HHS Public Access

Author manuscript

IEEE J Biomed Health Inform. Author manuscript; available in PMC 2018 November 16.

Published in final edited form as:

IEEE J Biomed Health Inform. 2018 September ; 22(5): 1476–1485. doi:10.1109/JBHI.2018.2791863.

Anatomical Landmark Based Deep Feature Representation for MR Images in Brain Disease Diagnosis

Mingxia Liu[#],

Department of Radiology and BRIC, University of North Carolina, Chapel Hill, NC 27599, USA.

Jun Zhang[#],

Department of Radiology and BRIC, University of North Carolina, Chapel Hill, NC 27599, USA.

Dong Nie,

Department of Radiology and BRIC, University of North Carolina, Chapel Hill, NC 27599, USA.

Pew-Thian Yap, and

Department of Radiology and BRIC, University of North Carolina, Chapel Hill, NC 27599, USA.

Dinggang Shen [Fellow, IEEE]

Department of Radiology and BRIC, University of North Carolina, Chapel Hill, NC 27599, USA, and also with the Department of Brain and Cognitive Engineering, Korea University, Seoul 02841, South Korea.

[#] These authors contributed equally to this work.

Abstract

Most automated techniques for brain disease diagnosis utilize hand-crafted (e.g., voxel-based or region-based) biomarkers from structural magnetic resonance (MR) images as feature representations. However, these hand-crafted features are usually high-dimensional or require regions-of-interest defined by experts. Also, because of possibly heterogeneous property between the hand-crafted features and the subsequent model, existing methods may lead to sub-optimal performances in brain disease diagnosis. In this paper, we propose a landmark-based deep feature learning (LDFL) framework to automatically extract patch-based representation from MRI for automatic diagnosis of Alzheimer's disease. We first identify discriminative anatomical landmarks from MR images in a data-driven manner, and then propose a convolutional neural network for patch-based deep feature learning. We have evaluated the proposed method on subjects from three public datasets, including the Alzheimer's disease neuroimaging initiative (ADNI-1), ADNI-2, and the minimal interval resonance imaging in alzheimer's disease (MIRIAD) dataset. Experimental results of both tasks of brain disease classification and MR image retrieval demonstrate that the proposed LDFL method improves the performance of disease classification and MR image retrieval.

Corresponding author: Dinggang Shen.

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org.

Keywords

Anatomical landmarks; convolutional neural network; classification; image retrieval; brain disease diagnosis

I. INTRODUCTION

ALZHEIMER'S disease (AD) is an increasingly prevalent disease, characterized by the accumulation of amyloid-β (Aβ) and hyperphosphorylated tau in the brain that eventually leads to neurodegeneration [1]. The accurate diagnosis of AD is of great importance for possible improvement in the treatment of the disease and is expected to help reduce costs associated with long-term care for patients. To support AD diagnosis, many computer-aided approaches have been proposed using various biomarkers. Compared with the accumulation of Aβ detected in cerebrospinal fluid (CSF) or by using positron emission tomography (PET) [2], [3], biomarkers based on structural magnetic resonance imaging (MRI) could suggest structural changes of the brain in a more sensitive manner [4], [5].

Currently, many global and relatively local biomarkers (shown in Fig. 1) have been proposed for AD diagnosis with MRI data, including global region-of-interest (ROI) based volumetric measures and local high-dimensional-morphological-analysis (HDMA) based measures. Specifically, ROI-based measures (e.g., cortical thickness [6]–[10], hippocampal volume [11]–[13], and gray matter volume [14], [15]) are traditionally adopted to measure regionally anatomical volume and to investigate abnormal tissue structures in the brain. However, the definition of ROIs usually requires *a priori* hypothesis on the abnormal regions from a structural/functional perspective, requiring expert knowledge in practice [16]. Also, an abnormal region might be only a small part of a pre-defined ROI or span over multiple ROIs, thereby leading to loss of discriminative information. In addition, ROI-based measurements depend largely on two time-consuming steps that reduce the feasibility of timely AD diagnosis: 1) non-linear registration across subjects, and 2) brain tissue segmentation [17]. As an alternative solution, HDMA-based measures capture localized structural changes in a hypothesis-free manner to quantify brain atrophy, among which voxel-based morphometry (VBM) [15], [18]–[20], deformation-based morphometry (DBM) [21], and tensor-based morphometry (TBM) [22] are the typical examples. Specifically, VBM directly measures local tissue (e.g., gray matter, white matter and cerebrospinal fluid) density of a brain via voxel-wise analysis, DBM detects morphological differences from non-linear deformation fields that align/warp images to a common anatomical template, and TBM identifies regionally structural differences from local Jacobians of deformation fields, respectively. While the number of subjects is often limited (e.g., in hundreds), HDMA-based measurement is generally of very high dimension (e.g., in millions), leading to the over-fitting problem in subsequent learning models [23]. Also, the HDMA-based measure is usually limited by registration errors or inter-subject anatomical variations. Therefore, it is desirable to extract discriminative biomarkers from MRI in a semi-global manner, independent of any pre-defined ROIs and time-consuming pre-processing procedures, which may result in better performance in brain disease diagnosis.

In this study, we propose an anatomical landmark based deep feature learning (LDFL) framework for AD diagnosis (see Fig. 2), by extracting patch-based measures from structural MR imaging data. Different from conventional ROI-based and voxel-based feature representations of MR images, we develop a novel patch-based feature extraction method for computer-aided brain disease diagnosis with MRI data. Specifically, we first identify discriminative anatomical landmarks via group comparison between AD and normal control (NC) subjects, and then learn patch-based feature representations from each landmark location via a deep convolutional neural network (CNN) [24]. The effectiveness of the proposed LDFL method is validated in both tasks of brain disease classification and MR image retrieval. The major contributions of this work can be summarized as follows. *First*, the proposed feature representation can be automatically learned from MR imaging data, without using pre-defined ROIs and time-consuming pre-processing procedures (e.g., brain tissue segmentation). *Second*, we propose to locate the most informative image patches from MRI with millions of patches based on anatomical landmarks. *Third*, we use the Alzheimer's Disease Neuroimaging Initiative (i.e., ADNI-1) [25] as the training set, and ADNI-2 and MIRIAD as *independent* testing sets.

II. MATERIALS AND METHODS

A. Data Preparation

Our analysis is based on three public datasets, including the Alzheimer's Disease Neuroimaging Initiative (ADNI-1) dataset [25], ADNI-2, and the Minimal Interval Resonance Imaging in Alzheimer's Disease (MIRIAD) dataset.[1] There are a total of 199 AD and 229 NC subjects with 1.5T T1-weighted structural MRI data in the baseline ADNI-1 dataset, whereas the baseline ADNI-2 dataset contains 159 AD and 200 NC subjects with 3T T1-weighted structural MRI data. Note that in our experiments, several subjects that appear in both ADNI-1 and ADNI-2 are removed from ADNI-2 for independent testing. The general inclusion/exclusion criteria used by ADNI-1 are summarized as follows: 1) NC subjects: Mini-Mental State Examination (MMSE) scores between 24–30 (inclusive), a Clinical Dementia Rating (CDR) of 0, non-depressed, non-MCI and non-demented; 2) mild AD: MMSE scores between 20–26 (inclusive), CDR of 0.5 or 1.0 and meets NINCDS/ADRDA criteria for probable AD. There are 23 NCs and 46 AD patients with baseline 1.5T T1-weighted MRI in MIRIAD. In the main experiments, ADNI-1 is used as the *training set*, while ADNI-2 and MIRIAD are adopted as *independent testing sets*. The demographic and clinical information of subjects is reported in Table I.

In this study, we process all MR images using a standard pipeline. More specifically, we first perform anterior commis-sure (AC)-posterior commissure (PC) correction for each MR image, using the MIPAV software package. We then re-sample each MR image to have a resolution of 256 × 256 × 256, followed by intensity inhomogeneity correction for images using the N3 algorithm [26]. Also, we perform skull stripping [27] for all MR images, as well as a process of manual editing, to ensure that both skull and dura are cleanly removed.

[1] http://www.ucl.ac.uk/drc/research/methods/miriad-scan-database.

We finally remove the cerebellum from each MR image, by warping a labeled template to each skull-stripped image.

B. Method

Fig. 2 illustrates a general framework of our proposed LDFL method, including four main components: 1) landmark discovery, 2) landmark-based patch extraction, 3) patch-based feature learning, and 4) applications of disease classification and image retrieval. For landmark discovery, we first identify AD landmarks that have statistically significant differences between AD and NC subjects in the training set and then apply a pre-trained landmark detection model to automatically detect landmarks in each testing image [28]. In the stage of landmark-based patch extraction, we extract multiple patches from each training image based on each of multiple anatomical landmarks. In the patch-based feature learning stage, we develop a CNN model to learn morphological features, where patches extracted from each landmark are used as input, and subject-level labels for patches are adopted as output. Note that each CNN model is corresponding to a specific landmark position. Finally, the patch-based deep feature representations learned from multiple landmarks are fed into subsequent disease classification and image retrieval models.

1) Anatomical Landmark Discovery: Our goal is to identify regions that have group differences in local brain structures between AD patients and normal controls (NCs). Following previous studies [28], [29], we performed a voxel-wise group comparison between AD and NC groups in ADNI-1. Specifically, we first linear-aligned all training images to the Colin27 template [30] to remove global translation, as well as the scale and rotation differences of MR images. Also, we re-sampled all images to have the same spatial resolution (i.e., $1 \times 1 \times 1$ mm^3). Then, we non-linearly aligned all training images to the template, to build the correspondence among voxels from different images, using the HAMMER algorithm [31]. By using the deformation field from non-linear registration, we can establish the correspondence between each voxel in the template and those in the linearly-aligned images. For each voxel in the template, we extracted two groups of morphological features (i.e., local energy pattern [32]) from corresponding voxels in all training images from the group of AD patients and the group of NCs, respectively. Then, based on the morphological features for each voxel, we performed a multivariate statistical test (i.e., Hotelling's T2 [33]) on AD and NC groups, and thus can obtain a p-value for each voxel in the template space. Given all voxels in the template, we generated a p-value map corresponding to the template. Finally, the local minima from the p-value map were identified as locations of discriminative anatomical landmarks in the template space. Finally, we projected these landmark locations to all linearly-aligned training images using their respective deformation fields (generated in the non-linear registration).

To fast locate anatomical landmarks for testing images, we further trained a regression forest based landmark detector, with landmark as output and those linearly aligned training images as input. For a new testing MR image, we can align it linearly to the template space and then use our trained landmark detector to identify landmarks in the linearly-aligned testing image. In this way, both training images and testing images would have the same landmarks. For

each MR image, there are approximately 1700 landmarks identified from AD and NC subjects in the ADNI-1 dataset, shown in Fig. 3(a).

From Fig. 3(a), we can observe that some landmarks are very close to each other. In such a case, directly using those landmark for patch extraction may lead to overlapped patches. To address this issue, for all identified landmarks ranked according to *p*-values in descending manner, we further define a spatial Euclidean distance threshold (i.e., 16) to control the distance between landmarks, to reduce the overlaps among image patches. Finally, we select the top 50 landmarks for deep feature learning [34], and show these landmarks in Fig. 3(b), Fig. S2 and Movie S1 of the *Supplementary Materials*.

2) Landmark-Based Patch Extraction: To suppress the negative influence of landmark localization errors introduced by image registration, we sample multiple image patches (with the size of 19 × 19 × 19) from each landmark location with displacements in a 3 × 3 × 3 cubic. Hence, there are 27 image patches extracted from a specific landmark position in an MR image of a specific subject. The class label of each image patch is assigned as the same label of that subject (i.e., subject-level label). The influence of parameters (i.e., the number of landmarks and size of patches) is shown in Fig. S1 in *Supplementary Materials*.

3) Patch-Based Feature Learning: We develop a CNN model [35] to extract discriminative patch-based biomarkers from MRI for AD diagnosis, with a schematic diagram shown in Fig. 4. As shown in Fig. 4, the input of the network include multiple image patches extracted from a specific landmark position in brain MR images, which are convolved by a series of 5 convolutional layers (i.e., Conv1, Conv2, Conv4, Conv5, and Conv7) with rectified linear unit (ReLU) activation. Here, Conv2, Conv5, and Conv7 are followed by max-pooling procedures to perform a down-sampling operation for their outputs, respectively. The size of a 3D kernel in each convolutional layer is 3 × 3 × 3. The resulting 256-dimensional features, which are equal to the number of feature maps of the last convolutional layer (i.e., Conv7), are fed into a series of 3 fully connected (FC) layers (i.e., FC9, FC10, and FC11) with dimensions of 256, 128, and 2, since 2 is the number of classes considered. The use of FC layers accelerates convergence, while the problem of over-fitting could be partly solved by adding a drop-out layer (with a ratio of 0.5) before FC10. The output of the last FC layer (i.e., FC11) is fed into a soft-max top-most output layer, to predict the probability of an input image patch belonging to AD or NC group. Note that the information given by the subject-level class labels are used in a back-propagation procedure.

Denote L as the number of landmarks, and the training set as $\mathcal{X} = \{\boldsymbol{X}_n\}_{n=1}^{N}$ that contains N subjects with the corresponding labels $\mathbf{y} = \{y_n\}_{n=1}^{N}$. We denote the n-th training image as $\mathrm{X}_n = \{\mathrm{X}_{n,1}, \mathrm{X}_{n,2}, \ldots, \mathrm{X}_{n,L}\}$ containing L image patches. As shown in Fig. 2, patches extracted from training images are the basic training samples for our proposed CNN model, and the labels of those patches are subject-level labels. That is, the subject-level label information (i.e., $y_n (n = 1, \ldots, N)$) is used as the supervision information for network training. More specifically, the proposed CNN aims to learn a mapping function Φ: $X \rightarrow \mathbf{y}$. For the l-th

CNN model corresponding to the l-th ($l = 1, \ldots, L$) landmark, the objective function is shown as follows:

$$\min_{\mathbf{W}_l} \sum_{\{\mathbf{x}_{n,l} \in \mathbf{X}_n\}_{n=1}^{N}} -\log\left(\mathbf{P}\left(y_n \middle| \mathbf{x}_{n,l}; \mathbf{W}_l\right)\right) \quad (1)$$

where $\mathbf{P}(y_n|\mathrm{x}_{n,l};\mathbf{W}_l)$ indicates the probability of the patch $\mathrm{x}_{n,l}$ being correctly classified as the class y_n using the network coefficients W_l.

The implementation of the proposed CNN model is based on Caffe [36], and the computer we used in the experiments contains a single GPU (i.e., NVIDIA GTX TITAN 12GB). The network is optimized by stochastic gradient descent (SGD) algorithm [37] with a momentum coefficient of 0.9. The learning rate is set to 10^{-2}, and the weight updates are performed in mini-batches of 30 samples per batch. Here, we further randomly select 10% subjects in ADNI-1 as validation data, while the remaining in ADNI-1 are regarded as the training data. The training process ends when the network does not significantly improve its performance on the validation set within 60 epochs.

4) Applications of Disease Classification and MR Image Retrieval: Given L (L = 50 in this study) anatomical landmarks, we can train L CNN models, and each model is corresponding to a specific landmark. In this way, for each MR image, we can obtain L feature vectors via those CNN models. For brain disease classification, we simply combine the estimated patch labels in 50 landmark locations achieved by CNNs using the majority voting strategy [38]. For the task of MR image retrieval, we use the outputs from three FC layers (i.e., FC9, FC10, and FC11) of each CNN as patch-based feature representations, and compare them with conventional representations for MRI in the experiments.

III. EXPERIMENTS

A. Methods for Comparison

We compare our proposed LDFL method with three state-of-the-art feature representations for MR images, including 1) ROI-based representation [10], [39], 2) voxel-based morphometry (VBM) representation [18], 3) landmark-based morphology (LMF) representation [28]. We briefly introduce these three methods as well as our method in the following.

1. *ROI-based (ROI) representation:* Following previous studies [10], [39], we extract ROI-based features from MR images in this method. To be specific, we first segment the studied MR image into three tissue types, including gray matter (GM), white matter (WM), and cerebrospinal fluid (CSF), by using the FAST algorithm in the FSL software. Then, we align the anatomical automatic labeling (AAL) atlas [40] with (90 pre-defined ROIs in the cerebrum) to the native space of each subject using a deformable registration algorithm (i.e., HAMMER [31]). We finally extract the volumes of GM tissue in 90 ROIs as feature representation for each MR image. Note that the volumes of GM tissue are normalized by the

total intracranial volume (estimated by the summation of GM, WM, and CSF volumes). In the classification task, we feed the ROI features into a linear support vector machine (LSVM) classifier [41] for disease classification. In the image retrieval task, we adopt these ROI features to represent each MRI.

2. *Voxel-based morphometry (VBM) representation [18]:* Similar to the pre-processing procedure used in the ROI-based method, we first normalize the studied MR images using the same registration and segmentation methods. Then, we extract the local tissue density of GM in a voxel-wise manner as feature representations for an MR images. To reduce the feature dimension, we further perform the *t*-test algorithm to select the most informative features. Those selected voxel-based features are finally fed into an LSVM for classification and used as the representation for each MRI in the task of image retrieval.

3. *Landmark based morphology (LMF) representation [28]* with engineered feature representations. In LMF, we first extract morphological features (i.e., local energy pattern [32]) from each local image patch centered at each landmark location, and then concatenates these fea tures extracted from multiple landmarks, followed by a *z*-score normalization [42] process. Finally, the normalized features are used in both tasks of disease classification (via LSVM) and image retrieval. As a landmark-based method, LMF shares the same landmark pool as our proposed LDFL method, shown in Fig. 3. It is worth noting that, different from our proposed LDFL approach that learns features automatically from MRI, LMF adopts human-engineered features for representing image patches around each landmark position.

4. *Proposed patch-based representation:* For each MR image, we can obtain L feature vectors via the proposed CNN model in Fig. 4. We then extract three types of features as the representation for MRI, including 1) features from FC9 layer in the proposed CNN that is denoted as *FC9* for short, 2) features from FC10 layer (denoted as *FC10*), and 3) features from FC11 layer (denoted as *FC11*). Similar to LMF method, we can simply concatenate the feature vectors from L landmarks for subsequent model learning.

In the disease classification task, we feed those ROI and VBM features to an LSVM classifier, respectively. It is worth noting that, for landmark-based features (i.e., LMF [28], FC9, FC10, and FC11), we further propose two strategies to utilize these features, including 1) *feature concatenation* and 2) *classifier ensemble*. Specifically, in feature concatenation methods (denoted as LMF_con, FC9_con, FC10_con, and FC11_con), features learned from 50 landmarks are simply concatenated into a long feature vector, followed by an LSVM classifier. In ensemble based methods (denoted as LMF_ens, FC9_ens, FC10_ens, and FC11_ens), we first train an LSVM using features obtained from each landmark position and can obtain 50 LSVMs given 50 landmarks. Then, the results of those LSVMs are combined by a majority voting strategy [38] for final classification. In the image retrieval task, we compare the proposed three types of patch-based feature representations via feature concatenation (i.e., FC9_con, FC10_con, and FC11_con) with three conventional representations for MRI (i.e., ROI, VBM, and LMF_con). Specifically, we first represent

each MR image (*w.r.t.* a particular subject) using a specific feature vector, and then compute the Euclidean distance between a new query MR image in ADNI-2 and each of MRI in ADNI-1.

B. Experimental Settings

The performance of disease classification is evaluated by the following metrics: 1) classification accuracy (ACC), 2) sensitivity (SEN), 3) specificity (SPE), 4) receiver operating characteristic curve (ROC), 5) area under ROC (AUC), and 6) F-Measure [44]. We denote TP, TN, FP, FN and PPV as true positive, true negative, false positive, false negative, and positive predictive value, respectively. These evaluation metrics are defined as: $\mathrm{ACC} = \frac{(\mathrm{TP}+\mathrm{TN})}{(\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN})}$, $\mathrm{SEN} = \frac{\mathrm{TP}}{(\mathrm{TP}+\mathrm{FN})}$, $\mathrm{SPE} = \frac{\mathrm{TN}}{(\mathrm{TN}+\mathrm{FP})}$, $\text{F-Measure} = \frac{(2\times SEN\times \mathrm{PPV})}{(\mathrm{SEN}+\mathrm{PPV})}$ where $\mathrm{PPV} = \frac{\mathrm{TP}}{(\mathrm{TP}+\mathrm{FP})}$. In image retrieval experiments, we utilize four metrics for performance evaluation, including 1) mean average precision (MAP) for a set of queries that is the mean of average precision scores for each query (where each subject with MR image in ADNI-2 is used as a specific query), 2) F-Measure, 3) Matthews correlation coefficient (MCC) [45] that is a balanced measure for binary classes, and 4) #Correct@K that is the number of correct results in top K returned results. The most relevant results should be ranked at top-most positions resulting in higher #Correct@K values.

C. Results

In this section, we first show the learned features and the learned kernels by the proposed LDFL method. Then, we report the experimental results in both tasks of disease classification and MR image retrieval, by comparing the proposed method with those competing methods.

1) Learned Feature Representations: It is worth noting that each layer of CNN combines the extracted low layer feature maps to learn higher level features at the next layer, in a hierarchical manner for describing more abstract anatomical variations of a brain. To analyze the discriminative capability of the patch-based features learned from our method, we visualize the learned features at 5 convolutional and 3 fully connected layers in CNN, by projecting those features down to 2 dimensions using the t-SNE dimension reduction algorithm [43]. As can be seen from Fig. 5(a)–(e), the convolutional layers (i.e., Conv1, Conv2, Conv4, Conv5, and Covn7) gradually enhance the discriminative power between AD and NC subjects along the hierarchy. Also, Fig. 5(f)–(h) indicate that the subsequent task-specific classification layers (i.e., FC9, FC10, and FC11) further enhance the separability between AD and NC subjects, and features at the top-most FC layers (FC10, and FC11) are most discriminative.

2) Learned Kernels in CNN: Fig. 6 shows 32 convolutional kernels learned at the Conv1 layer via the proposed CNN model (shown in Fig. 4). From Fig. 6, one may notice the different natures of these kernels in capturing fundamental 3D patterns. These patterns vary complexity while passing through consecutive convolutional layers, so that the last layer could have the ability to describe the structural differences between AD and NC

subjects. In addition, given an input image patch, we show the outputs of each convolutional layer in CNN in Figs. S6–S9 in the *Supplementary Materials*.

3) Results of Disease Classification: In the task of brain disease classification (i.e., AD vs. NC classification), we first learn features from MRI data in a supervised manner via our LDFL framework. In this group of experiments, we treat subjects in ADNI-1 as *training data*, and use subjects in ADNI-2 as *independent testing data*. Fig. 7 reports classification performances achieved by methods using conventional features and our patch-based features, as well as our LDFL method in AD vs. NC classification. As could be seen from Fig. 7, methods using our patch-level deep features usually achieve much better diagnosis results, compared with those using conventional feature representations. For instance, LDFL achieves the best accuracy of 90.56% and the best F-Measure of 89.10%, while ROI, VBM, LMF_con, and LMF_ens generally result in worse performance. On the other hand, ensemble based methods usually perform better than feature concatenation based methods. For instance, FC9_ens achieves an AUC of 94.72%, while the AUC of its counterpart (i.e., FC9_con) is only 91.40%. Also, among three patch-based measures learned from different FC layers in the proposed CNN model, the method using features extracted from the top-most FC layer (i.e., FC11) results in the best result. That is, FC11_ens achieves an accuracy of 88.61% and an AUC of 95.83%. It is worth noting that MR images in the training set (i.e., ADNI-1) were acquired by 1.5T T1-weighted scanners, while those in the testing set (i.e., ADNI-2) were acquired by 3T T1-weighted scanners. Despite the different signal-to-noise ratios of MRI in the training and the testing set, our LDFL method still achieves good results in AD classification. This implies that LDFL has good generalization capability.

In Fig. 8, we further plot the ROC curves achieved by the proposed methods and conventional approaches. As can be seen from Fig. 8, the overall best performances are achieved by the proposed FC11_con and LDFL methods among six feature concatenation based methods and the five ensemble based approaches, respectively. It further demonstrates the effectiveness of the proposed patch-based biomarkers learned from MR imaging data. More results are given in Figs. S4–S6 in the *Supplementary Materials*.

4) Results of MR Image Retrieval: In the task of MR image retrieval, subjects in the ADNI-1 dataset are used as the existing subjects' MRI database, while all the MR images in the ADNI-2 dataset are alternated used as the query image. Fig. 9 reports the performance comparison of methods using different feature representations. Fig. 9(a) implies that methods using our patch-based deep features (i.e., FC9_con, FC10_con, and FC11_con) usually outperform those using conventional features (i.e., ROI, and LMF_con) in MR image retrieval. Also, the best mean average precision (MAP), F-Measure, and Matthews correlation coefficient (MCC) are achieved by the proposed FC11_con, FC11_con, and FC9_con methods, respectively.

From Fig. 9(b), we could observe that features learned from the proposed methods result in better #Correct@K values, compared with ROI and LMF_con. For instance, FC11_con can return 8.27 relevant subjects in top 10 returned results and 38.23 relevant subjects in top 50 returned results. The better performance on #Correct@K indicates that the proposed method can return more relevant medical records. This could be due to the discriminative features

learned from landmark-based CNN models, which can precisely locate possible relevant subjects and then finally provide a fine ranking list. Given an MR image of a query subject represented by the learned patch-based features, it is possible to provide physicians with useful reference from historical data (e.g., existing patients' MRI database) to help design a subject-specific treatment plan.

IV. DISCUSSION

Although extensive studies investigate to extract different feature representations from structural MR imaging data for computer-aided AD diagnosis, most of them focus on global or local measures that require time-consuming pre-processing procedures or depend on pre-defined ROIs, respectively. To this end, we develop a landmark based deep feature learning (LDFL) framework to automatically extract patch-level representations from MR images, based on anatomical landmarks discovered from data via a data-driven algorithm. In particular, we first propose to identify the most discriminative anatomical landmarks via group comparison between AD and NC subjects. Given L (L = 50 in this study) landmarks, we then train L CNN models for patch-based deep feature learning, with each CNN corresponding to a specific landmark. Both disease classification and image retrieval experiments on three public datasets (i.e., ADNI-1, ADNI-2, and MIRIAD) suggest that the proposed method could help promote the performance of computer-aided AD diagnosis.

A. Discriminative Capability of Landmarks

Fig. 10 illustrates the patch classification accuracy in each landmark location achieved by our proposed LDFL method. One could see from Fig. 10 that the accuracy of patches in each CNN model is different from each other. For instance, the patch classification accuracies at the 1st landmark and the 50th landmark positions are 80.80% and 70.87%, respectively. Among those 50 landmarks, the best accuracies (i.e., > 80.00%) are achieved in a subset of landmarks (i.e., 1, 2, 4, 5, 9, 11, 13, 14, 16, 17, 21, 24, 31), implying the structural changes in these landmarks could be more discriminative in distinguishing AD from normal controls (NCs).

From Fig. 3(b), one may observe that landmarks in that subset mainly locate in the areas of bilateral hippocampus, bilateral parahippocampus, and bilateral fusiform. These areas are reported to be related to AD in previous studies [13], [47]–[49]. More results can be found in Figs. S2 and S3 in the *Supplementary Materials*. Besides, Fig. 10 suggests that the overall performance of patch classification gradually becomes worse with the increase of landmark index. The underlying reason could be that the p-values (in the group comparison between AD and NC subjects) of those 50 landmarks gradually increase, and thus their discriminative capabilities become worse in group comparison in the landmark discovery process [28]. Also, the most discriminative features are learned from patches located at the 13 th and the 14 th landmarks, other than from patches at the first landmark location shown in Fig. 3. The main reason is that we use hand-crafted features (i.e., local energy pattern [32]) of MRI to identify anatomical landmarks, while the proposed landmark-based patch-level features are learned automatically from data for AD diagnosis.

B. Diversity Analysis

We adopt a kappa measure [46] to analyze the diversity of classifiers in 5 ensemble-based methods for AD vs. NC classification. The kappa measure evaluates the level of agreement between the outputs of two classifiers. In Fig. 11, we show a diversity-error diagram achieved by different methods. For each method, the corresponding ensemble contains 50 individual classifiers for 50 landmarks. The value on the x-axis of a diversity-error diagram denotes the kappa measure of a pair of classifiers in the ensemble, whereas the value on the y-axis is the averaged individual error of a pair of classifiers. As a small value of kappa measure indicates better diversity and a small value of averaged individual error indicates a better accuracy, the most desirable pairs of classifiers will be close to the bottom left corner of the graph. We further plot the centroids of clouds achieved by different methods in Fig. 11 for visual evaluation of relative positions of kappa-error points.

Fig. 11 suggests that the proposed FC11_ens method outperforms the competing methods regarding the averaged classification error, while our FC10_ens method achieves the best diversity regarding kappa measure. Also, the proposed LDFL method gives the overall best trade-off between the classification error and the diversity of multiple classifiers, compared with the other four methods. That is, LDFL builds a classifier ensemble based on the reasonably accurate but markedly diverse individual components.

C. Computational Cost

We now analyze the computational costs of the proposed LDFL method. It is worth noting that, for our LDFL and three competing methods (i.e., ROI, VBM, and LMF [28]), all training processes are performed off-line. Hence, we analyze the on-line computational cost for a new testing subject with an MR image. Specifically, in the ROI method, we first linearly align the testing image to the template, and then use the non-linear registration algorithm [31] to map segmentations of gray matter (GM) and the 90 ROIs from template image to the testing image, followed by a linear support vector machine (SVM) classifier. Similar to ROI, we use the same registration and segmentation strategies for voxel-based method (VBM). In our LDFL method, we first linearly align the new testing MR image to the template, and then predict the landmark positions for this image. We further extract image patches from each landmark location and feed them to the proposed CNN model for joint feature learning and disease prediction. For LMF [28] method, we extract morphological features [32] of the linearly-aligned testing image based on the same landmarks as LDFL, and then perform prediction using the linear SVM classifier.

Table II reports the computational costs of different methods. From Table II, we can see that the total computational costs of both ROI and VBM are more than half an hour, which is much slower than our method (15 s). Although LMF has similar computational cost (20 s), its learning performance is worse than our proposed LDFL method (see results in Fig. 7 and Figs. S4–S6 in the *Supplementary Materials*).

D. Technical Limitations

Although the proposed LDFL method achieved promising results in both tasks of brain disease diagnosis and MR image retrieval, several technical issues need to be considered.

First, the anatomical landmarks used in this study is pre-defined in our previous study [28]. That is, the process of landmark definition is independent of our proposed patch-level feature learning, which may lead to sub-optimal performance. A reasonable solution is to integrate landmark identification and landmark-based feature learning into a unified deep learning framework, which will be one of our future works. *Second and more generally*, the anatomical landmarks used in this study were discovered in a data-driven manner. However, it remains unknown which subset of landmarks is the most informative for subsequent feature learning. Therefore, it could be interesting to investigate an optimal subset of identified landmarks for patch-based feature learning. As one of our future works, we will let experts refine these landmarks to make the used landmarks more compact. *Besides*, only the baseline MRI data in three datasets (i.e., ADNI-1, ADNI-2, and MIRIAD) are used in this work. In these datasets, there exist longitudinal MRI data that may provide complementary information for the proposed feature learning method. *Furthermore*, we learn multiple CNN models (*w.r.t.* multiple landmarks) separately in the current study, without considering the context information (e.g., spatial locations) of those identified landmarks. Hence, future work will cover the development of a joint deep learning model by considering the landmarks jointly and globally.

V. CONCLUSION

In this paper, we propose a landmark based deep feature learning (LDFL) framework, to automatically extract patch-based representations from MR images for AD-related brain disease diagnosis. Experimental results on three cohorts (i.e., ADNI-1, ADNI-2, and MIRIAD) demonstrate the effectiveness of the proposed method in both tasks of disease classification and MR image retrieval. This approach paves the way to discriminative biomarkers for computer-aided diagnosis of AD and the morphological analysis of MR images.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by the National Institutes of Health under Grants EB006733, EB008374, EB009634, MH100217, AG041721, AG042599, AG010129, and AG030514.

REFERENCES

[1]. Hardy J and Selkoe DJ, “The amyloid hypothesis of Alzheimer’s disease: Progress and problems on the road to therapeutics,” Science, vol. 297, no. 5580, pp. 353–356, 2002. [PubMed: 12130773]

[2]. Lian C, Ruan S, Denoeux T, Li H, and Vera P, “Spatial evidential clustering with adaptive distance metric for tumor segmentation in FDG-PET images,” IEEE Trans. Biomed. Eng, vol. 65, no. 1, pp. 21–30, Jan. 2018. [PubMed: 28371772]

[3]. Lian C, Ruan S, Denœux T, Jardin F, and Vera P, “Selecting radiomic features from FDG-PET images for cancer treatment outcome prediction,” Med. Image Anal, vol. 32, pp. 257–268, 2016. [PubMed: 27236221]

[4]. Frisoni GB, Fox NC, Jack CR, Scheltens P, and Thompson PM, “The clinical use of structural MRI in Alzheimer disease,” Nature Rev. Neurol, vol. 6, no. 2, pp. 67–77, 2010. [PubMed: 20139996]

[5]. Reiman EM, Langbaum JB, and Tariot PN, “Alzheimer’s prevention initiative: A proposal to evaluate presymptomatic treatments as quickly as possible,” Biomarkers Med, vol. 4, no. 1, pp. 3–14, 2010.

[6]. Fischl B and Dale AM, “Measuring the thickness of the human cerebral cortex from magnetic resonance images,” Proc. Nat. Acad. Sci. USA, vol. 97, no. 20, pp. 11050–11055, 2000. [PubMed: 10984517]

[7]. Cuingnet R et al., “Automatic classification of patients with Alzheimer’s disease from structural MRI: A comparison of ten methods using the ADNI database,” NeuroImage, vol. 56, no. 2, pp. 766–781, 2011. [PubMed: 20542124]

[8]. Lötjönen J et al., “Fast and robust extraction of hippocampus from MR images for diagnostics of Alzheimer’s disease,” NeuroImage, vol. 56, no. 1, pp. 185–196, 2011. [PubMed: 21281717]

[9]. Liu M, Zhang J, Yap P-T, and Shen D, “View-aligned hypergraph learning for Alzheimer’s disease diagnosis with incomplete multi-modality data,” Med. Image Anal, vol. 36, pp. 123–134, 2017. [PubMed: 27898305]

[10]. Liu M, Zhang D, and Shen D, “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.

[11]. Jack CR, Petersen RC, O’Brien PC, and Tangalos EG, “MR-based hippocampal volumetry in the diagnosis of Alzheimer’s disease,” Neurology, vol. 42, no. 1, pp. 183–183, 1992. [PubMed: 1734300]

[12]. Jack C et al., “Prediction of AD with MRI-based hippocampal volume in mild cognitive impairment,” Neurology, vol. 52, no. 7, pp. 1397–1397, 1999. [PubMed: 10227624]

[13]. Atiya M, Hyman BT, Albert MS, and Killiany R, “Structural magnetic resonance imaging in established and prodromal Alzheimer’s disease: A review,” Alzheimer Dis. Assoc. Disorders, vol. 17, no. 3, pp. 177–195, 2003.

[14]. Yamasue H et al., “Voxel-based analysis of MRI reveals anterior cingu-late gray-matter volume reduction in posttraumatic stress disorder due to terrorism,” Proc. Nat. Acad. Sci. USA, vol. 100, no. 15, pp. 9039–9043, 2003. [PubMed: 12853571]

[15]. Maguire EA et al., “Navigation-related structural change in the hippocampi of taxi drivers,” Proc. Nat. Acad. Sci. USA, vol. 97, no. 8, pp. 4398–4403, 2000. [PubMed: 10716738]

[16]. Small GW et al., “Cerebral metabolic and cognitive decline in persons at genetic risk for Alzheimer’s disease,” Proc. Nat. Acad. Sci. USA, vol. 97, no. 11, pp. 6037–6042, 2000. [PubMed: 10811879]

[17]. Liu M, Zhang D, Adeli E, and Shen D, “Inherent structure-based multiview learning with multitemplate feature representation for Alzheimer’s disease diagnosis,” IEEE Trans. Biomed. Eng, vol. 63, no. 7, pp. 1473–1482, Jul. 2016. [PubMed: 26540666]

[18]. Ashburner J and Friston KJ, “Voxel-based morphometry–The methods,” NeuroImage, vol. 11, no. 6, pp. 805–821, 2000. [PubMed: 10860804]

[19]. Baron J et al., “In vivo mapping of gray matter loss with voxel-based morphometry in mild Alzheimer’s disease,” NeuroImage, vol. 14, no. 2, pp. 298–309, 2001. [PubMed: 11467904]

[20]. Klöppel S et al., “Automatic classification of MR scans in Alzheimer’s disease,” Brain, vol. 131, no. 3, pp. 681–689, 2008. [PubMed: 18202106]

[21]. Gaser C, Nenadic I, Buchsbaum BR, Hazlett EA, and Buchsbaum MS, “Deformation-based morphometry and its relation to conventional volumetry of brain lateral ventricles in MRI,” NeuroImage, vol. 13, no. 6, pp. 1140–1145, 2001. [PubMed: 11352619]

[22]. Hua X et al., “Tensor-based morphometry as a neuroimaging biomarker for Alzheimer’s disease: An MRI study of 676 AD, MCI, and normal subjects,” NeuroImage, vol. 43, no. 3, pp. 458–469, 2008. [PubMed: 18691658]

[23]. Friedman J, Hastie T, and Tibshirani R, The Elements of Statistical Learning. Springer series in statistics, vol. 1. Berlin, Germany: Springer, 2001.

[24]. Zhang J, Liu M, and Shen D, “Detecting anatomical landmarks from limited medical imaging data using two-stage task-oriented deep neural networks,” IEEE Trans. Image Process, vol. 26, no. 10, pp. 4753–4764, Oct. 2017. [PubMed: 28678706]

[25]. Jack CR et al., “The Alzheimer’s disease neuroimaging initiative (ADNI): MRI methods,” J. Magn. Reson. Imag, vol. 27, no. 4, pp. 685–691, 2008.

[26]. Sled JG, Zijdenbos AP, and Evans AC, “A nonparametric method for automatic correction of intensity nonuniformity in MRI data,” IEEE Trans. Med. Imag, vol. 17, no. 1, pp. 87–97, Feb. 1998.

[27]. Wang Y, Nie J, Yap P-T, Shi F, Guo L, and Shen D, “Robust deformable-surface-based skull-stripping for large-scale studies,” in Proc. Med. Image Comput. Comput.-Assist. Intervention, 2011, pp. 635–642.

[28]. Zhang J, Gao Y, Gao Y, Munsell B, and Shen D, “Detecting anatomical landmarks for fast Alzheimer’s disease diagnosis,” IEEE Trans. Med. Imag, vol. 35, no. 12, pp. 2524–2533, Dec. 2016.

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

[30]. Holmes CJ, Hoge R, Collins L, Woods R, Toga AW, and Evans AC, “Enhancement of MR images using registration for signal averaging,” J. Comput. Assist. Tomography, vol. 22, no. 2, pp. 324–333, 1998.

[31]. Shen D and Davatzikos C, “HAMMER: Hierarchical attribute matching mechanism for elastic registration,” IEEE Trans. Med. Imag, vol. 21, no. 11, pp. 1421–1439, Nov. 2002.

[32]. Zhang J, Liang J, and Zhao H, “Local energy pattern for texture classification using self-adaptive quantization thresholds,” IEEE Trans. Image Process, vol. 22, no. 1, pp. 31–42, Jan. 2013. [PubMed: 22910113]

[33]. Mardia K, “Assessment of multinormality and the robustness of hotelling’s t2 test,” Appl. Statist, vol. 24, pp. 163–171, 1975.

[34]. Liu M, Zhang J, Adeli E, and Shen D, “Landmark-based deep multi-instance learning for brain disease diagnosis,” Med. Image Anal, vol. 43, pp. 157–168, 2018. [PubMed: 29107865]

[35]. Krizhevsky A, Sutskever I, and Hinton GE, “ImageNet classification with deep convolutional neural networks,” in Proc. 25th Int. Conf. Neural Inf. Process. Syst, 2012, pp. 1097–1105.

[36]. Jia Y et al., “Caffe: Convolutional architecture for fast feature embedding,” in Proc. ACM Int. Conf. Multimedia, 2014, pp. 675–678.

[37]. Boyd S and Vandenberghe L, Convex Optimization. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[38]. Dietterich TG, “Ensemble methods in machine learning,” in Proc. Int. Workshop Multiple Classifier Syst, 2000, pp. 1–15.

[39]. Zhang D, Wang Y, Zhou L, Yuan H, and Shen D, “Multimodal classification of Alzheimer’s disease and mild cognitive impairment,” NeuroImage, vol. 55, no. 3, pp. 856–867, 2011. [PubMed: 21236349]

[40]. Tzourio-Mazoyer N 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. [PubMed: 11771995]

[41]. Chang C-C and Lin C-J, “LIBSVM: A library for support vector machines,” ACM Trans. Intell. Syst. Technol, vol. 2, no. 3, 2011, Art. no. 27.

[42]. Jain A, Nandakumar K, and Ross A, “Score normalization in multimodal biometric systems,” Pattern Recog, vol. 38, no. 12, pp. 2270–2285, 2005.

[43]. Maaten L. v. d. and Hinton G, “Visualizing data using t-SNE,” J. Mach. Learn. Res, vol. 9, pp. 2579–2605, 2008.

[44]. Chinchor N and Sundheim B, “Muc-5 evaluation metrics,” in Proc. 5th Conf. Message Understand, 1993, pp. 69–78.

[45]. Matthews BW, “Comparison of the predicted and observed secondary structure of t4 phage lysozyme,” Biochim. Biophys. Acta, Protein Struct, vol. 405, no. 2, pp. 442–451, 1975.

[46]. Rodriguez JJ, Kuncheva LI, and Alonso CJ, “Rotation forest: A new classifier ensemble method,” IEEE Trans. Pattern Anal. Mach. Intell, vol. 28, no. 10, pp. 1619–1630, Oct. 2006. [PubMed: 16986543]

[47]. Hyman BT, Van Hoesen GW, Damasio AR, and Barnes CL, “Alzheimer’s disease: Cell-specific pathology isolates the hippocampal formation,” Science, vol. 225, no. 4667, pp. 1168–1170, 1984. [PubMed: 6474172]

[48]. De Jong L et al., “Strongly reduced volumes of putamen and thalamus in Alzheimer’s disease: An MRI study,” Brain, vol. 131, no. 12, pp. 3277–3285, 2008. [PubMed: 19022861]

[49]. Chan D et al., “Patterns of temporal lobe atrophy in semantic dementia and Alzheimer’s disease,” Ann. Neurol, vol. 49, no. 4, pp. 433–442, 2001. [PubMed: 11310620]

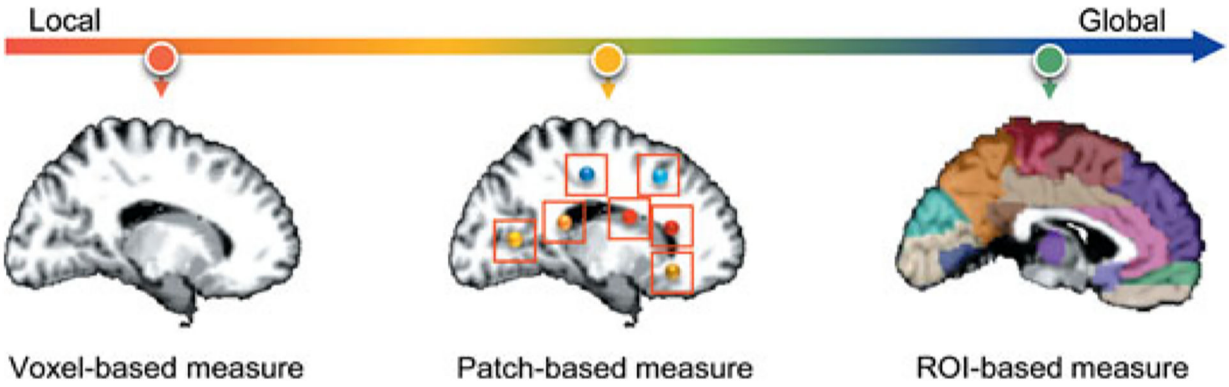

Fig. 1.

Illustration of MRI biomarkers for brain disease diagnosis shown in a local-to-global manner, including high-dimensional-morphological measure (e.g., voxel-based representation), region-of-interest (ROI)-based measure, and the proposed patch-based measure.

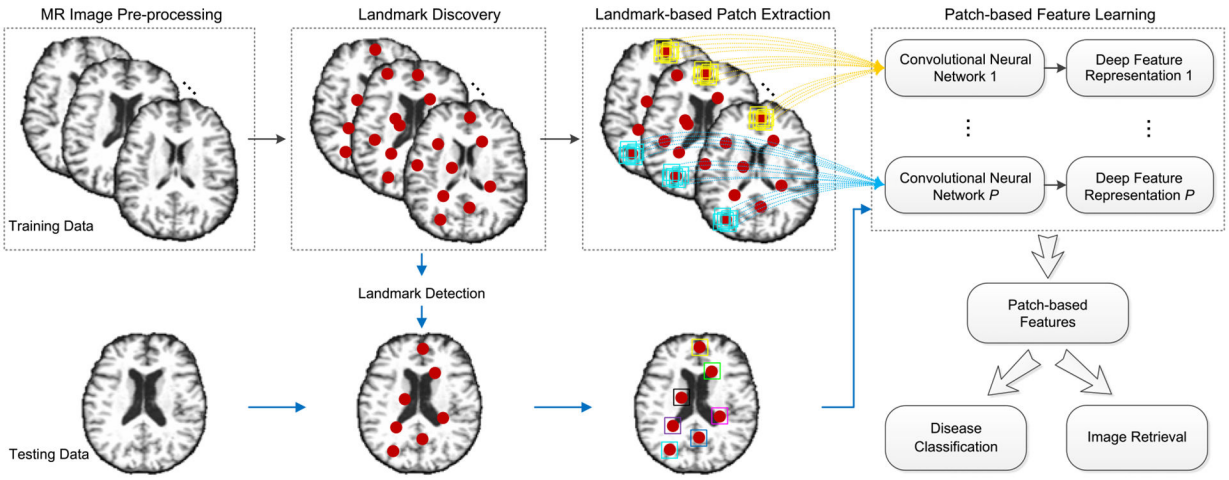

Fig. 2.

Illustration of the proposed anatomical Landmark based Deep Feature Learning (LDFL) framework. There are four main components: 1) landmark discovery, 2) landmark-based patch extraction, 3) patch-based feature learning, and 4) applications of disease classification and image retrieval.

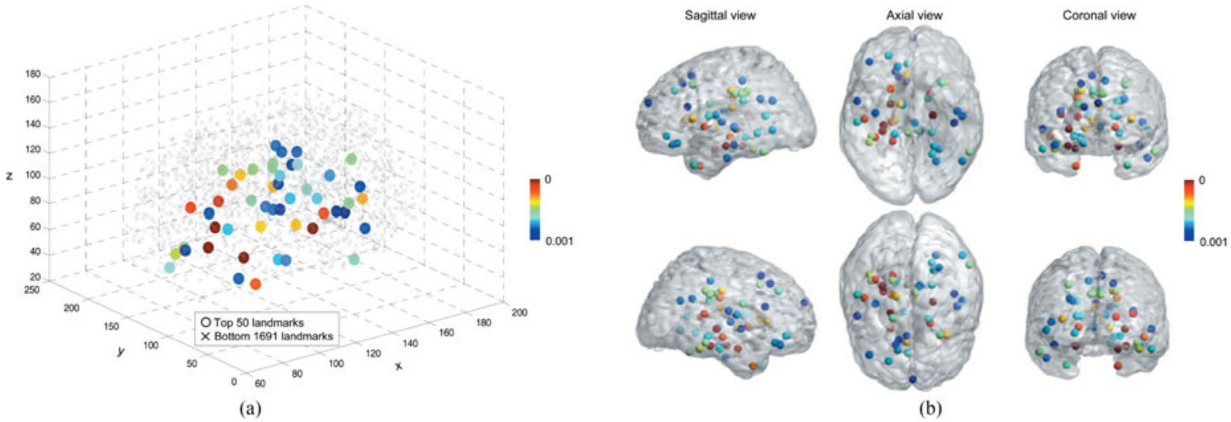

Fig. 3.

Illustration of (a) all identified AD-related anatomical landmarks from AD and NC subjects in ADNI-1, and (b) top 50 selected landmarks for patch-level feature learning in AD diagnosis. Here, different colors denote *p*-values in group comparison between AD and NC groups in ADNI-1.

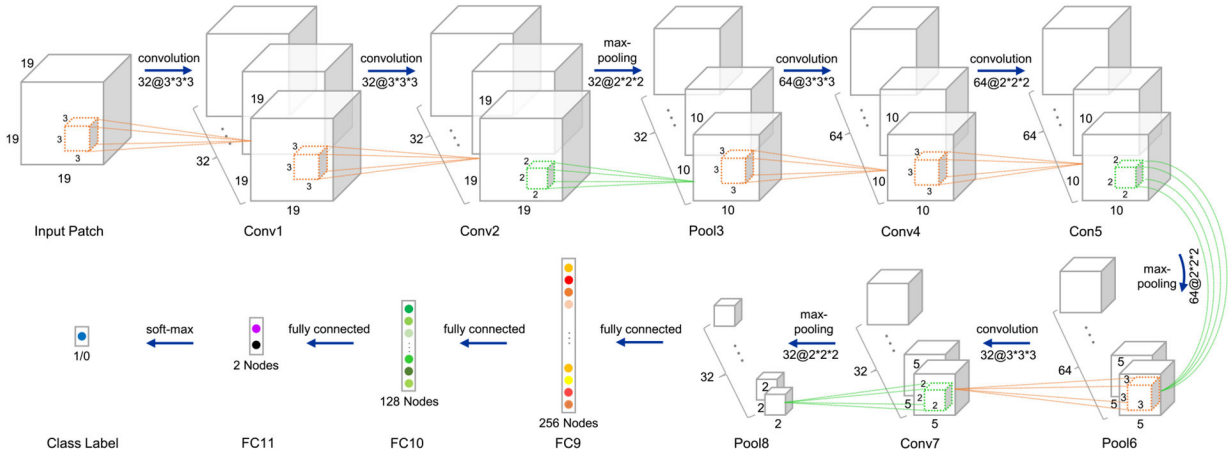

Fig. 4.

Schematic diagram of CNN for patch-based feature learning. Conv: Convolutional layer; Pool: Pooling layer; FC: Fully connected layer.

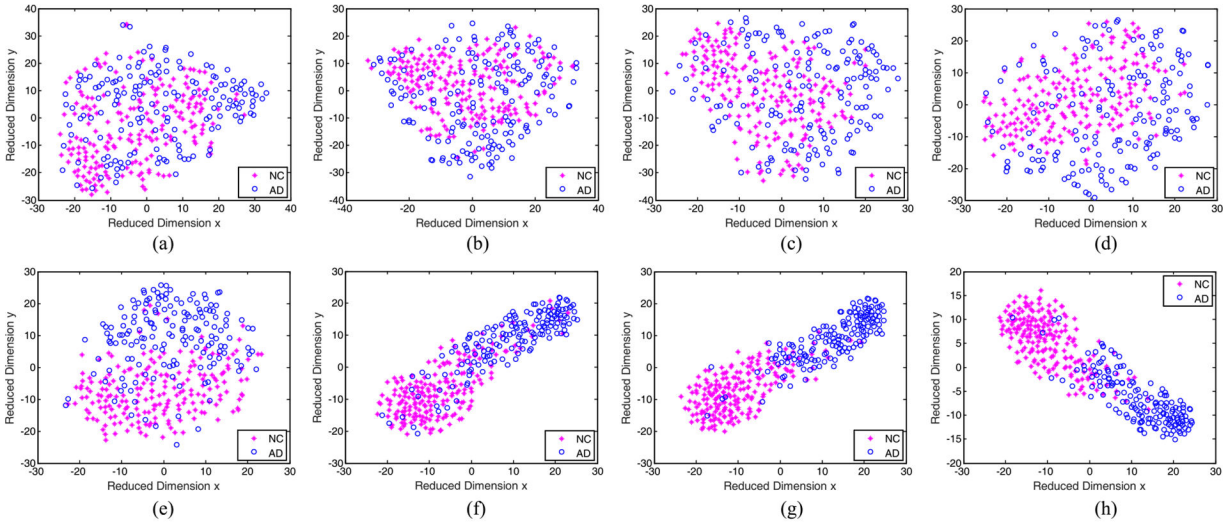

Fig. 5.

Manifold visualization of AD and NC subjects in the ADNI-2 dataset, by t-SNE projection [43] in learned 3D-CNN layers including (a) Conv1, (b) Conv2, (c) Conv4, (d) Conv5, (e) Conv7, (f) FC9, (g) FC10, and (h) FC11.

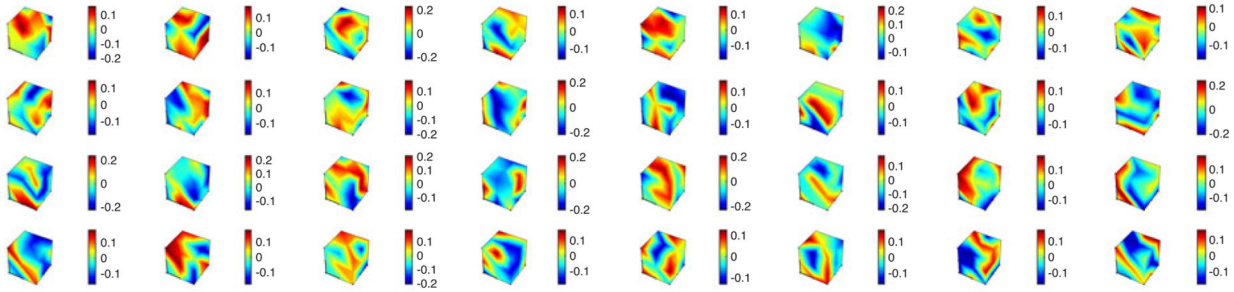

Fig. 6.

Illustration of the learned 32 convolutional kernels (with the size of 3 × 3 × 3) at the Conv1 layer in the proposed CNN architecture for AD vs. NC classification.

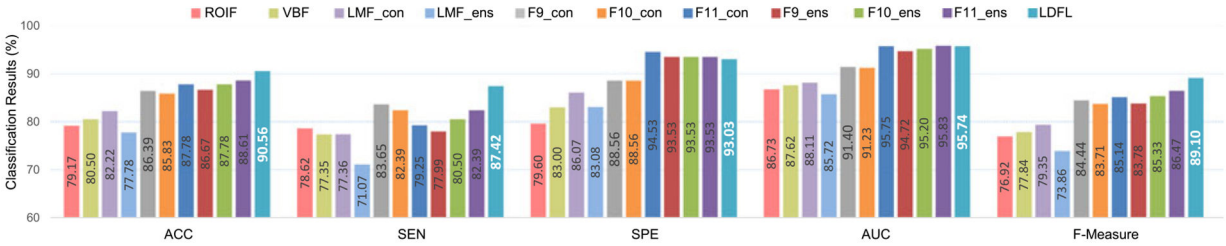

Fig. 7.

Results in AD vs. NC classification achieved by different methods, where subjects with MR imaging in ADNI-1 are used as *training data* and those in ADNI-2 are used as *independent testing data*. ACC: accuracy, SEN: sensitivity, SPE: specificity, AUC: area under the receiver operating characteristic curve.

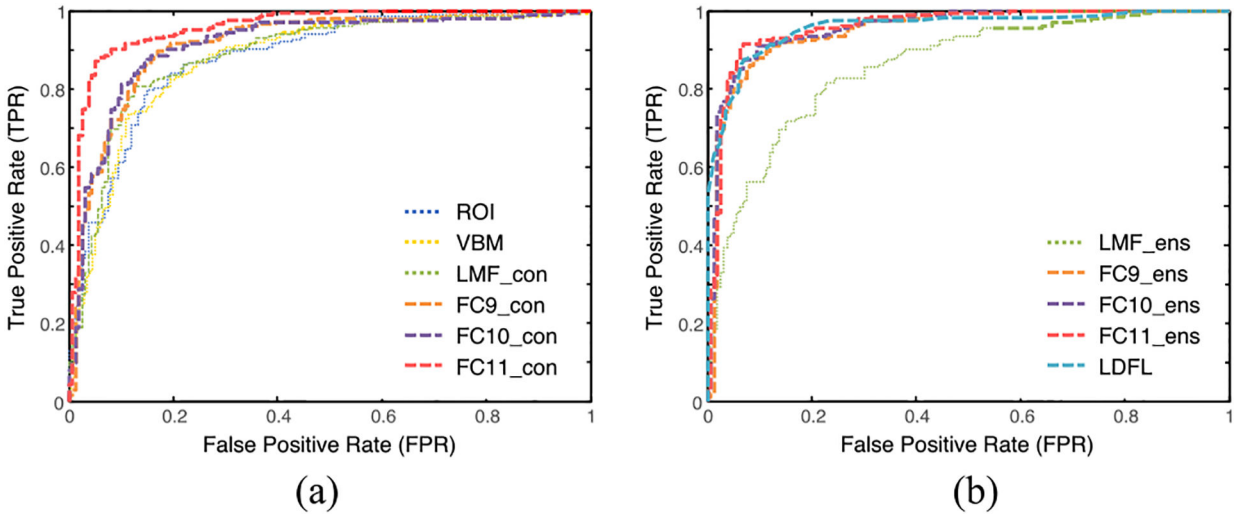

Fig. 8.

ROC curves achieved by (a) feature concatenation based methods and (b) ensemble based methods in AD vs. NC classification. The classifiers are trained on ADNI-1 and tested on ADNI-2.

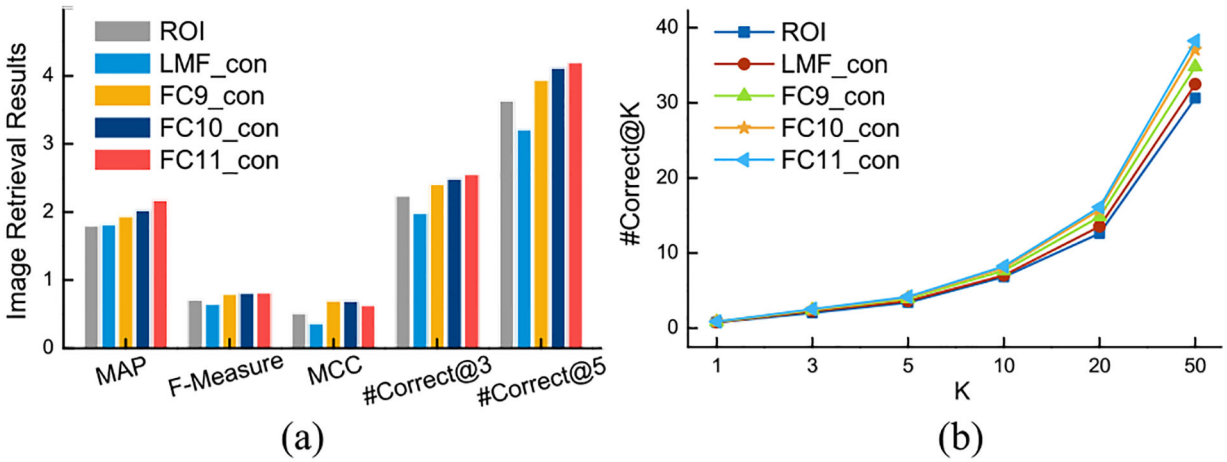

Fig. 9.

Image retrieval results achieved by different methods. All the MR images in the ADNI-2 dataset are alternated used as the query image. MAP: mean average precision; MCC: Matthews correlation coefficient; #Correct@K: number of correct results in top K returned results.

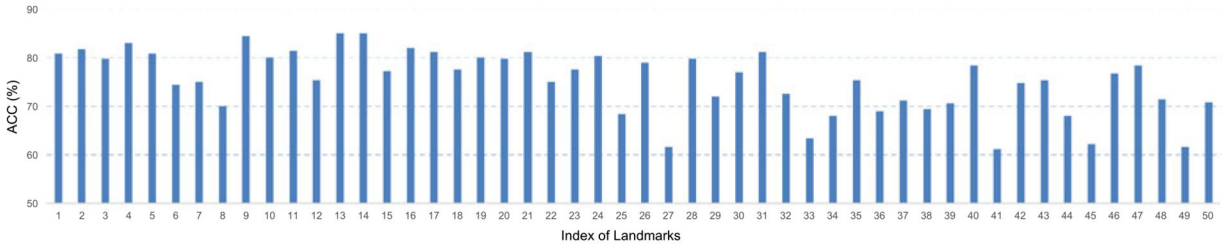

Fig. 10.

Classification accuracies of patches in each landmark position for the testing data in the ADNI-2 dataset.

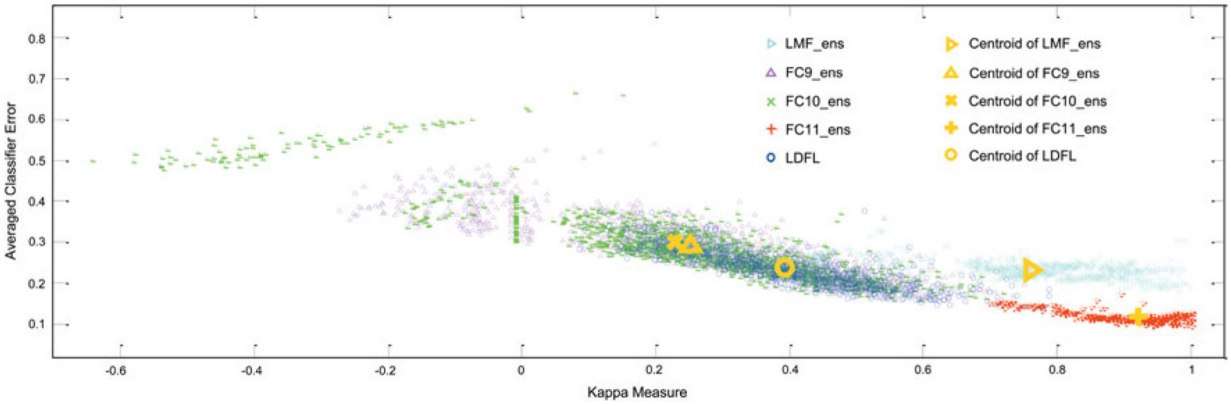

Fig. 11.

Kappa-error diagram achieved by five ensemble based methods in AD vs. NC classification. The value on the *x*-axis denotes the kappa measure [46] of a pair of classifiers in the ensemble, whereas the value on the *y*-axis is the averaged individual error of a pair of classifiers.

Table I

Demographic and Clinical Information of Subjects in the Baseline ADNI-1, ADNI2, and Miriad Datasets

Dataset	Category	Male/Female	Edu (Mean ± Std)	Age (Mean ± Std)	MMSE (Mean ± Std)	CDR-SB (Mean ± Std)
ADNI-1	AD	106/93	13.09 ± 6.83	69.98 ± 22.35	23.27 ± 2.02	0.74 ±0.25
	NC	127/102	15.71 ± 4.12	74.72 ± 10.98	29.11 ±1.01	0.00 ±0.00
ADNI-2	AD	91/68	14.19 ± 6.79	69.06 ± 22.04	21.66 ±6.07	4.16±2.01
	NC	113/87	15.66 ± 3.46	73.82 ± 8.41	27.12 ±7.31	0.05 ±0.22
MIRIAD	AD	19/27	-	69.95 ± 7.07	19.19 ±4.01	-
	NC	12/11	-	70.36 ± 7.28	29.39 ± 0.84	-

Values are reported as Mean± Standard Deviation (Std); Edu: Education years; MMSE: Mini-mental state examination; CDR-SB: Clinical Dementia Rating-Sum of Boxes.

Table II

Computational Costs of Different Methods in AD VS. NC Classification for a New Testing MR Image

Method	Tims of Each Process (Platform;		Total Time
ROI	1) Linear alignment	5.00 s (C++)	~32.00 min
	2) Non-linear registration	32.00 min (HAMMER [31])	
	3) Feature extraction	3.00 s (Matlab)	
	4) Classification	0.02 s (Matlab)	
VBM	1) Linear alignment	5.00 s (C++)	~32.00 min
	2) Non-linear registration	32.00 mm (HAMMER [31])	
	3) Feature extraction	4.00 s (Matlab)	
	4) Classification	0.05 s (Matlab)	
LMF	1) Linear alignment	5.00 s (C++)	~ 20.00 s
	2) Landmark prediction	10.00 s (Matlab)	
	3) Feature extraction	5.00s (Matlab)	
	4) Classification	0.03 s (Matlab)	
LDFL (Ours)	1) Linear alignment	5.00 s (C++)	~ 15,00 s
	2) Landmark prediction	10.00 s (Matlab)	
	3) Joint feature extraction and classification	0.31 s (Caffe[36])	