation.
- Graph Learning-Convolutional Networks (**GLCN**) [34] integrates the graph learning and the graph convolution to dynamically learn the optimal graph structure.
- Attention-based Graph Neural Network (**AGNN**) [35] replaces the intermediate fully-connected layers of the GCN model with the propagation layers under attention mechanisms, to learn a dynamic and adaptive local summary of the neighborhood.

RFS-LDA is a non-graph method while other comparisons as well as our method are graph methods. Among graph methods, GCN employs a fixed graph structure, while the others dynamically refine the graph structure during the training process. GLCN constructs a dynamic graph in the first layer relying on the graph Laplacian regularization, and AGNN adaptively updates the graph structure by computing the attention for the neighborhood of nodes. Our proposed method adaptively updates the graph structure, and more importantly, also focuses on interpretability, *i.e.,* via selecting important features to construct an intrinsic graph.

[2] http://www.loni.usc.edu/ADNI.



(a) AD-NC (b) AD-MCI (c) NC-MCI

(d) MCIn-MCIp (e) ADNI-3cla (f) ADNI-4cla

Fig. 2. The classification accuracy (ACC) and the corresponding STandard Deviation (STD) of all methods with different label ratios on all data sets.

3.1.3. Setting

In our experiments, the random splits strategy was used to conduct experiments for all methods. In specific, we randomly selected $[10\%, 20\%, 30\%, 40\%]$ of labeled subjects for training, 50% of the remainder for validation, and the other 50% for testing which is the counterpart for classification tasks. The random splits were repeated 20 times to reduce random errors, and the average of 20 results was set as the final result for each experiment. ADNI data sets do not provide the graph structure that presents the correlation of subjects, so we constructed it by the kNN strategy, where k is empirically set as $\sqrt{n}$ according to [43], and this graph was used as the initial graph for all graph-based methods. We adjusted hyper-parameters for each method by referring to the corresponding literature to output their best results. For our method, we set the maximum epochs as 5000, the learning rate as 0.001, and 0.005 for the graph learning layer and the GCN module, respectively. The training process stops if the validation loss does not decrease for 50 consecutive epochs. We also set $\lambda, \lambda_1 \in [10^{-3}, 10^{-2}, \dots, 10^{3}]$ and $\lambda_2 \in [0.25, 0.5, \dots, 2]$ for model selection. To conduct interpretable learning, we set $\|\mathbf{w}^i\|_2$ (where $\|\mathbf{w}^i\|_2 = \sqrt{\sum_j w_{i,j}^2}$) as the importance of the ith brain regions according to [44]. By sorting weights of each region (*i.e.*, feature), the top important regions was selected to interpret diagnosis results.

3.2. Personalized diagnosis

In this section, we conducted personalized diagnosis on all data sets with different settings of label ratios, and summarized the results in Fig. 2.

Our method achieved the best diagnosis performance, followed by GLCN, AGNN, GCN, and RFS-LDA. Compared to the best comparison method (*i.e.*, GLCN) and the worst comparison method (*i.e.*, RFS-LDA), our method on average improved by 1.16%, and 4.10%, on all six data sets at different label ratios. Moreover, we conducted the paired-sample t-tests at 95% significance level between our method and each comparison method in Fig. 2. Experimental results show that our method has statistically significant difference from each comparison method, *i.e.*, $p \leq 0.05$. The reasons could be that interpretable feature learning and dynamic graph leaning polish the graph structure for GCN to promote diagnosis performance. Compared to AGNN which uses an attention mechanism to aggregate the neighborhood information, our method on average improved by 1.42% on all six data sets. The reason could be that feature redundancy defers the method from finding inherent distribution of nodes as well as their neighbor. In contrast, our method uses an interpretable learning to refine important information, thus is more robust to redundancy. The advantages of dynamic graph learning can be testified from observation that two dynamic graph based methods (*i.e.*, GLCN and our method) improved by 2.01% than GCN on all data sets, which means dynamic graph learning outputs good graph structure to promote the classification performance of GCN. By embedding graph learning to adjust consistency between the training set and the testing set, dynamic graph methods refine classifiers for the testing set and output good personalized diagnosis results. Moreover, our method jointly optimizes the feature learning, the graph learning and the GCN, as a result, all three parts achieve their optimal status to output superior classification performance.

Our method slightly outperformed GLCN, even though two methods use similar dynamic graph learning mechanism in Eq. (5), *i.e.*, learning the new feature representation for the graph construction. Differently, our method adds a pre-classification (*i.e.*, Eq. (11)) to regulate new representation, while GLCN uses a graph Laplacian regularization. It means the new representation of our method is classification-oriented, while GLCN is graph-oriented. Therefore, the new feature representation of our method is easier to output better diagnosis performance, compared to the GLCN.

All graph-based methods outperformed the non-graph method, *i.e.*, RFS-LDA, which does not consider the correlation of subjects. For example, the worst graph-based comparison method (*i.e.*, GCN) has an average improvement of 0.54% than RFS-LDA. Actually, considering the correlation of subjects is important to improve diagnosis performance, because patients with the same disease usually have similar symptoms or features. Therefore, making use of the correlation through a graph structure could make more accurate diagnosis.

3.3. Interpretability analysis

To evaluate the interpretability of our method, we compared our proposed method with two feature selection methods, *i.e.*, Robust Feature-Sample Linear Discriminant Analysis (RFS-LDA) [32] and ANalysis Of VAriance (ANOVA) [45] in term of classification performance on all six data sets. More specifically, we used three feature selection methods to select 15 most important features (*i.e.*, top 15 features) and then reported the classification results in Fig. 4. As a result, our proposed method outperformed all comparison methods. For example, our method improved about 8% on data set AD-MCI, compared to ANOVA,

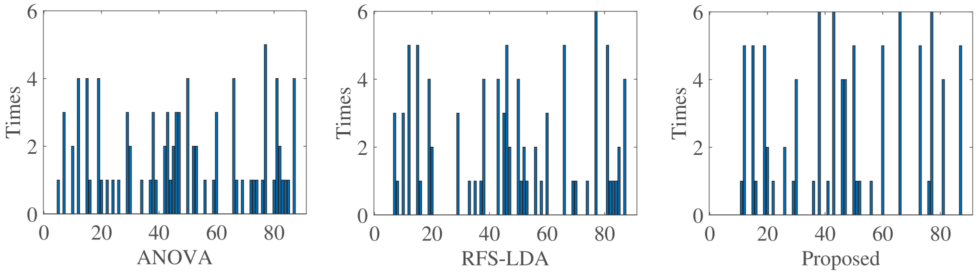


Fig. 3. The frequency of the features selected by three methods across six data sets.

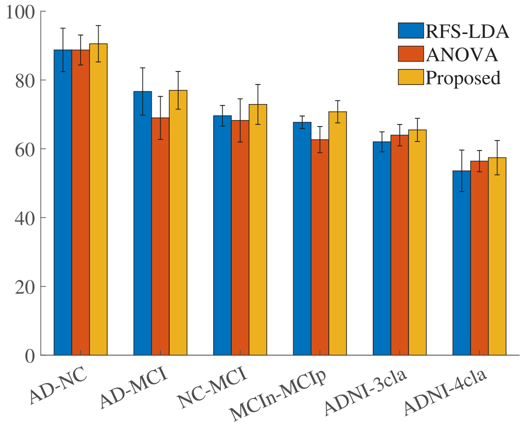


Fig. 4. The classification accuracy of three feature selection methods.

in terms of classification accuracy. This indicates the effectiveness of our proposed method in terms of interpretability.

In our experiments, we repeated the process of feature selection 20 times to obtain top 15 brain regions of all methods. Moreover, we visualized the selected brain regions of our method in Fig. 5, summarized the frequency of top 15 features of all methods in Fig. 3 and further listed the names of top 15 brain regions in Table 2.

From Fig. 5, the top 15 brain regions selected by our method have been shown relative to AD [31,46]. In particular, some brains regions were consistently selected across six data sets, such as the regions of middle temporal gyrus right, hippocampal formation left, precuneus left, and uncus left, which has been shown in [47].

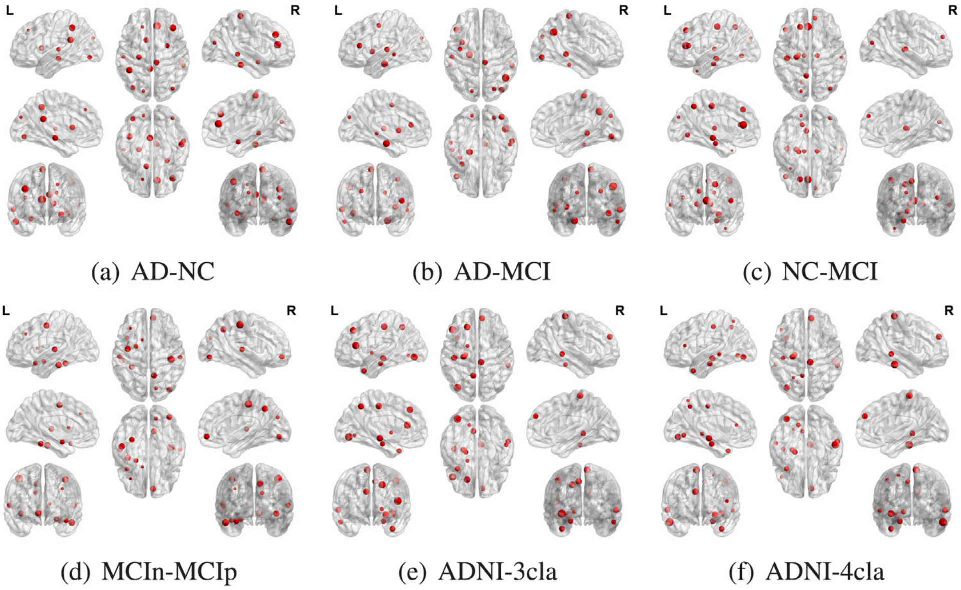
Based on Fig. 3, ANOVA and RFS-LDA, respectively, selected 44 and 36 regions across six data sets while our method only selected 27 regions. This implies that our method is more stable than the comparison methods. However, ANOVA and RFS-LDA selected some brain regions unrelated to AD [46], *e.g.,* postcentral gyrus. This verifies the effectiveness of our method again.

3.4. Ablation analysis

In this section, we analyze the effectiveness of each part in our method. To do this, we decompose our proposed objective function in Eq. (12) to form the comparison methods, *i.e.,* IDGCNnP, IDGCNnI, and IDGCNwL2. Compared to Eq. (12), IDGCNnP removes the $\mathcal{L}_{pre-clas}$ for investigating the effectiveness of the pre-classification, while IDGCNnI does not have the $\mathcal{L}_{sparse}$ for investigating the effectiveness of the reconstruction of the original $\mathbf{X}$. IDGCNwL2 replaces the $\ell_{2,1}$-norm with ℓ_2-norm in Eq. (7), aiming at investigating the interpretability. Besides, compared to our method, GCN without considering any above parts is also included in this section.

All experiments used 40% of labeled data in the experiments, and experimental results are shown in Fig. 6.

First, three methods considering partial components of our method, *i.e.,* IDGCNnP, IDGCNnI, and IDGCNwL2, outperformed GCN without considering any component in our method. This indicates that each part in our method is reasonable as well as effective. Second, IDGCNwL2 outperformed either IDGCNnP or IDGCNnI. The reason



(a) AD-NC
(b) AD-MCI
(c) NC-MCI
(d) MCIn-MCIp
(e) ADNI-3cla
(f) ADNI-4cla

Fig. 5. Top 15 brain regions selected by our method.

Table 2
Top 15 brain regions selected by three feature selection methods.

ANOVA	RFS-LDA	Proposed
Middle temporal gyrus right	Middle temporal gyrus right	Precuneus left
Inferior frontal gyrus left	Inferior frontal gyrus left	Uncus left
Parahippocampal gyrus left	Parahippocampal gyrus left	Hippocampal formation left
Nucleus accumbens right	Lingual gyrus left	Middle temporal gyrus right
Postcentral gyrus left	Hippocampal formation left	Inferior frontal gyrus left
Hippocampal formation left	Inferior temporal gyrus right	Parahippocampal gyrus left
Inferior temporal gyrus right	Nucleus accumbens right	Nucleus accumbens right
Lateral occipitotemporal gyrus left	Precuneus left	Postcentral gyrus left
Lateral ventricle right	Uncus left	Temporal pole left
Superior occipital gyrus right	Postcentral gyrus left	Amygdala left
Precuneus left	Lateral occipitotemporal gyrus left	Lateral occipitotemporal gyrus left
Uncus left	Medial frontal gyrus left	Caudate nucleus left
Lingual gyrus left	Globus palladus left	Lingual gyrus left
Superior frontal gyrus left	Superior occipital gyrus right	Superior frontal gyrus left
Temporal pole left	Middle temporal gyrus left	Inferior temporal gyrus right

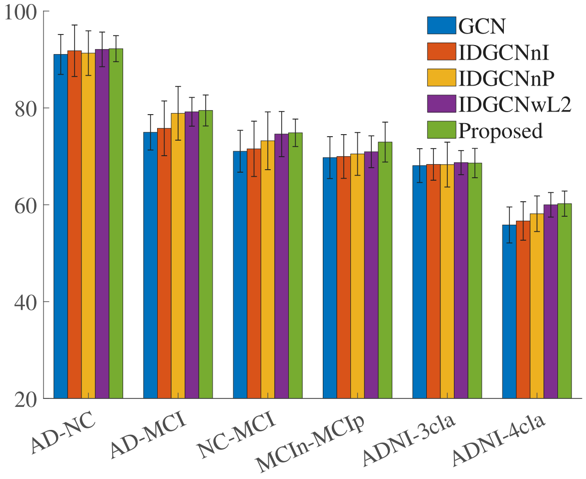


Fig. 6. The classification accuracy (ACC) and the corresponding STandard Deviation (STD) of ablation analysis on six data sets.

is that IDGCNwL2 considers both the feature learning and the pre-classification, while either IDGCNnP or IDGCNnI only considers one of them. Moreover, IDGCNwL2 is worse than our method, and this verifies the importance of the interpretability.

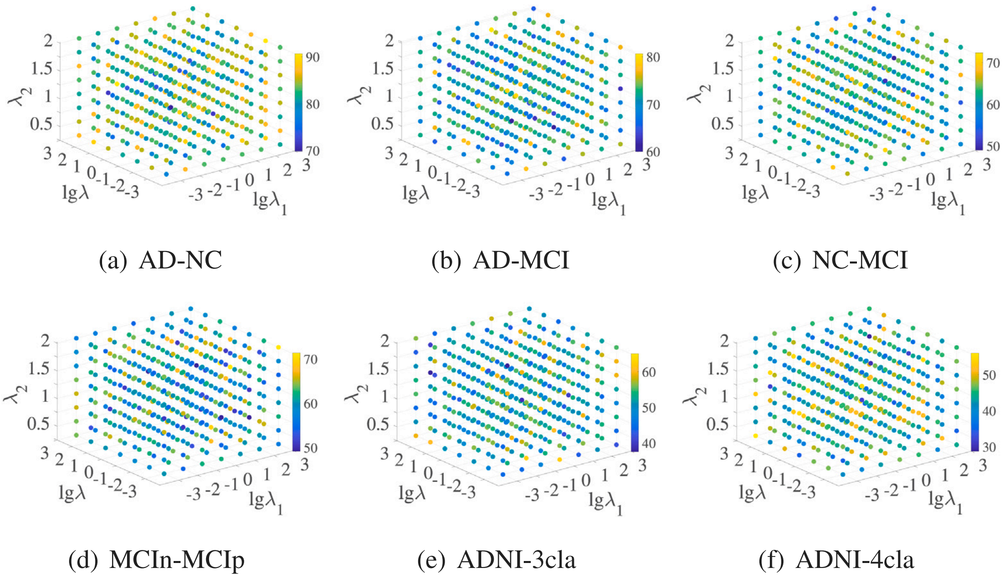
3.4.1. *Parameters' sensitivity analysis*

Fig. 7 shows the variations of the classification accuracy with different settings of three hyper-parameters of our objective function in Eq. (12), *i.e.*, λ, λ_1, and λ_2.

Obviously, our method is sensitive to parameters' setting as the classification accuracy on some data sets has the fluctuation of 20%. Moreover, we cannot figure out which parameter play a more important role as all three modules of our model jointly devote to the classification task. λ_2 is relatively less sensitive to final classification, compared to the others, as it is set to control the performance of pre-classification instead of final classification.

4. Conclusion

This paper proposed a new GCN method to exploit the interpretability of GCN model. To do this, the proposed method integrates feature selection with graph learning in the GCN model, in a semi-supervised learning manner, and then applied the proposed method for AD diagnosis. Experimental results demonstrated the effectiveness of our proposed method, compared to state-of-the-art semi-supervised methods, in terms of classification tasks.



(a) AD-NC (b) AD-MCI (c) NC-MCI

(d) MCIn-MCIp (e) ADNI-3cla (f) ADNI-4cla

Fig. 7. The classification accuracy of the proposed method with different parameters' setting.

Since incomplete data is very common in medical field, we plan to extend our proposed method to conduct personalize diagnosis on incomplete data in our future work.

CRediT authorship contribution statement

Yonghua Zhu: Conceptualization, Methodology, Software, Writing – original draft, Experiment. **Junbo Ma:** Conceptualization, Software, Validation, Methodology, Experiment. **Changan Yuan:** Supervision, Project administration, Writing – review & editing, Instruction. **Xiaofeng Zhu:** Supervision, Project administration, Writing – review & editing, Instruction.

Declaration of competing interest

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

Acknowledgments

This work was partially supported by the National Key Research and Development Program of China (Grant No. 2018AAA0102200); the National Natural Science Foundation of China (Grant No: 61876046); the Guangxi “Bagui” Teams for Innovation and Research, China; the Marsden Fund of New Zealand (MAU1721); and the Sichuan Province Science and Technology Support Program, China (Grants No: 2018GZDZX0032 and 2019YFG0535).

References

[1] Naomi Habib, Cristin McCabe, Sedi Medina, Miriam Varshavsky, Daniel Kitsberg, Raz Dvir-Szternfeld, Gilad Green, Danielle Dionne, Lan Nguyen, Jamie L. Marshall, et al., Disease-associated astrocytes in Alzheimer’s disease and aging, Nature Neurosci. 23 (6) (2020) 701–706.

[2] Heng Tao Shen, Xiaofeng Zhu, Zheng Zhang, Shui-Hua Wang, Yi Chen, Xing Xu, Jie Shao, Heterogeneous data fusion for predicting mild cognitive impairment conversion, Inf. Fusion 66 (2021) 54–63.

[3] Jiangzhang Gan, Ziwen Peng, Xiaofeng Zhu, Rongyao Hu, Junbo Ma, Guorong Wu, Brain functional connectivity analysis based on multi-graph fusion, Med. Image Anal. (2021) http://dx.doi.org/10.1016/j.media.2021.102057.

[4] Xiaofeng Zhu, Bin Song, Feng Shi, Yanbo Chen, Rongyao Hu, Jiangzhang Gan, Wenhai Zhang, Man Li, Liye Wang, Yaozong Gao, et al., Joint prediction and time estimation of COVID-19 developing severe symptoms using chest CT scan, Med. Image Anal. 67 (2021) 101824.

[5] Jay Penney, William T. Ralvenius, Li-Huei Tsai, Modeling Alzheimer’s disease with iPSC-derived brain cells, Mol. Psychiatry 25 (1) (2020) 148–167.

[6] Aliza P. Wingo, Yue Liu, Ekaterina S. Gerasimov, Jake Gockley, Benjamin A. Logsdon, Duc M. Duong, Eric B. Dammer, Chloe Robins, Thomas G. Beach, Eric M. Reiman, et al., Integrating human brain proteomes with genome-wide association data implicates new proteins in Alzheimer’s disease pathogenesis, Nature Genet. 53 (2) (2021) 143–146.

[7] Kenji Tagai, Maiko Ono, Manabu Kubota, Soichiro Kitamura, Keisuke Takahata, Chie Seki, Yuhei Takado, Hitoshi Shinotoh, Yasunori Sano, Yasuharu Yamamoto, et al., High-contrast in vivo imaging of tau pathologies in Alzheimer’s and non-Alzheimer’s disease tauopathies, Neuron 109 (1) (2021) 42–58.

[8] Yingying Zhu, Minjeong Kim, Xiaofeng Zhu, Jin Yan, Daniel Kaufer, Guorong Wu, Personalized diagnosis for Alzheimer’s disease, in: MICCAI, 2017, pp. 205–213.

[9] Daniel Giovinazzo, Biljana Bursac, Juan I. Sbodio, Sumedha Nalluru, Thibaut Vignane, Adele M. Snowman, Lauren M. Albacarys, Thomas W. Sedlak, Roberta Torregrossa, Matthew Whiteman, et al., Hydrogen sulfide is neuroprotective in Alzheimer’s disease by sulfhydrating GSK3β and inhibiting tau hyperphosphorylation, Proc. Natl. Acad. Sci. 118 (4) (2021).

[10] Mengxing Huang, Huirui Han, Hao Wang, Lefei Li, Yu Zhang, Uzair Aslam Bhatti, A clinical decision support framework for heterogeneous data sources, IEEE J. Biomed. Health Inf. 22 (6) (2018) 1824–1833.

[11] Xiaofeng Zhu, Jianye Yang, Chengyuan Zhang, Shichao Zhang, Efficient utilization of missing data in cost-sensitive learning, IEEE Trans. Knowl. Data Eng. (2019) http://dx.doi.org/10.1109/TKDE.2019.2956530.

[12] Liyuan Fan, Chengyuan Mao, Xinchao Hu, Shuo Zhang, Zhihua Yang, Zhengwei Hu, Huifang Sun, Yu Fan, Yali Dong, Jing Yang, et al., New insights into the pathogenesis of Alzheimer’s disease, Front. Neurol. 10 (2020) 1312.

[13] M. Tanveer, B. Richhariya, R.U. Khan, A.H. Rashid, P. Khanna, M. Prasad, C.T. Lin, Machine learning techniques for the diagnosis of Alzheimer’s disease: A review, ACM Trans. Multimedia Comput. Commun. Appl. (TOMM) 16 (1s) (2020) 1–35.

[14] Petrice M. Cogswell, Heather J. Wiste, Matthew L. Senjem, Jeffrey L. Gunter, Stephen D. Weigand, Christopher G. Schwarz, Arvin Arani, Terry M. Therneau, Val J. Lowe, David S. Knopman, et al., Associations of quantitative susceptibility mapping with Alzheimer’s disease clinical and imaging markers, Neuroimage 224 (2021) 117433.

[15] Juan Fortea, Eduard Vilaplana, Maria Carmona-Iragui, Bessy Benejam, Laura Videla, Isabel Barroeta, Susana Fernández, Miren Altuna, Jordi Pegueroles, Víctor Montal, et al., Clinical and biomarker changes of Alzheimer’s disease in adults with down syndrome: a cross-sectional study, Lancet 395 (10242) (2020) 1988–1997.

[16] Ammarah Farooq, SyedMuhammad Anwar, Muhammad Awais, Saad Rehman, A deep CNN based multi-class classification of Alzheimer’s disease using MRI, in: IST, 2017, pp. 1–6.

[17] Seyed Hani Hojjati, Ata Ebrahimzadeh, Ali Khazaee, Abbas Babajani-Feremi, Alzheimer’s Disease Neuroimaging Initiative and others, Predicting conversion from MCI to AD using resting-state fMRI, graph theoretical approach and SVM, J. Neurosci. Methods 282 (2017) 69–80.

[18] Rongyao Hu, Xiaofeng Zhu, Yonghua Zhu, Jiangzhang Gan, Robust SVM with adaptive graph learning, World Wide Web 23 (2020) 1945–1968.

[19] Xiaofeng Zhu, Jiangzhang Gan, Guangquan Lu, Jiaye Li, Shichao Zhang, Spectral clustering via half-quadratic optimization, World Wide Web 23 (2020) 1969–1988.

[20] Hongmin Zhu, Fuli Feng, Xiangnan He, Xiang Wang, Yan Li, Kai Zheng, Yongdong Zhang, Bilinear graph neural network with neighbor interactions, in: IJCAI, Vol. 5, 2020.

[21] Yuxiao Liu, Ning Zhang, Dan Wu, Audun Botterud, Rui Yao, Chongqing Kang, Guiding cascading failure search with interpretable graph convolutional network, 2020, arXiv preprint arXiv:2001.11553.

[22] Hai Shu, Xiao Wang, Hongtu Zhu, D-CCA: A decomposition-based canonical correlation analysis for high-dimensional datasets, J. Amer. Statist. Assoc. 115 (529) (2020) 292–306.

[23] Jonathan Graff-Radford, Keir X.X. Yong, Liana G. Apostolova, Femke H. Bouwman, Maria Carrillo, Bradford C. Dickerson, Gil D. Rabinovici, Jonathan M. Schott, David T. Jones, Melissa E. Murray, New insights into atypical Alzheimer’s disease in the era of biomarkers, Lancet Neurol. 20 (3) (2021) 222–234.

[24] Tong Tong, Katherine Gray, Qinquan Gao, Liang Chen, Daniel Rueckert, Alzheimer’s Disease Neuroimaging Initiative and others, Multi-modal classification of Alzheimer’s disease using nonlinear graph fusion, Pattern Recognit. 63 (2017) 171–181.

[25] Xiaofeng Zhu, Shichao Zhang, Yonghua Zhu, Pengfei Zhu, Yue Gao, Unsupervised spectral feature selection with dynamic hyper-graph learning, IEEE Trans. Knowl. Data Eng. (2020) http://dx.doi.org/10.1109/TKDE.2020.3017250.

[26] Xuelong Li, Han Zhang, Rui Zhang, Yun Liu, Feiping Nie, Generalized uncorrelated regression with adaptive graph for unsupervised feature selection, IEEE Trans. Neural Netw. Learn. Syst. 30 (5) (2018) 1587–1595.

[27] Jiaxin Wu, Sheng-hua Zhong, Yan Liu, Dynamic graph convolutional network for multi-video summarization, Pattern Recognit. 107 (2020) 107382.

[28] Shangran Qiu, Prajakta S Joshi, Matthew I. Miller, Chonghua Xue, Xiao Zhou, Cody Karjadi, Gary H. Chang, Anant S. Joshi, Brigid Dwyer, Shuhan Zhu, et al., Development and validation of an interpretable deep learning framework for Alzheimer’s disease classification, Brain (2020).

[29] E.E. Bron, M. Smits, W.J. Niessen, S. Klein, Feature selection based on the SVM weight vector for classification of dementia, IEEE J. Biomed. Health Inf. 19 (5) (2015) 1617–1626.

[30] Chunfeng Lian, Mingxia Liu, Jun Zhang, Dinggang Shen, Hierarchical fully convolutional network for joint atrophy localization and Alzheimer’s disease diagnosis using structural MRI, IEEE Trans. Pattern Anal. Mach. Intell. 42 (4) (2020) 880–893.

[31] Feng Liu, Chong-Yaw Wee, Huafu Chen, Dinggang Shen, Inter-modality relationship constrained multi-modality multi-task feature selection for Alzheimer’s disease and mild cognitive impairment identification, NeuroImage 84 (2014) 466–475.

[32] Ehsan Adeli, Kim-Han Thung, Le An, Guorong Wu, Feng Shi, Tao Wang, Dinggang Shen, Semi-supervised discriminative classification robust to sample-outliers and feature-noises, IEEE Trans. Pattern Anal. Mach. Intell. 41 (2) (2019) 515–522.

[33] Tong Tong, Robin Wolz, Qinquan Gao, Ricardo Guerrero, Joseph V. Hajnal, Daniel Rueckert, Multiple instance learning for classification of dementia in brain MRI, Med. Image Anal. 18 (5) (2014) 808–818.

[34] Bo Jiang, Ziyan Zhang, Doudou Lin, Jin Tang, Bin Luo, Semi-supervised learning with graph learning-convolutional networks, in: CVPR, 2019, pp. 11313–11320.

[35] Kiran K Thekumparampil, Chong Wang, Sewoong Oh, Li-Jia Li, Attention-based graph neural network for semi-supervised learning, 2018, arXiv preprint arXiv:1803.03735.

[36] Thomas N. Kipf, Max Welling, Semi-supervised classification with graph convolutional networks, in: ICLR, 2017.

[37] Xiaofei He, Deng Cai, Partha Niyogi, Laplacian score for feature selection, in: NIPS, 2006, pp. 507–514.

[38] Heng Tao Shen, Yonghua Zhu, Wei Zheng, Xiaofeng Zhu, Half-quadratic minimization for unsupervised feature selection on incomplete data, IEEE Trans. Neural Netw. Learn. Syst. (2020) http://dx.doi.org/10.1109/TNNLS.2020.3009632.

[39] Saúl Solorio-Fernández, J. Ariel Carrasco-Ochoa, José Fco Martínez-Trinidad, A review of unsupervised feature selection methods, Artif. Intell. Rev. 53 (2) (2020) 907–948.

[40] Xiao Cai, Feiping Nie, Heng Huang, Exact top-k feature selection via l2,0-norm constraint, in: IJCAI, 2013, pp. 1240–1246.

[41] Patrick L. Combettes, Jean-Christophe Pesquet, Proximal splitting methods in signal processing, in: Fixed-Point Algorithms for Inverse Problems in Science and Engineering, 2011, pp. 185–212.

[42] Nathalie Tzourio-Mazoyer, Brigitte Landeau, Dimitri Papathanassiou, Fabrice Crivello, Olivier Etard, Nicolas Delcroix, Bernard Mazoyer, Marc Joliot, Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain, Neuroimage 15 (1) (2002) 273–289.

[43] Upmanu Lall, Ashish Sharma, A nearest neighbor bootstrap for resampling hydrologic time series, Water Resour. Res. 32 (3) (1996) 679–693.

[44] Yi Yang, Heng Tao Shen, Zhigang Ma, Zi Huang, Xiaofang Zhou, $L_{2,1}$-norm regularized discriminative feature selection for unsupervised learning, in: IJCAI, 2011, pp. 1589–1594.

[45] Hui Ding, Peng-Mian Feng, Wei Chen, Hao Lin, Identification of bacteriophage virion proteins by the ANOVA feature selection and analysis, Mol. Biosyst. 10 (8) (2014) 2229–2235.

[46] M. Flint Beal, Michael F. Mazurek, Vinh T. Tran, Geetinder Chattha, Edward D. Bird, Joseph B. Martin, Reduced numbers of somatostatin receptors in the cerebral cortex in Alzheimer's disease, Science 229 (4710) (1985) 289–291.

[47] Xiaofeng Zhu, Heung-Il Suk, Seong-Whan Lee, Dinggang Shen, Subspace regularized sparse multitask learning for multiclass neurodegenerative disease identification, IEEE Trans. Biomed. Eng. 63 (3) (2015) 607–618.