Cost-Sensitive Weighted Contrastive Learning Based on Graph Convolutional Networks for Imbalanced Alzheimer's Disease Staging

Yan Hu, Jun Wang, *Senior Member, IEEE*, Hao Zhu, Juncheng Li, and Jun Shi, *Member, IEEE*

***Abstract*—Identifying the progression stages of Alzheimer's disease (AD) can be considered as an imbalanced multi-class classification problem in machine learning. It is challenging due to the class imbalance issue and the heterogeneity of the disease. Recently, graph convolutional networks (GCNs) have been successfully applied in AD classification. However, these works did not handle the class imbalance issue in classification. Besides, they ignore the heterogeneity of the disease. To this end, we propose a novel cost-sensitive weighted contrastive learning method based on graph convolutional networks (CSWCL-GCNs) for imbalanced AD staging using resting-state functional magnetic resonance imaging (rs-fMRI). The proposed method is developed on a multi-view graph constructed using the functional connectivity (FC) and high-order functional connectivity (HOFC) features of the subjects. A novel cost-sensitive weighted contrastive learning procedure is proposed to capture discriminative information from the minority classes, encouraging the samples in the minority class to provide adequate supervision. Considering the heterogeneity of the disease, the weights of the negative pairs are introduced into contrastive learning and they are computed based on the distance to class prototypes, which are automatically learned from the training data. Meanwhile, the cost-sensitive mechanism is further introduced into contrastive learning to handle the class imbalance issue. The proposed CSWCL-GCN is evaluated on 720 subjects (including 184 NCs, 40 SMC patients, 208 EMCI patients, 172 LMCI patients and 116 AD patients) from the ADNI (Alzheimer's Disease Neuroimaging Initiative). Experimental results show that the proposed CSWCL-GCN outperforms state-of-the-art methods on the ADNI database.**

***Index Terms*—Alzheimer's Disease Staging, cost-sensitive weighted contrastive learning, imbalanced classification, rs-fMRI**

I. INTRODUCTION

ALZHEIMER'S disease (AD) is a neurodegenerative disease characterized by decreased cognitive functions such as memory, language, and attention [1]. It progresses slowly and gradually and, eventually, it damages most areas of the brain. According to the severity of the disease, the AD progression always includes four stages, i.e. significant memory concern (SMC), early mild cognitive impairment (EMCI), late mild cognitive impairment (LMCI), and AD. Since the subjects in the adjacent AD stages present similar clinical symptoms, accurately identifying AD progression stages for a subject is challenging.

The resting-state functional magnetic resonance imaging (rs-fMRI) can capture slight functional changes in the brain by measuring the blood oxygenation level-dependent (BOLD) signals as a neurophysiological index of neural activity [2]. Based on rs-fMRIs, both functional connectivity (FC) and high-order functional connectivity (HOFC) [3] can be extracted as effective biomarkers to identify AD stages. Identifying AD stages based on rs-fMRIs can be considered as a multi-class classification problem in machine learning. Although many works have been developed for AD classification [4]–[7], most of them are limited to binary classification that differentiates AD patients from normal controls, without simultaneously considering multiple stages in AD progression.

One challenge in AD staging, i.e. multi-class AD classification, is that the class distribution in the training data is skewed, i.e. the numbers of subjects at different AD stages are quite different. For example, in ADNI dataset [8], only 5.9% subjects belong to the SMC class and more than 31.5% subjects belong to the EMCI class. Since most AD classification methods were developed without considering the potential of class imbalance, directly using them on multi-class AD classification would undermine the learned representations in minority classes.

Recently, graph neural networks (GNNs) have been proposed for AD classification based on rs-fMRIs [9]–[12]. It provides an intuitive way of modelling subjects and associations or similarities between them. In GNNs, a node is represented by a specific feature representation of a subject, and the edge weights are used to capture the similarities between each pair of nodes. Different from traditional AD classification methods that only used the features of individual patients,

This work was supported in part by the National Natural Science Foundation of China under Grant 62272289, in part by Shanghai Municipal Science and Technology Commission under Grant 20ZR1419900, and in part by 111 Project under Grant D20031

Corresponding author: J. Wang (wangjun_shu@shu.edu.cn)

Y. Hu, J. Wang, H. Zhu, J. Li and J. Shi are with the Key Laboratory of Specialty Fiber Optics and Optical Access Networks, Joint International Research Laboratory of Specialty Fiber Optics and Advanced Communication, Shanghai Institute for Advanced Communication and Data Science, School of Communication and Information Engineering, Shanghai University, China.

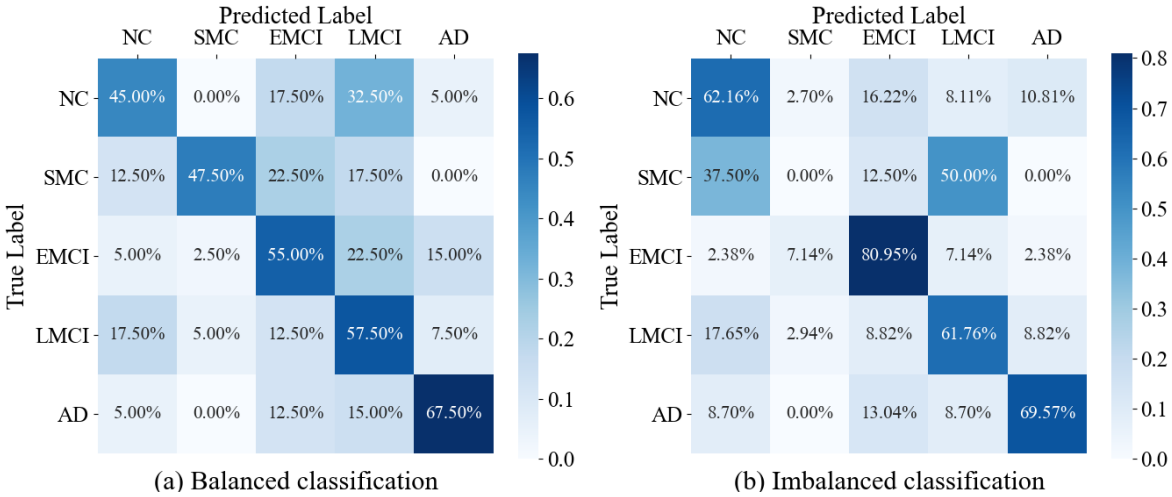

Fig.1. The confusion matrices of the results on classifying ADNI data using GCN. The color of the matrix elements indicates the classification accuracy. (a) Confusion matrix of the classification results of GCN on balanced ADNI data. (b) Confusion matrix of the classification results of GCN on imbalanced ADNI data, respectively.

GNNs consider the inter-subject correlation. However, previous GNNs-based studies assumed the training data were perfectly or nearly balanced across multiple classes [13]–[15]. When they are applied on imbalanced data, their performance degrades. Fig.1 shows an example that the classical graph convolutional network (GCN) performed multi-class classification on both balanced and imbalanced data generated from ADNI. The confusion matrices in both subfigures give the classification results of GCN on balanced data and imbalanced data, respectively. From both figures, one may observe that the diagonal elements in the confusion matrix of Fig. 1(a) are large, indicating that most samples were correctly classified by GCN on the balanced data. However, the performance of GCN on imbalanced data is significantly worse compared to its performance on the balanced data. This is because the training loss is dominated by the majority classes when GCN is trained on the imbalanced data, and the model is thus updated towards behaving significantly better on majority classes than minority ones. Therefore, the network tends to overfit to the samples in majority classes, whereas failed to differentiate the samples in the minority classes such as SMC.

To solve the aforementioned problem, several GNNs-based methods for imbalanced classification have been developed [16]–[23]. These works handled the data imbalance issue by data augmentation via under(over)-sampling [16]–[19], or weight assignment to adjust the portion of different classes in training loss [20]–[23]. However, these methods suffered the problems of over-fitting or discarding valuable information. Furthermore, they ignored to mine more knowledge in the minority classes, making the model fail to leverage adequate prior knowledge in minority classes.

Contrastive learning holds great promise for learning discriminative representations from insufficient training data [24]. It successfully mines the discriminative knowledge from the insufficient data by maximizing the agreement of samples in the positive pairs against that in the negative pairs [25], [26]. Recent studies have investigated contrastive learning on GCNs. Typical works include Graph Contrastive Learning (GraphCL), Graph Contrastive Coding (GCC) and Deep Graph Infomax (DGI) [27]–[29]. These methods aimed to learn effective representations from the overall training data. However, they cannot be directly applied to the task of identifying AD stages using rs-fMRIs. This is because AD is a progressive neurodegenerative disease with heterogeneity in diverse phenotypes. The patients even in the same AD stage may demonstrate different variations in rs-fMRIs [30], [31]. The FC and HOFC features extracted from the rs-fMRIs of the patients thus carry different amount of knowledge for classification. This will be an obstacle for existing contrastive learning methods to capture effective representations for classification.

To address the above issues, we propose a cost-sensitive weighted contrastive learning method based on GCN (CSWCL-GCN), a new learning framework that simultaneously considers both class imbalance and heterogeneity in multi-class AD classification. The proposed method is developed based on a dual-channel GCN structure whose input is a multi-view population graph. Considering the heterogeneity of the disease and different amount of information carried by each subject, the weights of the negative pairs are introduced into contrastive learning and they are computed based on the distance to class prototypes, which are automatically learned from the training data. Meanwhile, the cost-sensitive mechanism is also introduced into contrastive learning to balance the contribution of subjects in different classes. In summary, the main contributions of this paper can be concluded as follows:

1) We propose a novel CSWCL-GCN for identifying AD stages. Different from most AD classification methods that only consider binary classification settings, our method leverages multi-class classification that directly differentiates a subject from five classes of NC, SMC, EMCI, LMCI, and AD.

2) We propose a novel cost-sensitive weighted contrastive learning (CSWCL) procedure for imbalanced classification. Both issues of limited supervision caused by insufficient training samples in the minority classes and data heterogeneity are considered in the proposed CSWCL procedure.

3) We propose an automatic learning procedure to identify the class prototypes in CSWCL, where the weights for the negative pairs are adaptively evaluated based on the distance of the samples to the class prototypes.

The rest of the paper is organized as follows: In Section 2, the related works on imbalanced learning are reviewed. In Section 3, the details of the proposed CSWCL-GCN are presented. In Section 4, we report both ablation study and comparative study with experimental results. Finally, we draw the conclusion in Section 5.

II. Related Works

Imbalanced data is ubiquitous in many real-world applications. Traditional methods for handling imbalanced data include data-level approaches, algorithm-level approaches and hybrid approaches [32]. The data-level approaches try to rebalance the class distributions using a data pre-processing step including over-sampling minority classes or under-sampling majority classes. Representative works of data-level approaches include Synthetic Minority Over-sampling Technique (SMOTE) and Random Under-Sampling (RUS) [33], [34]. The algorithm-level approaches modify the contribution of different classes in the learning procedure by cost-sensitive learning or introducing a prior probability for each class [35], [36]. The hybrid approaches integrate both data-level and algorithm-level approaches [37], [38].

In practice, GNNs are also readily biased to majority classes in imbalanced classification [39]–[42]. This is because the class assignment for the node in GNNs is not simply determined by node features, but also affected by its neighbors in the graph. The nodes in the majority classes tend to have more frequent connections with its neighbors. In addition, the majority classes would dominate feature propagation between nodes. Therefore, node classification in graph is more challenging.

In literature, there are data-level approaches for GNNs to handle imbalanced node classification. Zhao et al. proposed GraphSMOTE that generated new synthetic nodes for minority classes and introduced an edge generator to simultaneously model the relations [16]. Qu et al. proposed ImGAGN that addressed imbalanced classification issue on graphs by introducing a graph generator and a GCN discriminator to discriminate between real and fake nodes [17]. Park et al. proposed GraphENS that synthesized the whole ego network for minor class to tackle class-imbalanced node classification issue [18]. Zhou et al. proposed SPARC that gradually learned the rare category-oriented network representation and the characterization model under the framework self-paced learning [20].

There are also algorithm-level approaches for GNNs to handle imbalanced node classification. Shi et al. proposed DR-GCN that incorporated both class-conditioned adversarial learning and latent distribution alignment between the labeled and unlabeled data into the training loss [19]. Chen et al. proposed ReNode to address the topology-imbalance issue by reweighting the influence of labeled nodes adaptively based on their relative positions to class boundaries [21]. Wang et al. proposed RSDNE and RECT for semi-supervised network embedding with completely imbalanced labels [22]. Yun et al. proposed LTE4G that jointly considered both class long-tailedness and degree long-tailedness for imbalanced node classification [23].

Although these approaches alleviated the imbalance issue in node classification to some extent, they did not capture more discriminative information from the minority classes. Besides, they did not consider the data heterogeneity issue in medical applications.

III. Method

A. Problem Statement

The notations used in this paper are shown in Table I.

TABLE I
NOTATIONS USED IN THE PAPER

Notations	Description
N	Number of nodes in the graph
K	Number of classes
$N^{(l)}$	Number of labeled nodes in the graph
$N_k^{(l)}$	Number of labeled nodes of the k-th class in the graph
$D^{(l)}$	The dimension of the node representation in the l-th layer
$.^T$	Transpose operation
$g(\cdot)$	The GCN-based encoder
$h(\cdot)$	shared distance layer across FC and HOFC views
$diag(\cdot)$	Operator that converts a vector into a diagonal matrix, with the input vector's elements forming the diagonal.
$diff(\cdot,\cdot)$	Difference function in the embedding space
y_i	True label for the i-th subject
I_k	The index set of the labeled samples in the k-th class
$\hat{y}_{i,k}$	The probability that the i-th subject predicted to be the k-th class
$\mathbf{A}^{FC}, \mathbf{A}^{HOFC}$	The adjacency matrices for $\mathcal{G}^{FC}$ and $\mathcal{G}^{HOFC}$
$\mathbf{X}^{FC}, \mathbf{X}^{HOFC}$	Node feature matrices for $\mathcal{G}^{FC}$ and $\mathcal{G}^{HOFC}$
$\mathcal{G}^{FC}, \mathcal{G}^{HOFC}$	The FC and HOFC graphs, $\mathcal{G}^{FC} = (\mathbf{X}^{FC}, \mathbf{A}^{FC})$, $\mathcal{G}^{HOFC} = (\mathbf{X}^{HOFC}, \mathbf{A}^{HOFC})$
$\mathbf{x}_i^{FC}, \mathbf{x}_i^{HOFC}$	The FC and HOFC feature vectors for the i-th subject
$\mathbf{p}_k^{FC}, \mathbf{p}_k^{HOFC}$	The prototype vectors of the k-th class in FC and HOFC views
$(\mathbf{z}_i^{FC})^{(l)}, (\mathbf{z}_i^{HOFC})^{(l)}$	The i-th node representations in the l-th layer of the FC and HOFC channels
$(\mathbf{Z}^{FC})^{(l)}, (\mathbf{Z}^{HOFC})^{(l)}$	The node matrices of the l-th layer in the FC and HOFC channels
$\mathbf{Z}^{global}$	The matrix of the global node representations
$(\mathbf{W}^{FC})^{(l)}, (\mathbf{W}^{HOFC})^{(l)}$	The learnable weight matrices of the l-th layer in the FC and HOFC channel
$\mathbf{S}, \mathbf{S}_k$	The support sets
$\mathbf{Q}, \mathbf{Q}_k$	The query sets

Let $\mathbf{x}_i^{FC}, \mathbf{x}_i^{HOFC} \in \mathbb{R}^D$ be the FC and HOFC feature vectors extracted from the rs-fMRI of the i-th subject, respectively, N is the total number of nodes in the graph, and $\mathbf{x}_i = ((\mathbf{x}_i^{FC})^T, (\mathbf{x}_i^{HOFC})^T)^T$, $i = 1, \cdots, N$. The $\mathbf{x}_i^{HOFC}$ can be considered as an augmentation view of $\mathbf{x}_i^{FC}$ for the i-th subject. Let $\mathbf{X}^{FC} = (\mathbf{x}_1^{FC}, \cdots, \mathbf{x}_N^{FC})^{\boldsymbol{T}} \in \mathbb{R}^{N\times D}$ and $\mathbf{X}^{HOFC} = (\mathbf{x}_1^{HOFC}, \cdots, \mathbf{x}_N^{HOFC})^{\boldsymbol{T}} \in \mathbb{R}^{N\times D}$ be the feature matrices for both FC and HOFC views, and $\mathbf{A}^{FC}$ and $\mathbf{A}^{HOFC}$ are the adjacency matrices in both views, respectively. We consider semi-supervised node classification based on a multi-view attributed

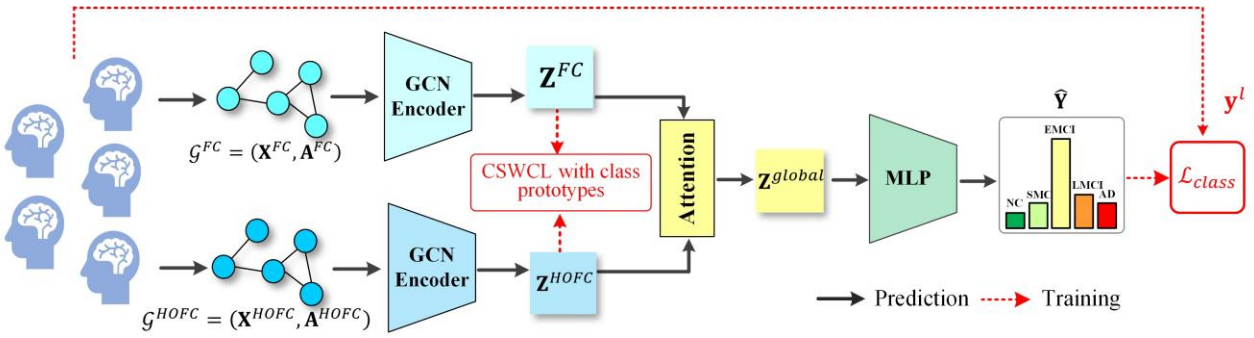

Fig.2. Framework of imbalanced AD staging using CSWCL-GCN

graph $\mathcal{G}=\{\mathcal{G}^{FC},\mathcal{G}^{HOFC}\}$, where $\mathcal{G}^{FC}=(\mathbf{X}^{FC},\mathbf{A}^{FC})$ and $\mathcal{G}^{HOFC}=(\mathbf{X}^{HOFC},\mathbf{A}^{HOFC})$. Following the common settings of semi-supervised node classification, only a part of nodes $\mathbf{X}^{l,FC}=\left(\mathbf{x}_1^{FC},\cdots,\mathbf{x}_{N^{(l)}}^{FC}\right)^{\boldsymbol{T}}$ and $\mathbf{X}^{l,HOFC}=\left(\mathbf{x}_1^{HOFC},\cdots,\mathbf{x}_{N^{(l)}}^{HOFC}\right)^{\boldsymbol{T}}$ have labels $\mathbf{y}^l=(y_1,\cdots,y_{N^{(l)}})$, where $N^{(l)}$ is the number of the labeled nodes in the graph, and $y_i\in\mathbb{R}$ is the label of the i-th subject, $i=1,\cdots,N^{(l)}$.

Given multi-view attributed graph $\mathcal{G}=\{\mathcal{G}^{FC},\mathcal{G}^{HOFC}\}$ and the labels y_is corresponding to a part of nodes, $i=1,\cdots,N^{(l)}$, the goal of CSWCL-GCN is to simultaneously optimize the parameters of $\mathcal{G}$ so that the unlabeled nodes can be classified well, especially when the labeled samples are imbalanced. The objective function of optimizing the network can be formulated as:

$$\min_{\theta}\mathcal{L}(\mathbf{X}^{FC},\mathbf{X}^{HOFC},\mathbf{y}^l)=\sum_{\substack{\mathbf{x}_i^{FC}\in\{\mathbf{x}_1^{FC},\ldots,\mathbf{x}_{N^{(l)}}^{FC}\}\\ \mathbf{x}_i^{HOFC}\in\{\mathbf{x}_1^{HOFC},\ldots,\mathbf{x}_{N^{(l)}}^{HOFC}\}}}\ell\left(f_\theta(\mathbf{x}_i^{FC},\mathbf{x}_i^{HOFC}),y_i\right)+R(\mathbf{X}^{l,FC},\mathbf{X}^{l,HOFC}) \quad (1)$$

where $f_\theta\colon\{\mathbf{x}_i^{FC},\mathbf{x}_i^{HOFC}\}\to y_i$ is a function learned by the network, $f_\theta(\mathbf{x}_i^{FC},\mathbf{x}_i^{HOFC})$ is the prediction of the i-th subject and $\ell(\cdot)$ is a loss (e.g. cross entropy) that measures the difference between the predictions and true labels, $R(\cdot,\cdot)$ is the regularization for cost-sensitive weighted contrastive learning.

B. CSWCL-GCN: The proposed model

Fig. 2 shows the framework of the proposed CSWCL-GCN. The network has two channels, each one of which is a node-level graph encoder that generates node representations using GCNs. The generated embeddings in both channels are then combined by attention and fed into the downstream network for classification. Following most GCN works [43], [44], we set three graph convolution layers for the GCN encoders in both channels.

To extract information in both channels of CSWCL-GCN, we design the propagation rule at the l-th layer as follows:

$$(\mathbf{Z}^{FC})^{(l)}=ReLU\left(\left(\tilde{\mathbf{D}}^{FC}\right)^{-\frac{1}{2}}\tilde{\mathbf{A}}^{FC}\left(\tilde{\mathbf{D}}^{FC}\right)^{-\frac{1}{2}}(\mathbf{Z}^{FC})^{(l-1)}(\mathbf{W}^{FC})^{(l)}\right),\quad l=1,\cdots,L \quad (2)$$

$$(\mathbf{Z}^{HOFC})^{(l)}=ReLU\left(\left(\tilde{\mathbf{D}}^{HOFC}\right)^{-\frac{1}{2}}\tilde{\mathbf{A}}^{HOFC}\left(\tilde{\mathbf{D}}^{HOFC}\right)^{-\frac{1}{2}}(\mathbf{Z}^{FC})^{(l-1)}(\mathbf{W}^{HOFC})^{(l)}\right),\quad l=1,\cdots,L \quad (3)$$

where both $(\mathbf{W}^{FC})^{(l)}$ and $(\mathbf{W}^{HOFC})^{(l)}$ are the trainable weight matrices in the l-th GCN layer of both channels, respectively; $\tilde{\mathbf{A}}^{FC}=\mathbf{A}^{FC}+\mathbf{I}$ and $\tilde{\mathbf{A}}^{HOFC}=\mathbf{A}^{HOFC}+\mathbf{I}$ are the enhanced adjacency matrices by adding the identity matrix $\mathbf{I}$ to $\mathbf{A}^{FC}$ and $\mathbf{A}^{HOFC}$, respectively, helping the node's information retained in network propagation. The $\tilde{\mathbf{D}}^{FC}$ and $\tilde{\mathbf{D}}^{HOFC}$ are the diagonal degree matrices of $\tilde{\mathbf{A}}^{FC}$ and $\tilde{\mathbf{A}}^{FC}$, respectively, $ReLU(\cdot)$ is the rectified linear unit (RELU) activation function, and $(\mathbf{Z}^{FC})^{(0)}=\mathbf{X}^{FC}$, $(\mathbf{Z}^{HOFC})^{(0)}=\mathbf{X}^{HOFC}$.

Since each node includes different amounts of FC and HOFC information, their contribution to the objective function is also different. To this end, we introduce the attention mechanism to encode different amount of node information into the global node representation. Let $\mathbf{Z}^{FC}=(\mathbf{z}_1^{FC},\cdots,\mathbf{z}_N^{FC})^{\boldsymbol{T}}=(\mathbf{Z}^{FC})^{(L)}$ and $\mathbf{Z}^{HOFC}=(\mathbf{z}_1^{HOFC},\cdots,\mathbf{z}_N^{HOFC})^{\boldsymbol{T}}=(\mathbf{Z}^{HOFC})^{(L)}$ be the output embedding matrices of GCNs in both FC and HOFC views, respectively, $\mathbf{z}_i^{FC},\mathbf{z}_i^{HOFC}\in\mathbb{R}^{D^{(L)}}$. The global node representations are obtained using the following fusion strategy:

$$\mathbf{Z}^{global}=diag(\boldsymbol{\alpha}_{FC})\cdot\mathbf{Z}^{FC}+diag(\boldsymbol{\alpha}_{HOFC})\cdot\mathbf{Z}^{HOFC} \quad (4)$$

where $\boldsymbol{\alpha}_{FC},\boldsymbol{\alpha}_{HOFC}\in\mathbb{R}^N$ record the attention values of the nodes in both views, respectively. Both $\boldsymbol{\alpha}_{FC}$ and $\boldsymbol{\alpha}_{HOFC}$ are learned as follows:

$$(\boldsymbol{\alpha}_{FC},\boldsymbol{\alpha}_{HOFC})=att(\mathbf{Z}^{FC},\mathbf{Z}^{HOFC}) \quad (5)$$

where $att(\mathbf{Z}^{FC},\mathbf{Z}^{HOFC})$ is computed in the same way as in [45].

The computed global node representations $\mathbf{Z}^{global}$ in Eq. (4) are fed into the fully connected layer for final classification. The prediction result is denoted as $\hat{\mathbf{Y}}=\left[\hat{y}_{i,k}\right]\in\mathbb{R}^{N\times K}$,where $\hat{y}_{i,k}$ is the probability that the i-th node predicted to be the k-th class, and $\hat{\mathbf{Y}}$ is calculated as follows:

$$\hat{\mathbf{Y}}=softmax(\mathbf{Z}^{global}\boldsymbol{\xi}+\boldsymbol{\eta}) \quad (6)$$

where $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$ are the parameters of the fully connected layer.

Suppose I_k is the indexed set of the labeled samples in the k-th class. For the i-th sample where $i\in I_k$, the true label is $y_{i,k}$. The cross-entropy loss function $\mathcal{L}_{class}$ for classifying all labeled nodes is calculated as follows:

$$\mathcal{L}_{class}=-\sum_{k=1}^{K}\frac{1}{N_k^{(l)}}\sum_{i\in I_k}y_{i,k}\ln\hat{y}_{i,k} \quad (7)$$

where $N_k^{(l)}$ is the number of training samples in class k.

C. Cost-sensitive weighted contrastive learning

Existing studies have indicated that both FC and HOFC reflect the correlation of two brain regions at different levels [3], [46]. Meanwhile, both FCs and HOFCs extracted from the same

subject share the same information. However, the samples in the minority classes are always limited and they provide inadequate supervision in the learning procedure. Although over-sampling or under-sampling strategies can be used to generate a balanced dataset to train the classifier, they cannot capture more knowledge from the minority classes. To this end, we propose cost-sensitive weighted contrastive learning (CSWCL) to capture more classification information without over-sampling minority classes. Specifically, given a set of multi-view subjects $\mathbf{z}_i = [(\mathbf{z}_i^{FC})^T, (\mathbf{z}_i^{HOFC})^T]^T, i = 1, \cdots, N$, we consider $(\mathbf{z}_i^{FC}, \mathbf{z}_i^{HOFC})$ as a positive pair and both $(\mathbf{z}_i^{FC}, \mathbf{z}_j^{HOFC})$ and $(\mathbf{z}_i^{HOFC}, \mathbf{z}_j^{FC})$ $i \neq j$, as negative pairs. Our goal is to optimize the GCN encoders in both channels by mapping multi-view embeddings of the same subject, i.e. $\mathbf{z}_i^{FC}$ and $\mathbf{z}_i^{HOFC}$, to nearby points while mapping multi-view embeddings of different subjects i.e. $\mathbf{z}_i^{FC}$ and $\mathbf{z}_j^{HOFC}$, $i \neq j$, to faraway points in the embedding space. To achieve this goal, $\mathbf{z}_i^{FC}$ and $\mathbf{z}_i^{HOFC}$ are used as anchors in the FC and HOFC views, respectively. We fix the anchor in one view and enumerate positive and negative pairs in the other view, and construct the following two losses for contrastive learning across two views:

$$\mathcal{L}_{cl}^{FC} = -\sum_{i=1}^{N} log \frac{exp\left(\frac{s_i^+}{\tau}\right)}{exp\left(\frac{s_i^+}{\tau}\right) + \sum_{j=1}^{N} \mathbb{1}_{i \neq j} \cdot exp\left(\frac{s_{i,j}^{FC}}{\tau}\right)} \quad (8)$$

$$\mathcal{L}_{cl}^{HOFC} = -\sum_{i=1}^{N} log \frac{exp\left(\frac{s_i^+}{\tau}\right)}{exp\left(\frac{s_i^+}{\tau}\right) + \sum_{j=1}^{N} \mathbb{1}_{i \neq j} \cdot exp\left(\frac{s_{i,j}^{HOFC}}{\tau}\right)} \quad (9)$$

where $s_i^+ = sim(\mathbf{z}_i^{FC}, \mathbf{z}_i^{HOFC}) = sim(\mathbf{z}_i^{HOFC}, \mathbf{z}_i^{FC})$ is the similarity of the positive pairs, $s_{i,j}^{FC} = sim(\mathbf{z}_i^{FC}, \mathbf{z}_j^{HOFC}), i \neq j$, and $s_{i,j}^{HOFC} = sim(\mathbf{z}_i^{HOFC}, \mathbf{z}_j^{FC}), i \neq j$, are the similarity of the negative pairs, τ is the temperature hyper-parameter; and $\mathbb{1}_{i \neq j}$ takes the value of 1 if $i \neq j$, otherwise it takes the value of 0.

The above losses equally consider each pair of graph nodes in contrastive learning. However, each graph node carries different amount of information, making negative pairs have different contributions from each other in the contrastive learning procedure. That is, if a graph node carries more class information, it should contribute more in the learning procedure. This can be achieved by assigning a larger weight in contrastive learning; In contrast, it contributes less to the loss and should be assigned a smaller weight. To this end, we further introduce weighting mechanism in contrastive learning and propose a weighted contrastive learning method, in which the negative pairs with different amount of information have various contributions. This is conducive to learning better node representations.

Considering different sizes of multiple classes in the training data, we further introduce the cost-sensitive mechanism based on the class sizes. The weights of the anchors in both FC and HOFC views, i.e. $\mathbf{z}_i^{FC}$ and $\mathbf{z}_i^{HOFC}$, can be computed as $w_i = \frac{N^{(l)}}{N_k^{(l)}}$, $i \in I_k$.

Conclusively, the losses of CSWCL for FC and HOFC features are formulated as follows:

$$\mathcal{L}_{cswcl}^{FC} = -\sum_{i=1}^{N^{(l)}} w_i log \frac{exp\left(\frac{s_i^+}{\tau}\right)}{exp\left(\frac{s_i^+}{\tau}\right) + \boldsymbol{E}^{HOFC} \cdot \sum_{j=1}^{N} \mathbb{1}_{i \neq j} \cdot \varphi_j^{HOFC} \cdot exp\left(\frac{s_{i,j}^{FC}}{\tau}\right)} \quad (10)$$

$$\mathcal{L}_{cswcl}^{HOFC} = -\sum_{i=1}^{N^{(l)}} w_i log \frac{exp\left(\frac{s_i^+}{\tau}\right)}{exp\left(\frac{s_i^+}{\tau}\right) + \boldsymbol{E}^{FC} \cdot \sum_{j=1}^{N} \mathbb{1}_{i \neq j} \cdot \varphi_j^{FC} \cdot exp\left(\frac{s_{i,j}^{HOFC}}{\tau}\right)} \quad (11)$$

where φ_j^{FC} and φ_j^{HOFC} denote the weights of negative pairs in the contrastive learning across two views, $\mathbf{E}^{FC} = \frac{1}{\sum_{j=1}^{N} \varphi_j^{FC}}$ and $\mathbf{E}^{HOFC} = \frac{1}{\sum_{j=1}^{N} \varphi_j^{HOFC}}$ are normalization factors; w_i is the weight of the i-th sample in the embedding space. The weight of positive pairs should also be considered in CSWCL. However, it will not affect $\mathcal{L}_{cswcl}^{FC}$ or $\mathcal{L}_{cswcl}^{HOFC}$ because all the weights of positive pairs will be normalized in the components of $\mathcal{L}_{cswcl}^{FC}$ or $\mathcal{L}_{cswcl}^{HOFC}$.

The loss function of CSWCL is obtained by combining $\mathcal{L}_{cswcl}^{FC}$ and $\mathcal{L}_{cswcl}^{HOFC}$ as follows:

$$\mathcal{L}_{cswcl} = \mathcal{L}_{cswcl}^{FC} + \mathcal{L}_{cswcl}^{HOFC} \quad (12)$$

D. Jointly learning multi-view prototypes for weights evaluation in CSWCL

The weights of negative pairs in CSWCL are evaluated based on the distance between the node representations and the prototypes of the class they belong to. A class prototype is closely surrounded by nodes of the same class, and it serves as the representative of that class. If a node is close to a class prototype in the embedding space, it carries large amount of class information, and vice versa. On the other hand, the class prototypes should be far from each other in the embedding space so that the nodes in different classes can be differentiated from each other.

To learn the class prototypes in both FC and HOFC views, we develop an adaptive learning procedure. Let $\mathbf{p}_k^{FC}$ and $\mathbf{p}_k^{HOFC}$ are the prototypes of the k-th class in FC and HOFC views, respectively. To compute the prototypes $\mathbf{p}_k^{FC}$ and $\mathbf{p}_k^{HOFC}$ in both views, we divide the training set into two parts, i.e., the support sets $\mathbf{S} = \{\mathbf{S}_k | k = 1, \cdots, K\}$ and the query sets $\mathbf{Q} = \{\mathbf{Q}_k | k = 1, \cdots, K\}$. $\mathbf{Q}_k$ only includes one query sample $\mathbf{q}_k$ in the k-th class of the training data, and $\mathbf{S}_k$ includes the remaining training samples in the k-th class except the query sample. Given class k, we compute the prototypes $\mathbf{P}_k^{FC}$ and $\mathbf{P}_k^{HOFC}$ for class k as follows:

$$\mathbf{p}_k^{FC} = \frac{1}{|\mathbf{S}_k|} \sum_{v_i \in \mathbf{S}_k} \mathbf{z}_i^{FC}, k = 1, \cdots, K \quad (13a)$$

$$\mathbf{p}_k^{HOFC} = \frac{1}{|\mathbf{S}_k|} \sum_{v_i \in \mathbf{S}_k} \mathbf{z}_i^{HOFC}, k = 1, \cdots, K \quad (13b)$$

Since $\mathbf{S}_k$ includes most samples in the class, both $\mathbf{p}_k^{FC}$ and $\mathbf{p}_k^{HOFC}$ computed with Eq. (13) are close to the class centers and they are not sensitive to the selection of the query sample.

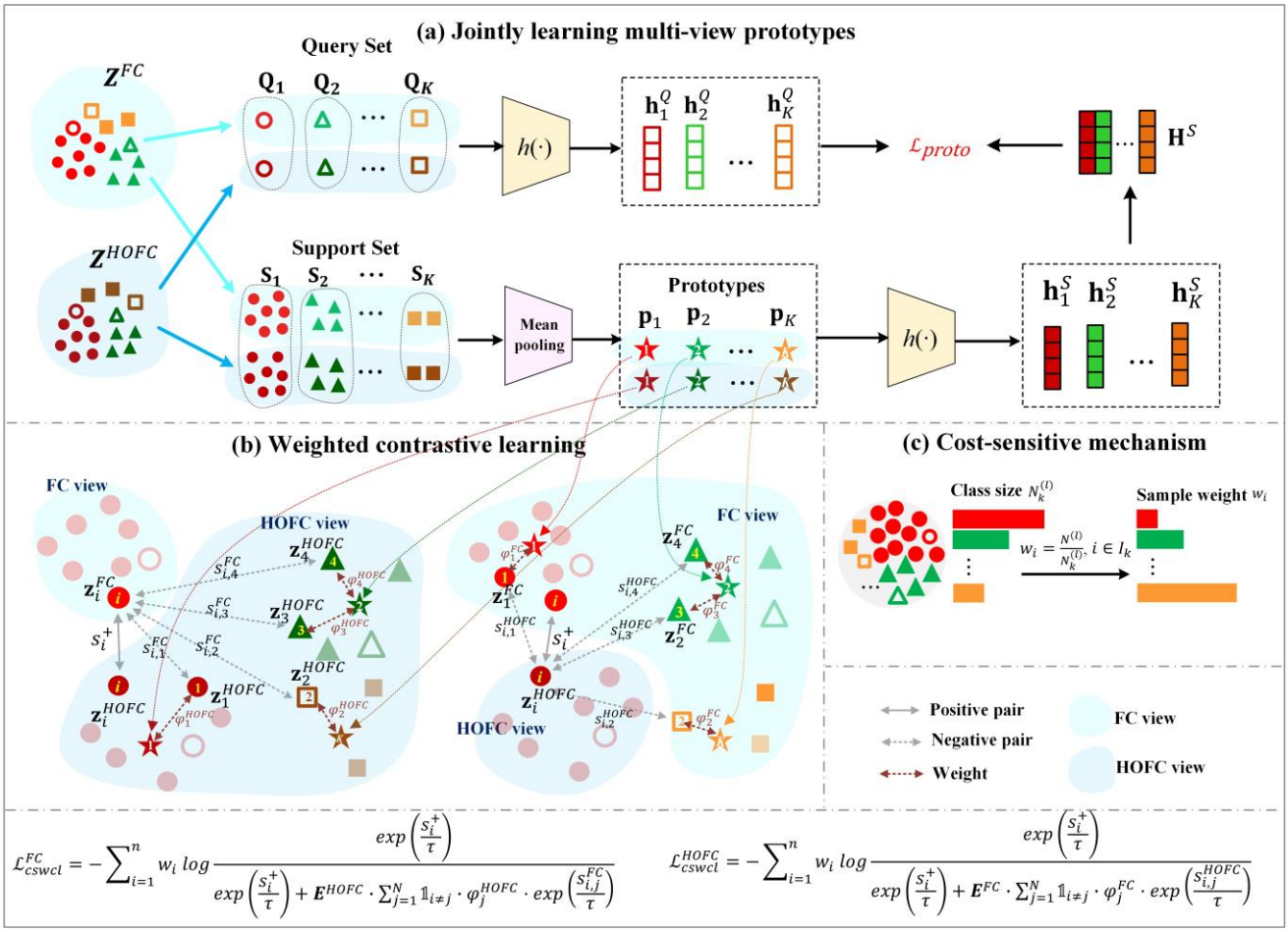

Fig. 3. Cost-sensitive weighted contrastive learning with class prototypes for imbalanced data. (a) D. Jointly learning multi-view prototypes; (b) Weighted contrastive learning; (c) Cost-sensitive mechanism.

Given $\mathbf{z}_i = [(\mathbf{z}_i^{FC})^T, (\mathbf{z}_i^{HOFC})^T]^T$ in the multi-view embedding space, its distance representation to all the prototypes can be computed as follows:

$$\mathbf{d}_{i,k}^{FC} = diff(\mathbf{z}_i^{FC}, \mathbf{p}_k^{FC}), k = 1, \cdots, K \quad (14a)$$

$$\mathbf{d}_{i,k}^{HOFC} = diff(\mathbf{z}_i^{HOFC}, \mathbf{p}_k^{HOFC}), k = 1, \cdots, K \quad (14b)$$

$$\mathbf{d}_i^{FC} = \|_{k\in\{1,\ldots,K\}}\, \mathbf{d}_{i,k}^{FC} \quad (15a)$$

$$\mathbf{d}_i^{HOFC} = \|_{k\in\{1,\ldots,K\}}\, \mathbf{d}_{i,k}^{HOFC} \quad (15b)$$

$$\mathbf{d}_i = [(\mathbf{d}_i^{FC})^T, (\mathbf{d}_i^{HOFC})^T]^T \quad (15c)$$

where $\mathbf{d}_{i,k}^{FC}$ and $\mathbf{d}_{i,k}^{HOFC}$ are the distance representations to all the prototypes in each view, $diff(\mathbf{z}, \mathbf{p}) = \mathbf{z} - \mathbf{p}$ is the distance function to compute the distance representation. The above procedure can be implemented by a shared distance layer denoted as $h(\cdot)$, i.e. $\mathbf{d}_i = h(\mathbf{z}_i)$.

The distance representation of a query sample in the k-th class should be similar to that of the prototype in the same class. To learn such representations, we feed both $\mathbf{q}_k = [(\mathbf{q}_k^{FC})^T, (\mathbf{q}_k^{HOFC})^T]^T$ and the corresponding prototype $\mathbf{p}_k = [(\mathbf{p}_k^{FC})^T, (\mathbf{p}_k^{HOFC})^T]^T$ into the shared distance layer, respectively. We have:

$$\begin{aligned} \mathbf{h}_k &= h(\mathbf{p}_k), k = 1, \cdots, K \\ \mathbf{h}_k^Q &= h(\mathbf{q}_k)\ k = 1, \cdots, K \end{aligned} \quad (16)$$

where $\mathbf{h}_k$ and $\mathbf{h}_k^Q$ are the distance representations of both prototypes and query sample of the k-th class, respectively. Let $\mathbf{H^S} = (\mathbf{h}_1, \cdots, \mathbf{h}_K)^T$ be a matrix whose rows are the distance representations of the prototypes. For $\mathbf{h}_l^Q$, we have:

$$\tilde{\mathbf{y}}_l = softmax\ ((\mathbf{H^S})^T \cdot \mathbf{h}_l^Q), \quad (17)$$

where $\tilde{\mathbf{y}}_l = (\tilde{y}_{l,1}, \cdots, \tilde{y}_{l,K})^T \in \mathbb{R}^K$ is a vector whose element records the similarity between $\mathbf{h}_l^Q$ and $\mathbf{h}_k$s. Obviously, if the network is optimized, $\tilde{y}_{l,k}$, $k = l$, tends to be 1, whereas $\tilde{y}_{l,k}$s, $k \neq l$, tend to be 0. We therefore have the following loss to learn the prototypes and the shared distance layer:

$$\mathcal{L}_{proto} = -\sum_{l=1}^{K}\sum_{k=1}^{K} y_{l,k}^Q \ln \tilde{y}_{l,k} \quad (18)$$

Based on the learned prototypes, the weights of negative pairs in Eqs.(10)-(11) are determined by the distance between the sample to the prototype of the class it belongs to in the embedding space, i.e.

$$\varphi_j^{FC} = \frac{1}{\left\|diff\left(\mathbf{z}_j^{FC}, \mathbf{p}_k^{FC}\right)\right\|_2} \quad (19a)$$

$$\varphi_j^{HOFC} = \frac{1}{\left\|diff\left(\mathbf{z}_j^{HOFC}, \mathbf{p}_k^{HOFC}\right)\right\|_2} \quad (19b)$$

where φ_j^{FC} and φ_j^{HOFC} denote the weight of negative pairs in Eqs. (10)-(11).

Fig. 3 shows the pipeline of cost-sensitive weighted contrastive learning with class prototypes.

E. Optimization

In summary, the overall objective function of CSWCL-GCN is formulated as follows:

$$\mathcal{L} = \mathcal{L}_{class} + \lambda_1 \mathcal{L}_{cswcl} + \lambda_2 \mathcal{L}_{proto} \tag{20}$$

where λ_1 and λ_2 are hyperparameters for $\mathcal{L}_{cswcl}$ and $\mathcal{L}_{proto}$, respectively, and $\mathcal{L}_{class}$, $\mathcal{L}_{cswcl}$, and $\mathcal{L}_{proto}$ are defined by Eq. (7), (12), and (18), respectively.

The following Algorithm 1 provides the training procedure of CSWCL-GCN.

Algorithm 1 The training procedure of CSWCL-GCN

Input: $\mathcal{G}^{FC} = (\mathbf{X}^{FC}, \mathbf{A}^{FC})$; $\mathcal{G}^{HOFC} = (\mathbf{X}^{HOFC}, \mathbf{A}^{HOFC})$; label matrix $\mathbf{Y} \in \mathbb{R}^{N\times K}$; maximal number of iterations Λ; the GCN layer number $L = 3$;
Output: The trained network.
Initialization: $(\mathbf{Z}^{FC})^{(0)} \leftarrow \mathbf{X}^{FC}$, $(\mathbf{Z}^{HOFC})^{(0)} \leftarrow \mathbf{X}^{HOFC}$, $L \leftarrow 3$, randomly set the network parameters $(\mathbf{W}^{FC})^{(l)}$, $(\mathbf{W}^{HOFC})^{(l)}$, $l = 1, \cdots, L$
Procedure:
for $\lambda = 1$ to Λ **do**
  Compute $\hat{\mathbf{Y}}$ using Eqs. (2), (3), (5), (4) and (6) in turn;
  Compute $\mathcal{L}_{class}$ using Eqs. (7);
  Divide the training sets into support set and query sets and compute $\mathbf{H}^{\mathbf{S}}$ using Eqs. (13), (16) in turn;
  Compute $\mathcal{L}_{proto}$ using Eqs. (17), (18) in turn;
  Compute $\mathcal{L}_{cswcl}$ by Eqs. (19), (10), (11), (12) in turn, respectively;
  Compute the overall loss $\mathcal{L}$ in Eq. (20);
  Update the network parameters using the gradient decent;
end for

IV. Experiments

A. Experimental settings

1) Datasets and preprocessing

Our proposed CSWCL-GCN is evaluated on the ADNI (Alzheimer's Disease Neuroimaging Initiative) [1], which is a multicenter database for AD study. It includes MRI, PET, biospecimen, genetic and clinical data. In this study, we only used the rs-fMRI data. We selected 720 subjects in this study, including 184 normal controls (NCs), 40 SMC subjects, 208 EMCI subjects, 172 LMCI subjects and 116 AD subjects. Table II shows the demographic information of the subjects.

TABLE II
Demographic Information of Data Used in Our Study

	Age(mean±SD, years)	Age range(years)	Sex(M/F)
NCs	75.7±6.1	[65, 96]	101/83
SMC	7 73.5±5.4	[66, 83]	22/18
EMCI	72.1±7.2	[56, 91]	118/90
LMCI	72.4±7.6	[57, 89]	59/113
AD	74.7±7.5	[56, 89]	56/60
Total	73.5±7.1	[56, 96]	364/356

In our experiments, we construct a multi-view population graph $\mathcal{G} = \{\mathcal{G}^{FC}, \mathcal{G}^{HOFC}\}$ using FC and HOFC features, respectively. Specifically, we adopt AAL atlas to parcellate the brain into 116 regions. Then we compute Pearson's correlation between the BOLD signals of two brain regions as the FC, and obtain a 116×116 correlation matrix of FCs [47]. We consider each row in the FC matrix as the description of a brain region, and further calculate the Pearson correlation between two rows, resulting in a new 116×116 correlation matrix, in which each element is the HOFC between two brain regions. After that, the upper triangular elements in both FC and HOFC matrices for each subject are reshaped into a 6670-dimensional feature vector, which includes redundant and noisy information for AD staging. We adopt the recursive feature elimination (RFE) method [48] to extract 500 discriminative FC and HOFC features, respectively, and use the extracted FC and HOFC features as the nodes in both $\mathcal{G}^{FC}$ and $\mathcal{G}^{HOFC}$, respectively.

To compute the adjacent matrices $\mathbf{A}^{FC}$ and $\mathbf{A}^{HOFC}$ for $\mathcal{G}^{FC}$ and $\mathcal{G}^{HOFC}$, we consider two factors. The first factor is the similarity between the nodes in both FC and HOFC views. We this study, we use the following Eq. (21) to compute the similarity between the nodes in each view:

$$Sim\left(\mathbf{x}_i^{FC}, \mathbf{x}_j^{FC}\right) = exp\left(-\frac{\rho^2\left(\mathbf{x}_i^{FC}, \mathbf{x}_j^{FC}\right)}{2\sigma^2}\right) \tag{21a}$$

$$Sim\left(\mathbf{x}_i^{HOFC}, \mathbf{x}_j^{HOFC}\right) = exp\left(-\frac{\rho^2\left(\mathbf{x}_i^{HOFC}, \mathbf{x}_j^{HOFC}\right)}{2\sigma^2}\right) \tag{21b}$$

where $\rho(\cdot,\cdot)$ is the Euclidean distance function, and σ is the kernel width.

The second factor is the difference between phenotypic measures of two subjects. Following [12], we use $\gamma(\cdot,\cdot)$ to compute the distance between phenotypic measures. For categorical phenotypic measures such as sex, $\gamma(\cdot,\cdot)$ is defined as the Kronecker function δ and, for quantitative phenotypic measures such as age, $\gamma(\cdot,\cdot)$ is defined as the following Eq. (22):

$$\gamma(M_h(i), M_h(j)) = \begin{cases} 1, & |M_h(i) - M_h(j)| < \theta \\ 0, & Otherwise \end{cases} \tag{22}$$

where $M_h(i)$ and $M_h(j)$ are the h-th phenotypic measure for subject i and j, respectively, and θ is the threshold.

Let $a_{i,j}^{FC}$ and $a_{i,j}^{HOFC}$ be the elements in the adjacent matrices $\mathbf{A}^{FC}$ and $\mathbf{A}^{HOFC}$, respectively. Considering the above two factors, $a_{i,j}^{FC}$ and $a_{i,j}^{HOFC}$ are computed as:

$$a_{i,j}^{FC} = Sim\left(\mathbf{x}_i^{FC}, \mathbf{x}_j^{FC}\right) \sum_h \gamma(M_h(i), M_h(j)) \tag{23a}$$

$$a_{i,j}^{HOFC} = Sim\left(\mathbf{x}_i^{HOFC}, \mathbf{x}_j^{HOFC}\right) \sum_h \gamma(M_h(i), M_h(j)) \tag{23b}$$

2) Experimental settings

All the models in this study are trained using Adam Optimizer in Pytorch. We use the following experimental settings for GCN-based models: the GCN layer number L=3, learning rate = 0.003, epochs = 500, and dropout rate = 0.3. In our model, both hyperparameters in Eq. (20) are set as $\lambda_1 = 0.02$ and $\lambda_2 = 0.1$ based on prior experience. When we construct the graph in both FC and HOFC views, both node similarity and sex are considered to compute the adjacent matrices using Eq. (23). All the models are evaluated using the stratified five-fold cross validation.

3) Evaluation metrics

Following previous works in imbalanced classification [19], we evaluate the performance of the models with F1-macro, F1-

weighted and AUC scores. Specifically, both F1-macro and F1-weighted scores evaluate model performance across different classes. F1-macro takes the average accuracy of each class and F1-weighted takes the weighted average to account for the issue of class imbalance. In this study, we use the Python *sklearn* package to calculate both metrics.

B. Ablation study

To verify the effectiveness of each component in the loss of CSWCL-GCN, we compare its performance with several variants. All the networks have the same architectures with different losses.

MV-GCN: A CSWCL-GCN variant that uses $\mathcal{L}_{class}$ as the loss.

P-GCN: A CSWCL-GCN variant that uses $\mathcal{L}_{class}$ and $\mathcal{L}_{proto}$ as the loss.

CP-GCN: A CSWCL-GCN variant that uses $\mathcal{L}_{class}$, $\mathcal{L}_{cl}$ and $\mathcal{L}_{proto}$ as the loss.

WCL-GCN: A CSWCL-GCN variant that uses $\mathcal{L}_{class}$, $\mathcal{L}_{wcl}$ and $\mathcal{L}_{proto}$ as the loss.

$$\mathcal{L}_{wcl}^{FC} = -\sum_{i=1}^{N^{(l)}} log \frac{exp(s_i^+/\tau)}{exp\left(\frac{s_i^+}{\tau}\right) + \boldsymbol{E}^{HOFC} \cdot \sum_{j=1}^{N} \mathbb{1}_{i \neq j} \cdot \varphi_j^{HOFC} \cdot exp\left(\frac{s_{i,j}^{FC}}{\tau}\right)} \tag{24a}$$

$$\mathcal{L}_{wcl}^{HOFC} = -\sum_{i=1}^{N^{(l)}} log \frac{exp(s_i^+/\tau)}{exp\left(\frac{s_i^+}{\tau}\right) + \boldsymbol{E}^{FC} \cdot \sum_{j=1}^{N} \mathbb{1}_{i \neq j} \cdot \varphi_j^{FC} \cdot exp\left(\frac{s_{i,j}^{HOFC}}{\tau}\right)} \tag{24b}$$

$$\mathcal{L}_{wcl} = \mathcal{L}_{wcl}^{FC} + \mathcal{L}_{wcl}^{HOFC} \tag{24c}$$

CSWCL-GCN: The proposed model.

Table III summarizes their features.

TABLE III
METHODS FOR COMPARISON

	$\mathcal{L}_{class}$	$\mathcal{L}_{cl}$	$\mathcal{L}_{proto}$	$\mathcal{L}_{wcl}$	$\mathcal{L}_{cswcl}$
MV-GCN	√	×	×	×	×
P-GCN	√	×	√	×	×
CP-GCN	√	√	√	×	×
WCL-GCN	√	×	√	√	×
CSWCL-GCN (proposed)	√	×	√	×	√

* $\mathcal{L}_{class}$: Objective function in Eq. (7);
$\mathcal{L}_{cl}$: Objective function computed by $\mathcal{L}_{cl} = \mathcal{L}_{cl}^{FC} + \mathcal{L}_{cl}^{HOFC}$;
$\mathcal{L}_{proto}$: Objective function in Eq. (18);
$\mathcal{L}_{wcl}$: Objective function in Eq. (24a),(24b),(24c);
$\mathcal{L}_{cswcl}$: Objective function in Eq. (20).

1) Effectiveness of learning prototypes in embedding space

CSWCL-GCN learns the class prototypes in the training procedure. Since each class is represented by the prototype, the imbalance issues caused by the majority and minority classes are solved. To fully show the effectiveness of learning prototypes by minimizing $\mathcal{L}_{proto}$, we compare the performance of MV-GCN with that P-GCN, which includes $\mathcal{L}_{proto}$ in the loss while MV-GCN does not. As Table III shows, P-GCN yields better performance compared with MV-GCN, indicating the effectiveness of learning prototypes by minimizing $\mathcal{L}_{proto}$ in handling the class imbalance issue in AD staging.

2) Effectiveness of weighted contrastive learning

Both CSWCL-GCN and WCL-GCN are developed based on weighted contrastive learning, in which the negative pairs have different contributions to the contrastive loss in the learning procedure. To verify the effectiveness of weighting mechanism in contrastive learning, we compare the performance of both methods with that of CP-GCN, which learns the node representations using classical contrastive learning. As shown in Table IV, both CSWCL-GCN and WCL-GCN yield better results than CP-GCN, indicating that weighted contrastive learning is conducive to learning better node representations for subsequent classification.

TABLE IV
RESULTS OF ABLATION STUDY

Methods	F1-macro (%)	F1-weight (%)	AUC (%)
MV-GCN	62.37 ± 0.87	64.54 ± 1.12	80.88 ± 0.30
P-GCN	64.26 ± 0.95	65.69 ± 0.83	82.30 ± 1.00
CP-GCN	65.38 ± 1.00	67.04 ± 1.77	83.13 ± 0.80
WCL-GCN	67.54 ± 0.78	69.50 ± 0.62	84.49 ± 0.95
CSWCL-GCN (proposed)	**75.09 ± 3.90**	**82.41 ± 2.4**	**88.90 ± 1.86**

C. Comparison with imbalanced classification methods

The proposed CSWCL-GCN leverages the knowledge in the minority classes as the supervision via cost-sensitive weighted contrastive learning. To verify its advantage in imbalanced classification, we compare the performance of CSWCL-GCN with existing strategies for imbalanced classification including over-sampling [49], reweighting [50] and GraphSMOTE [16]. The following methods are developed for comparison:

MV-GCN-OS: An MV-GCN variant that uses over-sampling strategy for imbalanced classification. The training samples in minority classes are randomly copied until the number of samples in minority classes are comparable to those in majority classes. A multi-view graph is then constructed using both training and testing samples in FC and HOFC views. Then MV-GCN is applied, and cross-entropy loss is used as the learning criterion.

MV-GCN-RW: An MV-GCN variant that reweights the samples in the cross-entropy loss. A multi-view graph is first constructed using both training and testing samples in FC and HOFC views. Then an MV-GCN is then applied, and the cost-sensitivity mechanism is introduced into the cross-entropy loss as the learning criterion, in which the samples are reweighted by the inverse ratio of the class sizes to the total number of training samples.

GraphSMOTE [16]**:** SMOTE on graphs. Both FC and HOFC features of each subject are first concatenated, and a graph is then constructed using both training and testing samples. The GraphSMOTE is applied on the graph.

Table V provides the experimental results. According to the table, one may observe that both MV-GCN-OS and MV-GCN-RW achieve better performance on the task of imbalanced AD staging than MV-GCN. However, their performance is still

limited in F1-macro, F1-weighted and AUC because both methods either simply repeat the training samples in minority classes or reweight the samples in the loss, without capturing more discriminative knowledge in the minority classes. The classification performance of GraphSMOTE is better than those of both MV-GCN-OS and MV-GCN-RW. It generates edges that connect synthetic and original nodes in the graph and encodes the similarity between the nodes in an embedding space. In addition, an edge generator is trained to model the relations between the newly generated samples. However, the discriminative knowledge in the minority classes is not fully considered in GraphSMOTE. Different from these methods, the proposed CSWCL-GCN captures more discriminative knowledge from the minority classes via self-supervised contrastive learning across two views, and further considers the heterogeneity of the disease via reweighting negative pairs in contrastive learning. Therefore, it outperforms other methods.

TABLE V
RESULTS OF COMPARISON WITH REPRESENTATIVE IMBALANCED CLASSIFICATION METHODS

Methods	F1-macro (%)	F1-weighted (%)	AUC (%)
MV-GCN-OS	62.69 ± 0.95	64.65 ± 0.97	81.28 ± 0.75
GraphSMOTE [16]	64.67 ± 1.21	66.40 ± 1.13	82.47 ± 0.67
MV-GCN-RW	63.66 ± 1.34	64.91 ± 0.70	81.91 ± 0.57
CSWCL-GCN (proposed)	**75.09 ± 3.90**	**82.41 ± 2.4**	**88.90 ± 1.86**

D. Comparison with graph contrastive learning methods

The proposed CSWCL-GCN reweights the multi-view node representations based on its distance to the prototypes, ensuring that negative pairs have different contributions in the contrastive learning. In addition, the cost-sensitive mechanism is added to balance the contributions of anchors in the loss of contrastive learning. These modifications to classical contrastive learning are conducive to learning better node representations. To verify the advantage of cost-sensitive weighted contrastive learning in our method, we compared the performance of CSWCL-GCN with other graph contrastive learning methods including GraphCL [27], GCC [28],and DGI [29], and provided the comparison results in Table VI.

TABLE VI
RESULTS OF COMPARISON WITH REPRESENTATIVE METHODS THAT COMBINES GRAPH AND CONTRASTIVE LEARNING

Method	F1-macro(%)	F1-weighted(%)	AUC(%)
GraphCL[27]	64.57±1.08	66.62±1.56	82.67±0.68
GCC[28]	65.45±0.88	67.42±1.28	82.85±0.55
DGI[29]	63.27±1.02	65.20±1.37	82.41±0.84
CSWCL-GCN (proposed)	**75.09 ± 3.90**	**82.41 ± 2.4**	**88.90 ± 1.86**

From the table, one may observe that our proposed method yields better results in F1-macro, F1-weighted and AUC scores, indicating that it is more effective than other graph contrastive learning methods in AD staging. This is because the proposed method utilizes cost-sensitive weighted contrastive learning to learn more discriminative node representations. In addition, the discriminative knowledge in the minority class is captured to provide more supervision for learning better node representations.

V. CONCLUSION

A novel CSWCL-GCN based on rs-fMRIs is proposed for imbalanced AD staging in this study. It is developed based on cross-view information including both FC and HOFC features. Meanwhile, a novel cost-sensitive weighted contrastive learning procedure is proposed to capture discriminative information from the minority classes. Both heterogeneity of the disease and data imbalanced issue are handled in the learning procedure. Experimental results show that CSWCL-GCN outperforms the state-of-the-art methods on the ADNI database. In summary, our proposed method is not only efficient and reliable for AD staging, but also provides a solid foundation for using GCNs on imbalanced AD staging.

Although our proposed method outperforms the state-of-the-art methods for AD staging, it still has limitations. For example, we use mean-pooling to compute the class prototype for each class, which assumes that the class information be well-represented by a single prototype. A single prototype may not adequately capture data in the class, particularly when dealing with non-unimodal distributions, which is common in medical applications due to data heterogeneity. In our future work, we plan to develop more advanced prototype learning methods to handle both class imbalance issue and data heterogeneity in medical applications.

REFERENCES

[1] C. R. Jack Jr. *et al*., "The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods," *Journal of Magnetic Resonance Imaging*, vol. 27, no. 4, pp. 685–691, 2008.

[2] C. I. Mark, E. L. Mazerolle, and J. J. Chen, "Metabolic and vascular origins of the BOLD effect: Implications for imaging pathology and resting-state brain function," *Journal of Magnetic Resonance Imaging*, vol. 42, no. 2, pp. 231–246, 2015.

[3] H. Zhang *et al*., "Topographical Information-Based High-Order Functional Connectivity and Its Application in Abnormality Detection for Mild Cognitive Impairment," *Journal of Alzheimer's Disease*, vol. 54, no. 3, pp. 1095–1112, Jan. 2016.

[4] D. Jin *et al*., "Grab-AD: Generalizability and reproducibility of altered brain activity and diagnostic classification in Alzheimer's Disease," *Human Brain Mapping*, vol. 41, no. 12, pp. 3379–3391, 2020.

[5] J. Zhao, X. Ding, Y. Du, X. Wang, and G. Men, "Functional connectivity between white matter and gray matter based on fMRI for Alzheimer's disease classification," *Brain and Behavior*, vol. 9, no. 10, p. e01407, Oct. 2019.

[6] A. Khazaee, A. Ebrahimzadeh, and A. Babajani-Feremi, "Classification of patients with MCI and AD from healthy controls using directed graph measures of resting-state fMRI," *Behavioural Brain Research*, vol. 322, pp. 339–350, Mar. 2017.

[7] F. Ramzan, M. U. G. Khan, A. Rehmat, S. Iqbal, T. Saba, A. Rehman, and Z. Mehmood, "A Deep Learning Approach for Automated Diagnosis and Multi-Class Classification of Alzheimer's Disease Stages Using Resting-State fMRI and Residual Neural Networks," *J Med Syst*, vol. 44, no. 2, p. 37, Dec. 2019.

[8] C. R. Jack *et al*., "NIA-AA Research Framework: Toward a biological definition of Alzheimer's disease," *Alzheimer's & Dementia*, vol. 14, no. 4, pp. 535–562, Apr. 2018.

[9] H. Jiang, P. Cao, M. Xu, J. Yang, and O. Zaiane, "Hi-GCN: A hierarchical graph convolution network for graph embedding learning of brain network

and brain disorders prediction," *Computers in Biology and Medicine*, vol. 127, p. 104096, Dec. 2020.

[10] Y. Zhu , Junbo Ma, "Interpretable learning based Dynamic Graph Convolutional Networks for Alzheimer's Disease analysis," *Information Fusion*, 2022.

[11] H. Zhou, L. He, Y. Zhang, L. Shen, and B. Chen, "Interpretable Graph Convolutional Network Of Multi-Modality Brain Imaging For Alzheimer's Disease Diagnosis," in *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, 2022, pp. 1–5.

[12] S. Parisot, S. I. Ktena, E. Ferrante, M. Lee, R. Guerrero, B. Glocker, and D. Rueckert, "Disease prediction using graph convolutional networks: Application to Autism Spectrum Disorder and Alzheimer's disease," *Medical Image Analysis*, vol. 48, pp. 117–130, Aug. 2018.

[13] H. Zhang *et al*., "Classification of Brain Disorders in rs-fMRI via Local-to-Global Graph Neural Networks," *IEEE Transactions on Medical Imaging*, vol. 42, no. 2, pp. 444–455, Feb. 2023.

[14] W. Tong *et al*., "fMRI-Based Brain Disease Diagnosis: A Graph Network Approach," *IEEE Transactions on Medical Robotics and Bionics*, vol. 5, no. 2, pp. 312–322, May 2023.

[15] D. Klepl, F. He, M. Wu, D. J. Blackburn, and P. Sarrigiannis, "EEG-Based Graph Neural Network Classification of Alzheimer's Disease: An Empirical Evaluation of Functional Connectivity Methods," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 30, pp. 2651–2660, 2022.

[16] T. Zhao, X. Zhang, and S. Wang, "GraphSMOTE: Imbalanced Node Classification on Graphs with Graph Neural Networks," in *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, Virtual Event Israel, 2021, pp. 833–841.

[17] L. Qu, H. Zhu, R. Zheng, Y. Shi, and H. Yin, "ImGAGN: Imbalanced Network Embedding via Generative Adversarial Graph Networks," in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, Virtual Event Singapore, 2021, pp. 1390–1398.

[18] J. Park, J. Song, and E. Yang, "GraphENS: Neighbor-Aware Ego Network Synthesis for Class-Imbalanced Node Classification," presented at the International Conference on Learning Representations, 2021.

[19] M. Shi, Y. Tang, X. Zhu, D. Wilson, and J. Liu, "Multi-Class Imbalanced Graph Convolutional Network Learning," in *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, Yokohama, Japan, 2020, pp. 2879–2885.

[20] D. Zhou, J. He, H. Yang, and W. Fan, "SPARC: Self-Paced Network Representation for Few-Shot Rare Category Characterization," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, London United Kingdom, 2018, pp. 2807–2816.

[21] D. Chen, Y. Lin, G. Zhao, X. Ren, P. Li, J. Zhou, and X. Sun, "Topology-Imbalance Learning for Semi-Supervised Node Classification," in *Advances in Neural Information Processing Systems*, 2021, vol. 34, pp. 29885–29897.

[22] Z. Wang, X. Ye, C. Wang, J. Cui, and P. S. Yu, "Network Embedding with Completely-imbalanced Labels," *IEEE Trans. Knowl. Data Eng.*, vol. 33, no. 11, pp. 3634–3647, Nov. 2021.

[23] S. Yun, K. Kim, K. Yoon, and C. Park, "LTE4G: Long-Tail Experts for Graph Neural Networks," in *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, 2022, pp. 2434–2443.

[24] X. Liu, F. Zhang, Z. Hou, L. Mian, Z. Wang, J. Zhang, and J. Tang, "Self-Supervised Learning: Generative or Contrastive," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 1, pp. 857–876, Jan. 2023.

[25] L. Wu, H. Lin, C. Tan, Z. Gao, and S. Z. Li, "Self-Supervised Learning on Graphs: Contrastive, Generative, or Predictive," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 4, pp. 4216–4235, Apr. 2023.

[26] X. Wang, N. Liu, H. Han, and C. Shi, "Self-supervised Heterogeneous Graph Neural Network with Co-contrastive Learning." arXiv, 19-May-2021.

[27] Y. You, T. Chen, Y. Sui, T. Chen, Z. Wang, and Y. Shen, "Graph Contrastive Learning with Augmentations," in *Advances in Neural Information Processing Systems*, 2020, vol. 33, pp. 5812–5823.

[28] J. Qiu, Q. Chen, Y. Dong, J. Zhang, H. Yang, M. Ding, K. Wang, and J. Tang, "GCC: Graph Contrastive Coding for Graph Neural Network Pre-Training," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, Virtual Event CA USA, 2020, pp. 1150–1160.

[29] P. Veličković, W. Fedus, and W. L. Hamilton, "Deep Graph Infomax," in *proceedings International Conference on Learning Representations (I CLR),* New Orleans, LA, USA, 2019..

[30] B. Lam, M. Masellis, and M. Freedman, "Clinical, imaging, and pathological heterogeneity of the Alzheimer's disease syndrome," *Alzheimer's Research & Therapy*, vol. 5, no. 1, p. 1, Jan. 2013.

[31] K. A. Jellinger, "Recent update on the heterogeneity of the Alzheimer's disease spectrum," *J Neural Transm*, vol. 129, no. 1, pp. 1–24, Jan. 2022.

[32] Y. Wang, C. Aggarwal, and T. Derr, "Distance-wise Prototypical Graph Neural Network for Imbalanced Node Classification," in *Proceedings of the 17th International Workshop on Mining and Learning with Graphs (MLG)*, Washington, 2022.

[33] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, Jun. 2002.

[34] M. A. Tahir, J. Kittler, and K. Mikolajczyk, "A Multiple Expert Approach to the Class Imbalance Problem Using Inverse Random under Sampling," in *Multiple Classifier Systems*, Berlin, Heidelberg, 2009, pp. 82–91.

[35] S. Lawrence, I. Burns, A. Back, A. C. Tsoi, and C. L. Giles, "Neural Network Classification and Prior Class Probabilities," in *Neural Networks: Tricks of the Trade*, G. B. Orr and K.-R. Müller, Eds. Berlin, Heidelberg: Springer, 1998, pp. 299–313.

[36] X. Wang, J. Wang, F. Shan, Y. Zhan, J. Shi, and D. Shen, "Severity prediction of pulmonary diseases using chest CT scans via cost-sensitive label multi-kernel distribution learning," *Computers in Biology and Medicine*, vol. 159, p. 106890, Jun. 2023.

[37] N. V. Chawla, A. Lazarevic, L. O. Hall, and K. W. Bowyer, "SMOTEBoost: Improving Prediction of the Minority Class in Boosting," in *Knowledge Discovery in Databases: PKDD 2003*, Berlin, Heidelberg, 2003, pp. 107–119.

[38] Xu-Ying Liu, Jianxin Wu, and Zhi-Hua Zhou, "Exploratory Undersampling for Class-Imbalance Learning," *IEEE Trans. Syst., Man, Cybern. B*, vol. 39, no. 2, pp. 539–550, Apr. 2009.

[39] X. Juan, F. Zhou, W. Wang, W. Jin, J. Tang, and X. Wang, "INS-GNN: Improving graph imbalance learning with self-supervision," *Information Sciences*, vol. 637, p. 118935, Aug. 2023.

[40] Z. Huang, Y. Tang, and Y. Chen, "A graph neural network-based node classification model on class-imbalanced graph data," *Knowledge-Based Systems*, vol. 244, p. 108538, May 2022.

[41] K. Wang, J. An, M. Zhou, Z. Shi, X. Shi, and Q. Kang, "Minority-Weighted Graph Neural Network for Imbalanced Node Classification in Social Networks of Internet of People," *IEEE Internet of Things Journal*, vol. 10, no. 1, pp. 330–340, Jan. 2023.

[42] Y. Wang, "Fair Graph Representation Learning with Imbalanced and Biased Data," in *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, New York, NY, USA, 2022, pp. 1557–1558.

[43] T. N. Kipf and M. Welling, "Semi-Supervised Classification with Graph Convolutional Networks." presented at the *5th International Conference on Learning Representations (ICLR)*, Toulon, France, 2017.

[44] J. Zhou, G. Cui, S. Hu, Z. Zhang, C. Yang, Z. Liu, L. Wang, C. Li, and M. Sun, "Graph neural networks: A review of methods and applications," *AI Open*, vol. 1, pp. 57–81, 2020.

[45] X. Wang, M. Zhu, and D. Bo, "AM-GCN: Adaptive Multi-channel Graph Convolutional Networks," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, New York, NY, USA, 2020, pp. 1243–1253.

[46] X. Chen, H. Zhang, Y. Gao, C. Wee, G. Li, and D. Shen, "High-order resting-state functional connectivity network for MCI classification," *Hum Brain Mapp*, vol. 37, no. 9, pp. 3282–3296, May 2016.

[47] D. Yao *et al.*, "Triplet Graph Convolutional Network for Multi-scale Analysis of Functional Connectivity Using Functional MRI," in *Graph Learning in Medical Imaging*, Cham, 2019, pp. 70–78.

[48] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene Selection for Cancer Classification using Support Vector Machines," *Machine Learning*, vol. 46, no. 1, pp. 389–422, Jan. 2002.

[49] R. Mohammed, J. Rawashdeh, and M. Abdullah, *Machine Learning with Oversampling and Undersampling Techniques: Overview Study and Experimental Results*. 2020, p. 248.

[50] B. Yuan and X. Ma, "Sampling + reweighting: Boosting the performance of AdaBoost on imbalanced datasets," in *The 2012 International Joint Conference on Neural Networks (IJCNN)*, 2012, pp. 1–6.