Multilevel Feature Representation of FDG-PET Brain Images for Diagnosing Alzheimer's Disease

Xiaoxi Pan, Mouloud Adel, Caroline Fossati, Thierry Gaidon, and Eric Guedj for the Alzheimer's Disease Neuroimaging Initiative*

***Abstract*—Using a single imaging modality to diagnose Alzheimer's disease (AD) or mild cognitive impairment (MCI) is a challenging task. FluoroDeoxyGlucose Positron Emission Tomography (FDG-PET) is an important and effective modality used for that purpose. In this paper, we develop a novel method by using single modality (FDG-PET) but multilevel feature, which considers both region properties and connectivities between regions to classify AD or MCI from normal control. First, three levels of features are extracted: statistical, connectivity, and graph-based features. Then, the connectivity features are decomposed into three different sets of features according to a proposed similarity-driven ranking method, which can not only reduce the feature dimension but also increase the classifier's diversity. Last, after feeding the three levels of features to different classifiers, a new classifier selection strategy, maximum Mean squared Error (mMsE), is developed to select a pair of classifiers with high diversity. In order to do the majority voting, a decision-making scheme, a nested cross validation technique is applied to choose another classifier according to the accuracy. Experiments on Alzheimer's Disease Neuroimaging Initiative database show that the proposed method outperforms most FDG-PET-based classification algorithms, especially for classifying progressive MCI (pMCI) from stable MCI (sMCI).**

***Index Terms*—Alzheimer's disease (AD), ensemble classification, FDG-PET, multilevel feature representation.**

Manuscript received December 16, 2017; revised May 3, 2018 and June 19, 2018; accepted July 14, 2018. Date of publication September 19, 2018; date of current version July 1, 2019. *(Corresponding author: Mouloud Adel.)*

X. Pan, C. Fossati, and T. Gaidon are with the Ecole Centrale de Marseille, Institut Fresnel, UMR 7249, Marseille 13013, France (e-mail: xiaoxi.pan@fresnel.fr; caroline.fossati@fresnel.fr; thierry.gaidon@fresnel.fr).

M. Adel is with the Aix-Marseille Université, Institut Fresnel, UMR 7249, Marseille 13013, France (e-mail: mouloud.adel@univ-amu.fr).

E. Guedj is with the Aix-Marseille Université, Institut Fresnel, UMR 7249, Marseille 13013, France, and also with the Centre Européen de Recherche en Imagerie Médicale, Marseille 13013, France (e-mail: Eric.GUEDJ@ap-hm.fr).

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

Digital Object Identifier 10.1109/JBHI.2018.2857217

I. Introduction

ALZHEIMER's Disease (AD) is a dominant neurodegenerative brain disease and the main cause of dementia in elderly people worldwide. It is expected that 115 million people will be affected by this disease in 2050 [1]. The National Institute on Aging and Alzheimer's Association (NIA-AA) criteria distinguish 3 clinical stages: asymptomatic preclinical phase (pre-clinical stage of AD), amnestic Mild Cognitive Impairment (MCI) phase due to AD, and AD dementia phase [2]–[4]. These criteria introduce the utility of different biomarkers of the pathophysiological process to weight the diagnostic probability of the disease [2], [5]. One of them is FDG-PET, which is effective in diagnosing AD [6]. It can reveal pathophysiological changes before irreversible anatomical changes and provide useful information about the cerebral glucose metabolic rate [7].

Machine learning techniques offer an automatic and objective classification framework for high-dimensional data processing and can learn complex patterns of changes across various imaging modalities [8]. Computer-Aided Diagnosis (CAD) based on machine learning approaches is a useful method for doctors, and can bring a quantitative evaluation to better detect brain diseases. Therefore, developing a method that can be used to distinguish AD and MCI from Normal Control (NC) automatically is important yet challenging.

In recent papers [8], [9], multi-modality-based algorithms, specifically combining MRI and FDG-PET, are the most commonly used methods, since different modalities can provide complementary information [10], [11]. Shi *et al.* [12] devised a coupled feature representation based on MRI and FDG-PET to diagnose AD and MCI. Liu *et al.* [13] developed a method under deep learning architecture which used a zero-masking strategy for data fusion to extract complementary information from MRI and FDG-PET. There has been a growing interest in using FDG-PET as a single modality to diagnose AD and MCI as well. These FDG-PET-based methods can be classified into 2 main categories according to the type of used features: 1) voxel-based methods, which used voxels as features [14], [15]; 2) atlas-based methods, which segmented a subject into different regions and the region information was then used as features [16]–[18]. But they take only region properties into consideration without connectivities between regions. In fact, a human brain is a complex system and multiple regions interact with

each other [19], [20]. Therefore, connectivities between regions are important in AD and MCI diagnosis and cannot be ignored.

In this study, we investigate the multi-level feature representation for FDG-PET data to diagnose AD and MCI. The major contributions can be summarized as three folds: 1) the multi-level feature representation considers not only region properties (1st-Level), but also the connectivity between any pair of regions (2nd-Level) and an overall connectivity between one region and the other regions (3rd-Level); 2) a similarity-driven ranking method is proposed to rank regions from highly affected to slightly affected by the disease, which can decompose the 2nd-level feature, thereby reducing the feature dimension and increasing the classifier's diversity to a certain degree; 3) a classifier selection strategy, maximum Mean squared Error (mMsE), is proposed to choose a pair of classifiers with high diversity to enhance the ensemble effect, especially for the case that sub-classifiers do not perform well.

The remaining of the paper is organized as follows. Section II describes the novel multi-level representation method for diagnosing AD and MCI. Section III reports and analyzes the experimental results. Finally, a conclusion of this work is given in Section IV.

II. Methods

The proposed multi-level feature representation method is described from 3 aspects in details, including feature extraction, feature selection and ensemble classification, as shown in Fig. 1. First, after segmenting each subject into 116 Regions of Interest (ROIs) according to an Automated Anatomical Labeling (AAL) atlas [21], 3 levels of features are extracted, specifically, the 1st-Level feature, which comprises ROI's mean intensity and standard deviation. The 2nd-Level feature, the similarity-based connectivity between any pair of ROIs, is decomposed into 3 sets according to a proposed similarity-driven ranking method. The 3rd-Level feature is composed of graph-based features. Next, Least Absolute Shrinkage and Selection Operator (LASSO) [22] is applied to do the feature selection for each set of features, respectively. Then different classifiers are trained using different sets of features. Final prediction is obtained through an ensemble classifier decided by a proposed maximum Mean squared Error (mMsE) strategy and a nested cross validation technique.

A. Dataset

Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial MRI, PET, other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of MCI and early AD.

ADNI provides different imaging modalities, such as structural MRI, functional MRI, Diffusion Tensor Imaging (DTI) and PET, for researchers to develop methods for early detection of AD. In this study, we focus on using FDG-PET data to diagnose AD and MCI. After having acquired original data, there are usually 3 steps in data processing: spatial normalization, smoothing and intensity normalization. In the existing literatures, most researchers do these procedures independently. In fact, besides those original data, ADNI also provides processed data. There are 2 kinds of FDG-PET data images, pre-processed data and post-processed data. Specifically, for the pre-processed data, there are 4 different groups [23], including 1) Co-registered Dynamic; 2) Co-registered, Averaged; 3) Co-reg, Avg, Standardized Image and Voxel Size; 4) Co-reg, Avg, Std Img and Vox Siz, Uniform Resolution. The post-processed data was processed on the basis of group 4) data mentioned above and then spatially normalized to MNI template using SPM [24] with $2 \times 2 \times 2$ mm voxel size and $79 \times 95 \times 69$ matrix dimension. The intensity normalization is done by using the global mean value. It should be noted that the reason why we use the post-processed data is to avoid the impact of pre-treatments as far as possible and pay more attention to the influence of features and classification methods on results. Therefore, 272 post-processed baseline FDG-PET data were obtained from ADNI, including 94 subjects with AD, 88 subjects with MCI and 90 subjects under NC. MCI subjects were clinically further subdivided into 44 progressive MCI (pMCI), who progressed to AD in 24 months, and 44 stable MCI (sMCI), who did not progress to AD. Demographic and clinical information of subjects are provided in Table I.

B. Feature Extraction

Before extracting features, each subject is segmented into 116 ROIs using AAL atlas. Many methods in the existing literatures used mean gray level intensities of some ROIs as features [16], [18], [25]. However, only ROI's information is not enough. Therefore, in this paper, we explore to expand the feature pool computed on FDG-PET data.

1) First-Level Feature: Since each ROI's mean intensity and standard deviation can reflect the FDG uptake and its corresponding distribution, the 1st-Level feature for the n-th sample can be represented as:

$$\mathbf{r}_n^m = [r_{n1}^m, r_{n2}^m, \ldots, r_{np}^m] \tag{1}$$

$$\mathbf{r}_n^s = [r_{n1}^s, r_{n2}^s, \ldots, r_{np}^s] \tag{2}$$

where $\mathbf{r}_n^m$ and $\mathbf{r}_n^s$ are the mean intensity and standard deviation, respectively, and p is the number of ROIs, here $p = 116$.

2) Second-Level Feature: The 2nd-Level feature is the similarity-based connectivity between ROIs. Hereafter, connectivity is used to refer to similarity-based connectivity. First, the 1st-Level feature is used to represent each ROI, and the i-th ROI is represented by:

$$\mathbf{x}_i = [r_i^m, r_i^s] \tag{3}$$

then the connectivity between any two ROIs is computed through:

$$w_{ij} = \begin{cases} e^{-\|\mathbf{x}_i - \mathbf{x}_j\|^2} & i \neq j, \\ 0 & i = j. \end{cases} \tag{4}$$

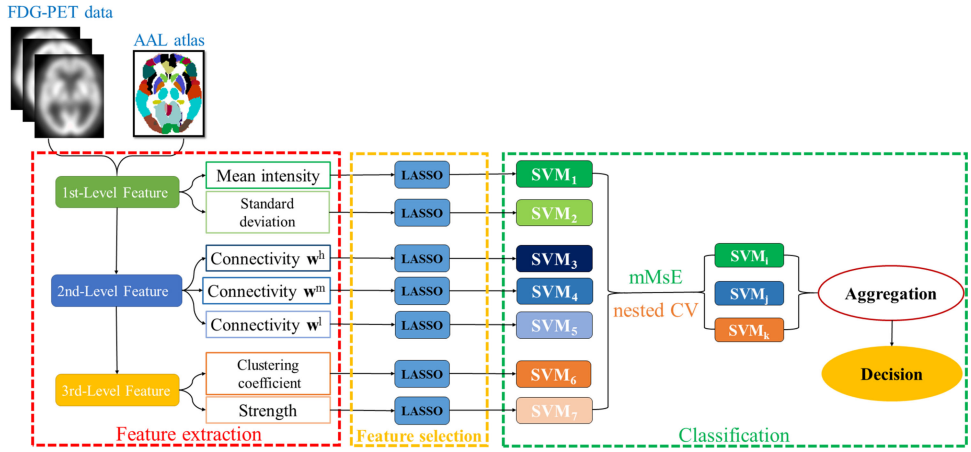

Fig. 1. Framework of the proposed method.

TABLE I
DEMOGRAPHIC AND CLINICAL INFORMATION OF THE SUBJECTS

Characteristic	AD	NC	MCI*	pMCI	sMCI
Number of subjects	94	90	88	44	44
Female/male	38/56	34/56	32/56	14/30	18/26
Age(Mean ± SD)	75.83 ± 7.37	76.08 ± 5.01	76.71 ± 6.63	75.86 ± 7.37	77.56 ± 5.75
MMSE(Mean ± SD)	23.46 ± 2.14	28.97 ± 1.15	26.92 ± 1.62	26.77 ± 1.78	27.07 ± 1.45

*Including 44 pMCI and 44 sMCI that are described in the last two columns.

where w_{ij} is the connectivity of the i-th ROI and the j-th ROI, and the higher the value of w_{ij}, the more similar the two ROIs. It should be noted that before computing w_{ij} through (4), each type of the 1st-Level feature is normalized over ROIs. The 2nd-Level feature of any subject is:

$$\mathbf{W_r} = \begin{bmatrix} 0 & w_{r12} & \cdots & w_{r1j} & \cdots & w_{r1p} \\ w_{r21} & 0 & \cdots & w_{r2j} & \cdots & w_{r2p} \\ \vdots & \vdots & \ddots & \vdots & & \vdots \\ w_{ri1} & w_{ri2} & \cdots & 0 & \cdots & w_{rip} \\ \vdots & \vdots & & \vdots & \ddots & \vdots \\ w_{rp1} & w_{rp2} & \cdots & w_{rpj} & \cdots & 0 \end{bmatrix} \tag{5}$$

where $\mathbf{W_r}$ is a symmetric matrix.

The 2nd-Level feature is composed of connectivities between all the 116 ROIs, totally 6670 dimensions $(116 \times (116-1)/2$, only considering the values on the upper triangle). Clearly, it is not an optimal dimension for the subsequent classification. Therefore, $\mathbf{W_r}$ is further decomposed into 3 subsets of features according to a proposed similarity-driven ranking method.

Similar to the way of computing connectivities between ROIs, we can obtain the similarity coefficients between subjects for a specific ROI:

$$w_{uv} = \begin{cases} e^{-\|\mathbf{x}_u - \mathbf{x}_v\|^2} & u \neq v, \\ 0 & u = v. \end{cases} \tag{6}$$

where u, v stands for the u-th and v-th subjects.

For any ROI, a symmetric matrix for subjects, $\mathbf{W_s}$, is obtained from:

$$\mathbf{W_s} = \begin{bmatrix} 0 & w_{s12} & \cdots & w_{s1v} & \cdots & w_{s1N} \\ w_{s21} & 0 & \cdots & w_{s2v} & \cdots & w_{s2N} \\ \vdots & \vdots & \ddots & \vdots & & \vdots \\ w_{su1} & w_{su2} & \cdots & 0 & \cdots & w_{suN} \\ \vdots & \vdots & & \vdots & \ddots & \vdots \\ w_{sN1} & w_{sN2} & \cdots & w_{sNv} & \cdots & 0 \end{bmatrix} \tag{7}$$

The dimension of $\mathbf{W_s}$ is determined by the number of subjects, N, in a group (AD, NC, MCI, pMCI and sMCI). For example, there are 94 subjects in AD group, so $N = 94$, then the dimension of $\mathbf{W_s}$ is 94×94. Each subject is segmented into 116 ROIs, thus there are 116 matrices like $\mathbf{W_s}$.

If taking NC subjects (including training and testing samples) as a reference, in one hand, for a ROI which is not affected by AD, the similarity coefficients between AD subjects are supposed to be close to those of NC subjects. In the other hand, for a ROI affected by AD, the similarity coefficients of AD subjects are different from NC group. In order to quantify the difference, we first make a statistic on the upper triangle values of $\mathbf{W_s}$ to get the frequency distribution histogram of those values. Then the cumulative probability curve of similarity coefficients can be obtained, as shown in Fig. 2, where (a), (b) and (c) stand for region Angular_L, region Hippocampus_L and region Cerebelum_10_R, respectively. It can be seen that there is a clear difference between the AD and NC groups in

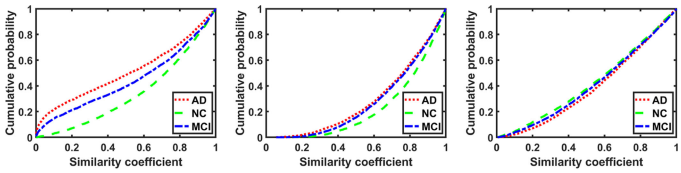

Fig. 2. Statistics of the similarity coefficients between subjects for a certain ROI. (a) ROI: Angular_L. (b) ROI: Hippocampus_L. (c) ROI: Cerebelum_10_R.

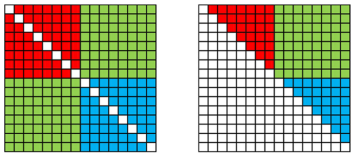

Fig. 3. Instance of the division for a similarity matrix.

Fig. 2(a), and for the other two ROIs, the difference decreases gradually. It implies that among the experimental subjects, region Cerebelum_10_R is almost unaffected by AD, while region Angular_L has a great chance of getting influenced, therefore region Angular_L is ranked before region Cerebelum_10_R, and region Hippocampus_L is placed between them. The difference between curves is computed through the difference of area under curve, which is denoted ΔS. The larger the ΔS, the greater the impact generated by AD for a ROI. At last, all the ROIs can be ranked according to ΔS from high to low. It should be noted that we highly recommend using a balance number of subjects in 2 groups for the comparison and the more the better.

After ranking all the ROIs, the similarity matrix $\mathbf{W_r}$ is recalculated according to the new order of ROIs. Then $\mathbf{W_r}$ is divided into 4 equal parts, as shown in Fig. 3(a), where the red part stands for the sets which are highly influenced by AD, denoted $\mathbf{W}^h$, while the blue part stands for ROIs with less impact of AD, denoted $\mathbf{W}^l$, and the green part represents the connectivities between highly influenced ROIs and slightly influenced ROIs, which is denoted $\mathbf{W}^m$. Since $\mathbf{W_r}$ is symmetric, only upper triangular matrix is taken into consideration, like in Fig. 3(b). Therefore, the 2nd-Level feature $\mathbf{W_r}$ is divided into 3 sets, and after converting them to vectors, the 2nd-Level feature for the n-th sample is represented as:

$$\mathbf{w}_n^h = [w_{n1}^h, w_{n2}^h, \ldots, w_{np^h}^h] \tag{8}$$

$$\mathbf{w}_n^m = [w_{n1}^m, w_{n2}^m, \ldots, w_{np^m}^h] \tag{9}$$

$$\mathbf{w}_n^l = [w_{n1}^l, w_{n2}^l, \ldots, w_{np^l}^l] \tag{10}$$

where p^h, p^m and p^l are the dimension of each subset of features. p^h and p^l are the same (red and blue parts in Fig. 3(b)), both equal to 1653 $(58 \times (58-1)/2)$, and p^m (green part) is 3364 (58×58). Apparently, compared to 6670 (red, blue and green parts), the dimension is decreased by about 50%–75%.

3) Third-Level Feature: The 3rd-Level feature is extracted from a graph point of view, which stands for an overall connectivity between a ROI and the other ROIs. Generally, a graph $G = (V, E)$ consists of a finite set V of vertices and a finite set of edges $E \subseteq V \times V$. A vertex in a graph is equivalent to a ROI in a brain. Therefore, the connectivity between the i-th ROI and the j-th ROI, w_{ij}, can be viewed as the weight of an edge which connects the i-th vertex and the j-th vertex. In this paper, we analyze the undirected graph, which means $w_{ij} = w_{ji}$. Then a subject can be represented by a graph, as shown in Fig. 4 which represents a subject from ADNI database.

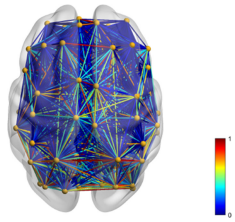

Fig. 4. Instance of the brain connectivity network from the axial view [26].

After constructing a graph for a subject, several graph measures can be computed, such as degree, strength, clustering coefficient, betweenness centrality [27]. According to [19], [28], the metrics strength and clustering coefficient are effective in discriminating AD, therefore the 3rd-Level feature is represented by these two graph measures. Specifically,

strength: the sum of a vertex's neighboring link weights [27].

$$s_i = \sum_{j=1}^{p} w_{ij} \tag{11}$$

where s_i is the strength of a vertex or a ROI.

clustering coefficient: the geometric mean of all triangles associated with each vertex [27].

$$\mathbf{c} = \frac{diag((\mathbf{W_r}^{\cdot\frac{1}{3}})^3)}{\mathbf{d}(\mathbf{d}-1)} \tag{12}$$

where $diag(\cdot)$ is a operator which takes the diagonal values from a matrix, $\mathbf{c}$ is a clustering coefficient vector, and $\mathbf{d}$ is a degree vector in which the element d_i is,

$$d_i = \sum_{j=1}^{p} a_{ij} \tag{13}$$

where a_{ij} is the connection status between the i-th vertex and the j-th vertex: $a_{ij} = 0$ when $w_{ij} = 0$, otherwise $a_{ij} = 1$.

Thus, the 3rd-Level feature consists of 2 sets of features, and each of them for the n-th sample is represented as:

$$\mathbf{g}_n^s = [s_{n1}, s_{n2}, \ldots, s_{np}] \tag{14}$$

$$\mathbf{g}_n^c = \mathbf{c}_n \tag{15}$$

These features exhibit different ranges of values. Thus a procedure of feature normalization is necessary by z-score prior to classification:

$$z_{nm} = \frac{f_{nm} - \mu_m}{\delta_m} \tag{16}$$

where f_{nm} is the value of the m-th feature of the n-th sample, and $f \in \{r^m, r^s, w^h, w^m, w^l, g^s, g^c\}$, μ_m and δ_m are the mean value and standard deviation of the m-th feature, respectively. Most of f_{nm} values can be transformed to the range $[-1, 1]$

through (16), while out-of-range values are clamped to either -1 or 1.

C. Feature Selection

In this paper, there are 3 levels of features. For the 1st-Level and 3rd-Level features, the dimension is 116 for each type of feature. For the 3 subsets of features in 2nd-Level, the dimension is 1653 ($\mathbf{w}^h$), 3364 ($\mathbf{w}^m$), 1653 ($\mathbf{w}^l$), respectively. Therefore, it is necessary to select representative features to reduce the feature dimension. A good strategy of feature reduction or selection is to remove irrelevant, redundant and noisy features and meanwhile improve classification performances. Least Absolute Shrinkage and Selection Operator (LASSO) is one of the popular techniques for dimension reduction and feature selection. It uses l_1 regularization to get a sparsity solution, thereby achieving the goal of feature selection. In this paper, feature selection is accomplished by using LASSO.

D. Ensemble Classification

The support vector machine (SVM) classifier is a popular and effective method in distinguishing subjects with AD or MCI from NC. In this study, 3 levels of features, which then are decomposed into 7 types of features, are fed into 7 linear SVMs to train 7 individual models, respectively. The motivation of training in this way is to ensure a model focus on one type of feature of the data. The margin parameter C of all the SVMs is fixed to 1 for a fair comparison, like [29], [30].

The effectiveness of an ensemble classifier depends on the number of individual classifiers and the diversity between them. The more the number of classifiers and the higher the diversity, the more effective the ensemble classifier is. However, if the subclassifier doesn't perform well (the accuracy is usually between 50% and 60%), the increase of the number of classifiers cannot improve the ensemble classifier's performance, because as the number of classifiers increases, the possibility that misclassified results accounted for the majority also increases. Thus, in order to enhance the ensemble effect and meanwhile, avoid misclassified results taken up the majority, a strategy of selecting models, maximum Mean square Error (mMsE), is proposed. Let $\mathbf{y}_i$ and $\mathbf{y}_j$ denote the output labels of SVM_i and SVM_j, respectively, then the Mean Square Error (MSE) between $\mathbf{y}_i$ and $\mathbf{y}_j$ is computed through,

$$M(i,j) = \frac{1}{K}\|\mathbf{y}_i - \mathbf{y}_j\|^2 \tag{17}$$

where K is the number of the testing samples and each element in $\mathbf{y}_i$ belongs to $\{-1, 1\}$. The higher the MSE, the greater the diversity between the outputs of classifiers. Then a pair of classifiers with high diversity can be achieved by finding the maximum MSE,

$$(i,j) = \arg\max_{i,j} M(i,j) \tag{18}$$

In addition, another classifier, $\mathbf{y}_k$, is determined through nested cross validation on the training set and the one with the highest accuracy is selected. Last, the final decision is made through a majority voting of the 3 selected classifiers' outputs:

$$\mathbf{Y} = \text{sgn}(\mathbf{y}_i + \mathbf{y}_j + \mathbf{y}_k) \tag{19}$$

where $\text{sgn}(\cdot)$ is a sign function. Even though the number of classifiers for decision making decreases, the classifiers with high diversity and high accuracy are kept. Therefore, the strategy can enhance the ensemble effect, especially in the case where all the classifiers do not have a good performance, since it can avoid misclassified results accounted for the majority.

Algorithm 1: Workflow of the Proposed Method.

1: Dividing the dataset into 10 parts, one of them is used as testing data and the remaining parts are for training;
2: Extracting 7 types of features for the training and testing data, respectively;
3: Selecting features by LASSO for each type of features;
4: Training different models using different types of features on training data;
5: Using the proposed mMsE method and the nested cross validation technique to choose 3 models;
6: Applying the 3 models on testing data and then the evaluation metrics (ACC, SEN, SPE, AUC) can be computed;
7: Returning to step 1, choosing another part as the testing data till all the 10 parts are used for testing;
8: Repeating step 1 to step 7 ten times, then computing the average value of each metric.

III. Experiments and Results

A. Experimental Setup

Experiments are conducted on 3 different kinds of classifications, including 1) AD vs. NC, 2) MCI vs. NC and 3) pMCI vs. sMCI. In order to evaluate the performance of the proposed method, 4 different metrics, classification accuracy (ACC), sensitivity (SEN), specificity (SPE), and area under curve (AUC) are used. The higher the values are, the better the corresponding method is. Specifically, ACC is the proportion of samples that are properly predicted. SEN implies the proportion of correctly classified AD or MCI samples. SPE means the proportion of NC samples that are correctly classified. Because of a limited number of samples, we use a 10-fold cross validation technique to assess the performance, and repeat 10 times to reduce the possible bias. The parameter in LASSO, λ, is decided by nested cross validation on the training dataset within the range $\{10^{-5}, 10^{-4}, \ldots, 10^{-1}\}$ for the 1st-Level and 3rd-Level features, and $\{10^{-9}, 10^{-8}, \ldots, 10^{-1}\}$ for the 2nd-Level feature. The parameter is chosen separately, which can help reduce the computation cost in a great extent. It should be noted that all the results shown in following parts are obtained after LASSO. The whole procedure is shown in Algorithm 1.

B. Single-Type Feature Representation Evaluation

The 3 levels of features are decomposed to 7 different types of features, and the performance of each type of feature is shown

TABLE II
PERFORMANCE OF DIFFERENT TYPE OF FEATURE FOR AD VS. NC(%)

Method	Feature	ACC	SEN	SPE	AUC
1st-Level	Mean intensity	85.13	86.61	83.97	93.39
1st-Level	Standard deviation	85.49	84.98	**86.24**	93.84
2nd-Level	Connectivity $\mathbf{w}^h$	85.05	86.24	84.56	93.01
2nd-Level	Connectivity $\mathbf{w}^m$	**86.88**	**88.82**	85.17	**93.88**
2nd-Level	Connectivity $\mathbf{w}^l$	83.98	84.31	83.37	91.37
3rd-Level	Strength	80.77	80.29	81.50	88.63
3rd-Level	Clustering coefficient	83.89	84.03	84.26	92.05

TABLE III
PERFORMANCE OF DIFFERENT TYPE OF FEATURE FOR MCI VS. NC(%)

Method	Feature	ACC	SEN	SPE	AUC
1st-Level	Mean intensity	73.55	75.01	72.87	81.36
1st-Level	Standard deviation	**78.19**	**78.31**	**78.69**	**86.67**
2nd-Level	Connectivity $\mathbf{w}^h$	72.78	70.63	74.35	83.19
2nd-Level	Connectivity $\mathbf{w}^m$	74.67	76.06	73.65	83.27
2nd-Level	Connectivity $\mathbf{w}^l$	74.89	77.01	72.68	78.94
3rd-Level	Strength	71.12	70.62	72.01	80.07
3rd-Level	Clustering coefficient	72.31	74.73	70.26	80.36

TABLE IV
PERFORMANCE OF DIFFERENT TYPE OF FEATURE FOR PMCI VS. SMCI(%)

Method	Feature	ACC	SEN	SPE	AUC
1st-Level	Mean intensity	**59.85**	**61.14**	**60.88**	**62.40**
1st-Level	Standard deviation	53.57	55.63	52.50	53.47
2nd-Level	Connectivity $\mathbf{w}^h$	52.01	55.04	51.27	55.82
2nd-Level	Connectivity $\mathbf{w}^m$	57.26	55.03	59.95	60.88
2nd-Level	Connectivity $\mathbf{w}^l$	56.35	55.65	56.00	58.64
3rd-Level	Strength	53.76	57.35	52.08	56.76
3rd-Level	Clustering coefficient	56.28	60.57	53.69	61.84

TABLE V
PERFORMANCE OF DIFFERENT LEVEL OF FEATURE FOR AD VS. NC(%)

Method	ACC	SEN	SPE	AUC
1st-Level	**87.87**	**88.17**	**88.09**	**95.59**
2nd-Level	87.11	86.37	87.80	94.30
3rd-Level	82.49	82.77	82.93	91.41
1st & 2nd & 3rd	89.09	89.60	88.49	95.38

in Table II, Table III and Table IV for AD vs. NC, MCI vs. NC and pMCI vs. sMCI, respectively. It can be seen that the 1st-Level feature (either the mean intensity or the standard deviation) outperforms the other 2 levels of features for all the 3 kinds of classifications. Even though it doesn't give the best result in classifying AD from NC, the difference from the best one ($\mathbf{w}^m$) is small in terms of ACC and AUC, about 1.39% and 0.04%, respectively. Furthermore, the SPE of the feature standard deviation (belongs to 1st-Level feature) is the highest. The graph metric, strength, which belongs to the 3rd-Level feature is inferior among all the types of features in AD diagnosis and MCI diagnosis.

C. Feature Concatenation Evaluation

In this part, the evaluation for different levels of features are given. Different types of features within the same level are concatenated to a long vector and the results are shown in Table V to Table VII (the first 3 lines). As can be seen, among all the 3 levels of features, the 1st-Level feature is still superior to other features in three tasks. In addition, it can be seen from Table II and Table V (AD diagnosis) that concatenation of two types of

TABLE VI
PERFORMANCE OF DIFFERENT LEVEL OF FEATURE FOR MCI VS. NC(%)

Method	ACC	SEN	SPE	AUC
1st-Level	**77.14**	75.04	**79.71**	**84.30**
2nd-Level	76.42	**75.29**	78.66	83.76
3rd-Level	71.92	73.35	71.51	80.43
1st & 2nd & 3rd	77.39	76.19	78.23	83.42

TABLE VII
PERFORMANCE OF DIFFERENT LEVEL OF FEATURE FOR PMCI VS. SMCI(%)

Method	ACC	SEN	SPE	AUC
1st-Level	**58.26**	**60.25**	**59.48**	**64.42**
2nd-Level	55.17	54.87	57.07	58.14
3rd-Level	55.81	57.62	57.41	54.85
1st & 2nd & 3rd	53.38	52.91	56.42	57.46

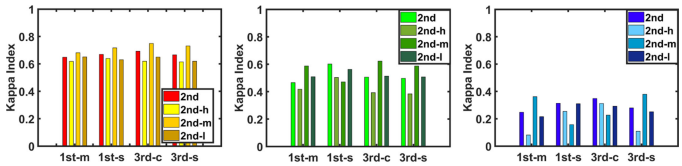

Fig. 5. Performance evaluation of the similarity-driven ranking method. (a) AD vs. NC. (b) MCI vs. NC. (c) pMCI vs. sMCI.

1st-Level features can improve the performance of AD classification, and increase by about 2.38% (ACC), 1.56% (SEN), 1.85% (SPE), 1.75% (AUC). Concatenation of 2nd-Level features also has some improvements, but concatenation of 3rd-Level features has an inverse effect and all the four metrics are lower than the results obtained using the optimal sub-feature (clustering coefficient) in 3rd-Level. In MCI diagnosis, only concatenation of 2nd-Level features improves the classification effectiveness. But in classifying pMCI from sMCI, concatenation of sub-features within the same level cannot improve the performance. In addition, the performances of concatenating all the 3 levels of features are also shown in Table V to Table VII (the last line). It can be seen that there is a significant improvement only for AD diagnosis, and for MCI diagnosis, the improvement is small. For pMCI vs. sMCI, concatenation of 3 levels of features fails to improve the performance. It is because that those added features may be effective, or may be redundant. Therefore, the strategy of concatenating features is not an effective method to improve the classification performance for the all 3 tasks.

D. Effectiveness of the Similarity-Driven Ranking Method

The similarity-driven ranking method can not only reduce the 2nd-Level feature's dimension, but also improve the classifier's diversity. Here, Kappa index [29] is applied to measure the diversity and a small value indicates a high diversity, which is computed through:

$$Ka(i,j) = \frac{p_1 - p_2}{1 - p_2} \tag{20}$$

where p_1 denotes the observed agreement of $\mathbf{y_i}$ and $\mathbf{y_j}$, and p_2 stands for the chance agreement. Fig. 5 shows the effectiveness

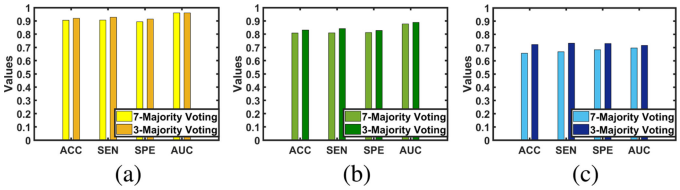

Fig. 6. Performance evaluation of the ensemble classification. (a) AD vs. NC. (b) MCI vs. NC. (c) pMCI vs. sMCI.

the proposed ranking method on the improvement of classifier's diversity, where '2nd' denotes the original 2nd-Level feature, '2nd-h', '2nd-m' and '2nd-l' denote the decomposed 3 subsets of features and '1st-m', '1st-s', '3rd-c' and '3rd-s' denote the mean intensity, standard deviation (1st-Level feature), clustering coefficient and strength (3rd-Level feature), respectively. As can be seen, the decomposed features can achieve a higher diversity (a smaller value) than the original 2nd-Level feature for all the 3 tasks, especially for classification of pMCI. The higher diversity benefited from the similarity-driven ranking method can ensure the ensemble classifier has a good performance.

E. Ensemble Classification Evaluation

The increase of the number of classifiers and their diversities can improve the performance of the ensemble classifier in theory. Obviously, the maximum number of classifiers (7 classifiers) is fixed in this paper. If the sub-classifiers do not perform well and all of them are used to do the final decision through majority voting, there will be a high probability that misclassified results accounted for the majority. In order to avoid this situation and enhance the ensemble effect, a strategy of selecting models with high diversity is proposed. In this experiment, we compare majority voting using outputs from all the 7 SVMs (noted as 7-Majority Voting) with the proposed method which using 3 selected SVMs' decisions (noted as 3-Majority Voting), and the results are shown in Fig. 6. It can be seen that the proposed method outperforms the 7-Majority Voting, specifically, it improves by 1.42% (ACC), 2.20% (SEN), 2.00% (SPE), and 0.03% (AUC) in AD diagnosis and 1.42% (ACC), 3.22% (SEN), 0.67% (SPE), -0.55% (AUC) in MCI diagnosis. For pMCI vs. sMCI, the proposed method increases by 6.64% (ACC), 6.44% (SEN), 4.71% (SPE), and 1.93% (AUC). Clearly, the proposed method shows an effective improvement for classifying pMCI from sMCI. It is because that a single type of feature in the classification of pMCI does not perform well, and the highest accuracy is only 59.85% (Table IV). The probability that misclassified results dominate the majority voting will be high, if considering all the 7 classifiers' outputs. And another reason is that the improvement of performance in classifying pMCI from sMCI benefits from the increase of diversity brought by the decomposition of 2nd-Level feature.

F. Comparison With the State-of-the-Art Methods

We also compare the classification performance of the proposed method with the state-of-the-art methods, including Hinrichs's method [14], Gray's method [16], Li's method [17], Padilla's method [31], which are designed on FDG-PET data and use classical machine learning techniques. The results are shown in Table VIII to Table X. It can be seen that our method outperforms the other methods regarding MCI diagnosis and classifying pMCI from sMCI. For AD vs. NC, the proposed method is superior to the compared method in terms of ACC and SEN. The difference with the best result in respect of SPE is 2.22%, and for AUC, it is 1.02%. But our method is inferior to Lu's method [32], which uses deep neural network and reports outstanding results in AD diagnosis and pMCI diagnosis, 93.85% (ACC) and 82.51% (ACC), respectively.

TABLE VIII
PERFORMANCE COMPARISON FOR AD VS. NC(%)

Method	Subjects	ACC	SEN	SPE	AUC
Hinrichs *et al.*[14]	89AD+94NC	84	84	82	87.16
Gray *et al.*[16]	50AD+54NC	88.4	83.2	**93.6**	−−
Li *et al.*[17]	25AD+30NC	89.1	92	86	**97**
Padilla *et al.*[29]	53AD+52NC	86.59	87.50	85.36	−−
Our method	94AD+90NC	**91.90**	**92.78**	91.38	95.98

TABLE IX
PERFORMANCE COMPARISON FOR MCI VS. NC(%)

Method	Subjects	ACC	SEN	SPE	AUC
Gray *et al.*[16]	53pMCI+54NC	81.3	79.8	**82.9**	−−
Li *et al.*[17]	29MCI+30 NC	63.2	65	62	72
Our method	88MCI+90NC	**83.18**	**84.20**	82.83	**88.93**

TABLE X
PERFORMANCE COMPARISON FOR PMCI VS. SMCI (%)

Method	Subjects	ACC	SEN	SPE	AUC
Gray *et al.*[16]	53pMCI+64sMCI	63.1	52.2	**73.2**	−−
Our method	44pMCI+44sMCI	**72.33**	**73.27**	73.11	**71.66**

IV. CONCLUSION

AD and MCI diagnoses under FDG-PET single modality are challenging. In this paper, a novel ensemble method which uses multi-level features is proposed to address the problem. First, 3 levels of features that represent properties of ROIs and their connectivities are extracted gradually. Then a proposed similarity-driven ranking method is applied to decompose the 2nd-Level feature to 3 different sets of features, which reduces the feature dimension to a great extent and increases the classifier's diversity. Next, different models are trained by using different types of features. In order to enhance the ensemble effect, a pair of models with high diversity are selected through the proposed mMsE method and another model with high accuracy is chosen by nested cross validation. The final decision is made through the majority voting of the 3 selected models' outputs. According to experiments on the public dataset (ADNI), the proposed method can improve the performance of AD and MCI diagnoses and especially classifying pMCI from sMCI when compared with those state-of-the-art methods developed by using classical machine learning techniques, but our approach does not outperform the deep learning based methods, which will be included in our future work.

ACKNOWLEDGMENT

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

REFERENCES

[1] World Health Organization, "Dementia: A public health priority," Geneva, World Health Organization, 2012.

[2] G. M. McKhann *et al.*, "The diagnosis of dementia due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association Workgroups on diagnostic guidelines for Alzheimer's disease," *Alzheimers Dementia*, vol. 7, no. 3, pp. 263–269, 2011.

[3] M. S. Albert *et al.*, "The diagnosis of mild cognitive impairment due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association Workgroups on diagnostic guidelines for Alzheimer's disease," *Alzheimers Dementia*, vol. 7, no. 3, pp. 270–279, 2011.

[4] R. A. Sperling *et al.*, "Toward defining the preclinical stages of Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association Workgroups on diagnostic guidelines for Alzheimer's disease," *Alzheimers Dementia*, vol. 7, no. 3, pp. 280–292, 2011.

[5] C. R. Jack Jr *et al.*, "Introduction to revised criteria for the diagnosis of Alzheimer's disease: National Institute on Aging and the Alzheimer Association Workgroups," *Alzheimers Dementia*, vol. 7, no. 3, pp. 257–262, 2011.

[6] L. K. Ferreira *et al.*, "Support vector machine-based classification of neuroimages in Alzheimer's disease: Direct comparison of FDG-PET, rCBF-SPECT and MRI data acquired from the same individuals," *Revista Brasileira de Psiquiatria*, vol. 40, pp. 181–191, 2017.

[7] L. Mosconi, V. Berti, L. Glodzik, A. Pupi, S. De Santi, and M. J. Leon, "Pre-clinical detection of Alzheimer's disease using FDG-PET, with or without amyloid imaging." *J. Alzheimer's Disease*, vol. 20, no. 3, pp. 843–854, 2010.

[8] S. Rathore, M. Habes, M. A. Iftikhar, A. Shacklett, and C. Davatzikos, "A review on neuroimaging-based classification studies and associated feature extraction methods for Alzheimer's disease and its prodromal stages," *NeuroImage*, vol. 155, pp. 530–548, 2017.

[9] M. R. Arbabshirani, S. Plis, J. Sui, and V. D. Calhoun, "Single subject prediction of brain disorders in neuroimaging: Promises and pitfalls," *NeuroImage*, vol. 145, pp. 137–165, 2017.

[10] K. B. Walhovd *et al.*, "Multi-modal imaging predicts memory performance in normal aging and cognitive decline," *Neurobiol. Aging*, vol. 31, no., 7, pp. 1107–1121, 2010.

[11] S. M. Landau *et al.*, "Comparing predictors of conversion and decline in mild cognitive impairment," *Neurology*, vol. 75, no. 3, pp. 230–238, 2010.

[12] Y. Shi *et al.*, "Joint coupled-feature representation and coupled boosting for AD diagnosis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 2721–2728.

[13] S. Liu *et al.*, "Multimodal neuroimaging feature learning for multiclass diagnosis of Alzheimer's disease," *IEEE Trans. Biomed. Eng.*, vol. 62, no. 4, pp. 1132–1140, Apr. 2015.

[14] C. Hinrichs, V. Singh, L. Mukherjee, G. Xu, M. K. Chung, and S. C. Jhonson, "Spatially augmented LPboosting for AD classification with evaluations on the ADNI dataset," *NeuroImage*, vol. 48, no. 1, pp. 138–149, 2009.

[15] C. Cabral, P. M. Morgado, D. C. Costa, and M. Silveira, "Predicting conversion from MCI to AD with FDG-PET brain images at different prodromal stages," *Comput. Biol. Med.*, vol. 58, pp. 101–109, 2015.

[16] K. R. Gray, R. Wolz, R. A. Heckemann, P. Aljabar, A. Hammers, and D. Reuckert, "Multi-region analysis of longitudinal FDG-PET for the classification of Alzheimer's disease," *NeuroImage*, vol. 60, no. 1, pp. 221–229, 2012.

[17] R. Li *et al.*, "Gaussian mixture models and model selection for [18F] fluorodeoxyglucose positron emission tomography classification in Alzheimer's disease," *PLoS One*, vol. 10, no. 4, 2015, Art. no. e0122731.

[18] M. Pagani *et al.*, "Volume of interest-based [^{18}F] fluorodeoxyglucose PET discriminates MCI converting to Alzheimer's disease from healthy controls. A European Alzheimer's Disease Consortium (EADC) study," *NeuroImage: Clin.*, vol. 7, pp. 34–42, 2015.

[19] K. L. Arnemann, F. Stoeber, S. Narayan, G. D. Rabinovici, and W. J. Jagust, "Metabolic brain networks in aging and preclinical Alzheimer's disease," *NeuroImage: Clin.*, vol. 17, pp. 987–999, 2017.

[20] I. Yakushev, A. Drzezga, and C. Habeck, "Metabolic connectivity: Methods and applications," *Current Opinion Neurol.*, vol. 30, no. 6, pp. 677–685, 2017.

[21] N. Tzourio-Mazoyer *et al.*, "Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain," *NeuroImage*, vol. 15, no. 1, pp. 273–289, 2002.

[22] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Statist. Soc. B*, vol. 58, no. 1, pp. 267–288, 1996.

[23] Alzheimers Disease Neuroimaging Initiative. [Online]. Available: http://adni.loni.usc.edu/methods/pet-analysis/pre-processing/

[24] Statistical Parametric Mapping. [Online]. Available: http://www.fil.ion.ucl.ac.uk/spm/software/

[25] I. Garali, M. Adel, S. Bourennane, and E. Guedj, "Brain region ranking for 18FDG-PET computer-aided diagnosis of Alzheimer's disease," *Biomed. Signal Process. Control*, vol. 27, pp. 15–23, 2016.

[26] M. Xia, J. Wang, and Y. He, "BrainNet viewer: A network visualization tool for human brain connectomics," *PLoS One*, vol. 8, no. 7, 2013, Art. no. e68910.

[27] M. Rubinov and O. Sporns, "Complex network measures of brain connectivity: Uses and interpretations," *NeuroImage*, vol. 52, no. 3, pp. 1059–1069, 2010.

[28] J. Chung, K. Yoo, E. Kim, D. L. Na, and Y. Jeong, "Glucose metabolic brain networks in early-onset vs. late-onset Alzheimer's disease," *Frontiers Aging Neurosci.*, vol. 8, p. 159, 2016.

[29] M. Liu, D. Zhang, and D. Shen, "Relationship induced multi-template learning for diagnosis of Alzheimer's disease and mild cognitive impairment," *IEEE Trans. Med. Imag.*, vol. 35, no. 6, pp. 1463–1474, Jun. 2016.

[30] J. Zhang, M. Liu, L. An, Y. Gao, and D. Shen, "Alzheimer's disease diagnosis using landmark-based features from longitudinal structural MR images," *IEEE J. Biomed. Health Inform.*, vol. 21, no. 6, pp. 1607–1616, Nov. 2017.

[31] P. Padilla, M. López, J. M. Górriz, R. Ramírez, D. Salas-González, and I. Álvarez, "NMF-SVM based CAD tool applied to functional brain images for the diagnosis of Alzheimer's disease," *IEEE Trans. Med. Imag.*, vol. 31, no. 2, pp. 207–216, Feb. 2012.

[32] D. Lu, K. Popuri, G. W. Ding, R. Balachandar, and M. F. Beg, "Multiscale deep neural network based analysis of FDG-PET images for the early diagnosis of Alzheimers disease," *Med. Image Anal.*, vol. 46, pp. 26–34, 2018.