NeuroImage 202 (2019) 116113

Contents lists available at ScienceDirect

NeuroImage

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

# Toward an interpretable Alzheimer's disease diagnostic model with regional abnormality representation via deep learning

Eunho Lee [a,1], Jun-Sik Choi [a,1], Minjeong Kim [c], Heung-Il Suk [a,b,*], the Alzheimer's Disease Neuroimaging Initiative[2]

[a] Department of Brain and Cognitive Engineering, Korea University, Republic of Korea
[b] Department of Artificial Intelligence, Korea University, Republic of Korea
[c] Department of Computer Science, University of North Carolina at Greensboro, USA

ARTICLE INFO

*Keywords:*
Abnormality representation
Interpretable diagnostic model
Magnetic resonance imaging
Deep neural network
Alzheimer's disease
Mild cognitive impairment

ABSTRACT

In this paper, we propose a novel method for magnetic resonance imaging based Alzheimer's disease (AD) or mild cognitive impairment (MCI) diagnosis that systematically integrates voxel-based, region-based, and patch-based approaches into a unified framework. Specifically, we parcellate the brain into predefined regions based on anatomical knowledge (*i.e.*, templates) and derive complex nonlinear relationships among voxels, whose intensities denote volumetric measurements, within each region. Unlike existing methods that use cubical or rectangular shapes, we consider the anatomical shapes of regions as atypical patches. Using complex nonlinear relationships among voxels in each region learned by deep neural networks, we extract a "regional abnormality representation." We then make a final clinical decision by integrating the regional abnormality representations over the entire brain. It is noteworthy that the regional abnormality representations allow us to interpret and understand the symptomatic observations of a subject with AD or MCI by mapping and visualizing these observations in the brain space. On the baseline MRI dataset from the Alzheimer's Disease Neuroimaging Initiative (ADNI) cohort, our method achieves state-of-the-art performance for four binary classification tasks and one three-class classification task. Additionally, we conducted exhaustive experiments and analysis to validate the efficacy and potential of our method.

## 1. Introduction

Alzheimer's disease (AD) is a chronic neurodegenerative disease and the most common cause of dementia (Barker et al., 2002). In the gradual progression of AD, neurons in the broad regions of the brain are irreversibly damaged or destroyed and patients suffer from evolving symptoms accordingly. Apathy and depression often appear in the early stages and major cognitive function issues, including impaired communication, disorientation, confusion, poor judgment, and behavior changes, follow as later symptoms. Because neuronal destruction is not limited to the brain area dedicated to cognitive function, the disease ultimately affects life-critical functions, such as swallowing. This makes AD a fatal disease. The main pathogenesis of the disease is the progressive accumulation of plaques composed of amyloid $\beta$ and tangles composed of hyper-phosphorylated $\tau$ protein. A familial form of AD is known to be caused by mutations in the genes linked to amyloid $\beta$ metabolism. However, the cause of the sporadic form of AD is still unknown (Blennow et al., 2006).

Among the various causes of dementia, AD accounts for an estimated 60–80% percent of all cases. There are approximately 5.5 million individuals suffering from dementia caused by AD (Association et al., 2017). In addition to the prevalence of AD, the fact that most of patients are eventually bedridden and require 24-h care makes the disease an enormous burden on society. Considering that aging is a major risk factor for AD and that the elderly population is projected to double in size to 1.6 billion people globally, the societal cost of AD will increase rapidly (He

[*] Corresponding author. Department of Brain and Cognitive Engineering, Korea University, Republic of Korea.
*E-mail address:* hisuk@korea.ac.kr (H.-I. Suk).

[1] Equally contributed.

[2] Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (http://www.loni.ucla.edu/ADNI). 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. ADNI investigators include (complete listing available at http://adni.loni.ucla.edu/wpcontent/uploads/how_to_apply/ADNI_Authorship_List.pdf).

https://doi.org/10.1016/j.neuroimage.2019.116113
Received 11 December 2018; Received in revised form 26 July 2019; Accepted 19 August 2019
Available online 22 August 2019

et al., 2016). For the treatment of AD, although many candidates have been and are being tested in clinical trials, there have been no successful treatments for preventing, slowing, or halting the progression of AD. It is known that changes in the brain caused by AD begin approximately 20 or more years before the onset of dementia (Villemagne et al., 2013; Reiman et al., 2012). In the preclinical stage of AD, people often experience mild, but measurable changes in cognitive abilities, which do not interfere with their daily life. This condition, called mild cognitive impairment (MCI), does not always lead to dementia, but an average of 32% percent of individuals with MCI develop AD-related dementia within five years (Ward et al., 2013). Additionally, the amnestic subtype of MCI has a high risk of progression to AD, meaning it could represent a prodromal stage of the disease (Gauthier et al., 2006; Kantarci et al., 2009; Mitchell and Shiri-Feshki, 2009). Therefore, the development of methods for the early diagnosis and prediction of AD is of paramount clinical importance. The early detection of AD can provide patients with better a chance to fight the disease because potential treatments are likely to be most efficacious when applied in early stages.

Biomarkers can play an important role in the diagnosis of AD in its preclinical and MCI stages. With the proper use of biomarkers in diagnosis, it is expected that AD can be detected before any cognitive symptoms occur. The diagnostic guidelines provided by the National Institute of Aging and the Alzheimer's Association also reflect the importance of biomarkers in AD diagnosis. The identification of such biomarkers is highly dependent on cognitive tests (Sperling et al., 2011; Albert et al., 2011; McKhann et al., 2011). Among the various types of biomarkers, imaging-based biomarkers from brain imaging (*e.g.*, magnetic resonance imaging (MRI) and positron emission tomography) are of significant interest based on their widespread availability. Brain imaging has served as a key auxiliary tool for clinical diagnosis and brain studies. In AD, cortical neuronal loss leaves disease-related pathological patterns that can be captured by structural MRI (sMRI) (Zarow et al., 2005). Various types of structural alteration patterns, such as medial temporal lobe atrophy and hippocampal volume reduction, can be observed when investigating brain images. As imaging-biomarkers, these patterns can be used to measure the progression of the disease or differentiate causes of cognitive decline. However, because the human brain is highly complex and different regions are functionally or structurally related to each other, changes in multiple regions or all regions in a brain must be considered jointly.

To investigate the complex relationships between neuropathological patterns across brain regions, many studies have developed machine-learning- or deep-learning-based methods (Klöppel et al., 2008; Moradi et al., 2015; Shen et al., 2014, 2017; Liu et al., 2014b; Suk et al., 2015).

One of the many challenges in building an imaging-based diagnostic model as a clinical decision support system is the ultra-high dimensionality of input images and extremely limited number of samples for tuning the learnable parameters of a model. For example, there are more than five million voxels in a 256 × 256 × 256-voxel sMRI image, but only hundreds of sMRI samples in a dataset. Under such conditions, it is very difficult to train machine-learning or deep-learning models without overfitting or a lack of generalization. One straight-forward solution can be to construct a large-scale image dataset like the ImageNet in the field of computer vision (Deng et al., 2009). However, it is generally very difficult to collect a sufficient number of images based on the high cost of scanning, diverse imaging modality and inter-scanner, and inter-site variance, although there are a number of national-scale projects to build huge neuroimaging datasets as the Human Connectome Project (Van Essen et al., 2013).

Because it is paramount to extract well-designed or target-task-relevant features to enhance a clinical accuracy, there have been many efforts to overcome the limitation of composing a large dataset computationally and algorithmically. In (Suk et al., 2014), existing methods of imaging-based feature extraction were divided into three categories: voxel-based approaches, region-based approaches, and patch-based approaches. There are tradeoffs in terms of learning complexity and information gain between those approaches. A voxel-based approach uses voxel intensities, the most fine-grained information, but suffers from high dimensionality issues (Rathore et al., 2017). A region-based approach considers structurally or functionally predefined brain regions and extracts representative handcrafted features (*e.g.*, mean volume) from each region (Liu et al., 2015). Although this approach significantly reduces the dimensionality of a feature space, based on the limitations of coarse-grained information, it may miss small or subtle changes in the early stages of the AD progression spectrum. A patch-based approach attempts to find relationships between voxels within the predefined-form of a patch, thereby combining the merits of voxel-based and region-based approaches. However, a patch is generally a manually defined cubical or rectangular form. (Tong et al., 2014). There have also been efforts to construct a set of classifiers, rather than a single classifier, such that each classifier finds different patterns, which are eventually combined through simple averaging (Liu et al., 2012b) or a more systematic method to reach an ensemble decision (Ithapu et al., 2015; Suk et al., 2017).

Previous works have contributed to improving diagnostic accuracy using public datasets (*e.g.*, Alzheimer's Disease Neuroimaging Initiative (ADNI)) and identifying potential imaging biomarkers. However, to the best of our knowledge, no studies have focused on providing a quantitative and investigative method for measuring brain abnormality in the spectrum of AD progression. We believe that the quantitative abnormality measurement or indexing of brain regions can provide clinically useful information regarding AD progression by allowing clinicians to link changes in the brain to symptomatic observations.

In this paper, we propose a novel framework that systemically combines the three aforementioned approaches to feature representation or extraction. Specifically, we parcellate the brain into predefined regions (region-based approach) and identify complex nonlinear relationships between voxels (voxel-based approach) within a patch whose form is determined by the anatomical shape of each region (patch-based approach). Note that in our method, a patch is atypically shaped based on the forms of regions, rather than cubical or rectangular (Fan et al., 2007). Using the complex voxel relationships in each region learned by deep neural networks (DNNs), we measure abnormalities in the spectrum of AD progression based on regional volume states. We then make a final clinical decision by integrating the region-based abnormality measurements over the entire brain. The morphological changes in a pathological brain (*i.e.*, regional atrophy, possibly caused by AD) can be small and subtle in multiple sub-regions within a region or span over multiple regions. In this regard, it should be possible to detect small and subtle changes, which generally can not be detected based on the mean volumes of regions. Furthermore, by considering region-based abnormality measurements as high-level information extracted from each region, we construct a robust and generalized classifier. Finally, our region-based abnormality measurement provides intuitive interpretation and understanding regarding the pathological statuses of various regions, thereby allowing us to make connections to symptomatic behaviors observed in a subject.

Compared to recent machine-/deep-learning based work for AD/MCI diagnosis, one of the major differences in our method is the interpretability of the prediction. In particular, while our method allows clinicians to interpret the output prediction via regional abnormality representation, thus to link any symptomatic behavioral observations in an individual level, the previous work mostly focused on either discriminative feature extraction or feature selection at the group level with no consideration of interpreting the model outputs. In terms of model learning, the independent work of (Adeli et al., 2019) and (An et al., 2017) focused on selecting class-discriminative features and/or samples, instead of feature representation learning. In the application of deep learning, mostly convolutional neural networks (CNNs), for AD/MCI diagnosis (Esmaeilzadeh et al., 2018), and (Lin et al., 2018) validated the use of CNNs for feature learning with promising results in their own experiments. Compared to those, our method exploits the anatomical

**Table 1**
Demographics and clinical information of subjects (pMCI: progressive MCI, sMCI: stable MCI, SD: standard deviation).

| | AD | pMCI | sMCI | CN |
|---|---|---|---|---|
| Number of Subjects | 198 | 160 | 214 | 229 |
| Sex (Female/Male) | 94/104 | 68/92 | 65/149 | 108/121 |
| Age (Mean ± SD) | 75.37 ± 7.55 | 74.89 ± 6.83 | 75.00 ± 7.63 | 75.96 ± 5.04 |
| Education (Mean ± SD) | 14.70 ± 3.13 | 15.69 ± 2.87 | 15.62 ± 3.18 | 16.03 ± 2.88 |
| Race | | | | |
| Asian | 2 | 3 | 5 | 3 |
| Black | 8 | 4 | 10 | 16 |
| White | 174 | 143 | 197 | 207 |
| More than One | 2 | 0 | 0 | 0 |
| Am Indian/Alaskan | 0 | 0 | 1 | 0 |
| Unknown | 12 | 10 | 0 | 3 |
| MMSE (Mean ± SD) | 23.28 ± 2.02 | 26.59 ± 1.71 | 27.28 ± 1.77 | 29.11 ± 1.00 |
| ADAS-Cog (Mean ± SD) | 18.44 ± 6.71 | 13.30 ± 4.05 | 10.33 ± 4.31 | 6.21 ± 2.93 |
| CDR (Mean ± SD) | 0.75 ± 0.25 | 0.50 ± 0.00 | 0.50 ± 0.03 | 0.00 ± 0.00 |

knowledge about a human brain to build a much simple model, while still achieving the state-of-the-art performance over the ADNI dataset.

## 2. Dataset and preprocessing

### 2.1. Dataset

We used a 1.5-T T1-weighted MRI dataset with images from 801 subjects from the ADNI-1[3] (Jack et al., 2008). Specifically, we considered a baseline dataset consisting of 229 cognitively normal (CN), 374 MCI, and 198 AD subjects. The MCI subjects were further categorized into 214 stable MCI (sMCI) and 160 progressive MCI (pMCI) subjects based on their AD progression over 18 months. Clinical information regarding each subject, such as their mini-mental state examination score (MMSE), Alzheimer's disease assessment scale-cognition (ADAS-Cog), and clinical dementia rating (CDR), is presented in Table 1 with other demographic information. Table 2 presents the statistical significance for each pair of the four groups in terms of age, gender, education, and race.

### 2.2. Preprocessing

The MRI images were preprocessed by applying the common procedures of anterior commissure (AC)-posterior commissure (PC) correction, skull stripping (Wang et al., 2011), and cerebellum removal. Specifically, we used the MIPAV software[4] for AC-PC correction, resampled images to 256 × 256 × 256 voxels, and applied the N3 algorithm (Sled et al., 1998) for intensity inhomogeneity correction. Following skull stripping and cerebellum removal, we checked the quality of the preprocessed images manually. The MRI images were then segmented into three tissue types, namely gray matter (GM), white matter, and cerebrospinal fluid by using the FAST (FMRIB's automated segmentation tool) (Zhang et al., 2001) from the FSL package.[5] Next, the segmented images were parcellated into 93 regions by warping each subject's brain into each subject's space using HAMMER (Shen and Davatzikos, 2002) based on Kabani et al.'s atlas (Kabani, 1998), which has been broadly applied in AD/MCI diagnosis. The full list of the ROIs in this template are provided in Appendix D. Finally, we acquired regional volumetric maps, referred to as RAVENS maps, by using a tissue-preserving image warping method (Davatzikos et al., 2001). To focus on effects of brain atrophy caused by neural destruction during the progression of AD, we used GM images for diagnosis. Note that each voxel intensity in a RAVENS map denotes a quantitative volumetric measurement, representing the volume in the image prior to warping based on the atlas. We consider this volume information as very fine-grained information that can be extracted from an input sMRI image.

**Table 2**
Summary of the statistical significance ($p$-value) for each pair of the four groups.

| Group | Age | Gender | Education | Race |
|---|---|---|---|---|
| AD vs. CN | 0.2938 | 0.4981 | $0.6773 \times 10^{-7}$ | 0.9064 |
| MCI vs. CN | 0.2943 | 0.4854 | 0.0752 | 0.6481 |
| AD vs. MCI | 0.5546 | 0.5267 | 0.001 | 0.7203 |
| pMCI vs. sMCI | 0.9324 | 0.4599 | 0.5663 | 0.8258 |

## 3. Methods

In this section, we describe our novel framework for assessing and representing regional abnormalities to identify the clinical status of a brain within the spectrum of AD progression. From a clinical perspective, it is of paramount importance to understand which parts of the brain are pathologic and how different brain regions are related to symptomatic observations. In this regard, it is desirable to construct a diagnostic model that can identify those parts of a brain that are affected by a pathologic disease and then quantify their abnormalities. Motivated by such interpretability issues, we propose a method for regional abnormality assessment based on DNNs.

Our method was inspired by two main factors. First, to enhance clinical accuracy, particularly in the early stages of AD progression, a diagnostic model must be sensitive to subtle changes in brain volume. Second, to the best of our knowledge, there are high variations in regional atrophies among individuals with no known general patterns. Therefore, we believe that the voxel-level analysis is extremely useful and that the relationships between voxels in different locations can provide helpful discriminative information for clinical diagnosis. However, based on the unfavorable high-dimensional nature of an MRI image in a voxel-level space and the limited number of samples available, it is challenging to find representations and train a diagnostic model for brain disorder identification using such fine-grained level relationships directly.

To circumvent the ultra-high dimensionality problem, we adopt a '*divide-and-conquer*' strategy (Cormen et al., 2009) based on cascaded two-level computations. Specifically, in the first level, we partition the brain (*i.e.*, the original high-dimensional input space) into a set of regions (*i.e.*, relatively small subspaces) based on a predefined anatomical atlas. For each brain region, we estimate a proxy representation of the voxels in the respective subspace. The proxy representations from all regions (*i.e.*, the entire brain) are then fed into a classifier to derive a final decision.

Fig. 1 illustrates the overall framework of the proposed method. Given an MRI input, we first preprocess it as described in Section 2 to obtain voxel-level GM volume density maps, which have been effectively used for analyzing morphological changes in the brain (Good et al., 2001; Hirata et al., 2005), then parcellate these maps into a set of regions based on an anatomical template. For each of these regions, we then construct a chunk of randomized DNNs (rDNNs), each of which is constructed based on a number of randomly selected voxels within the respective region. After training, the rDNNs output the regional probabilities of belonging to the clinical labels for a target classification task (*i.e.*, AD vs. CN, MCI vs. CN, pMCI vs. sMCI, or AD vs. MCI vs. CN). By considering the probability of a clinical status that appears in the late stages of the AD progression spectrum for the target task as a regional abnormality score, we define an abnormality representation for the entire brain by concatenating the regional abnormalities with one abnormality score for each region. By using this abnormality representation, we construct a classifier and make clinical decisions based on the target task.

[3] Available at 'http://www.loni.ucla.edu/ADNI.

[4] Available at 'http://mipav.cit.nih.gov/clickwrap.php.

[5] Available at 'http://fsl.fmrib.ox.ac.uk/fsl/fslwiki/

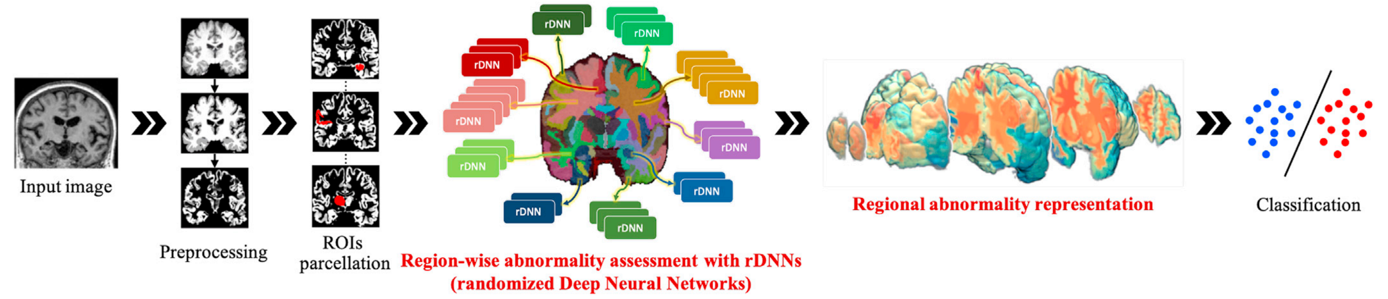

Input image
Preprocessing
ROIs parcellation
rDNN
Region-wise abnormality assessment with rDNNs (randomized Deep Neural Networks)
Regional abnormality representation
Classification

**Fig. 1.** A conceptual description of the proposed framework for assessing regional abnormalities and predicting a clinical statuses in the AD progression spectrum.

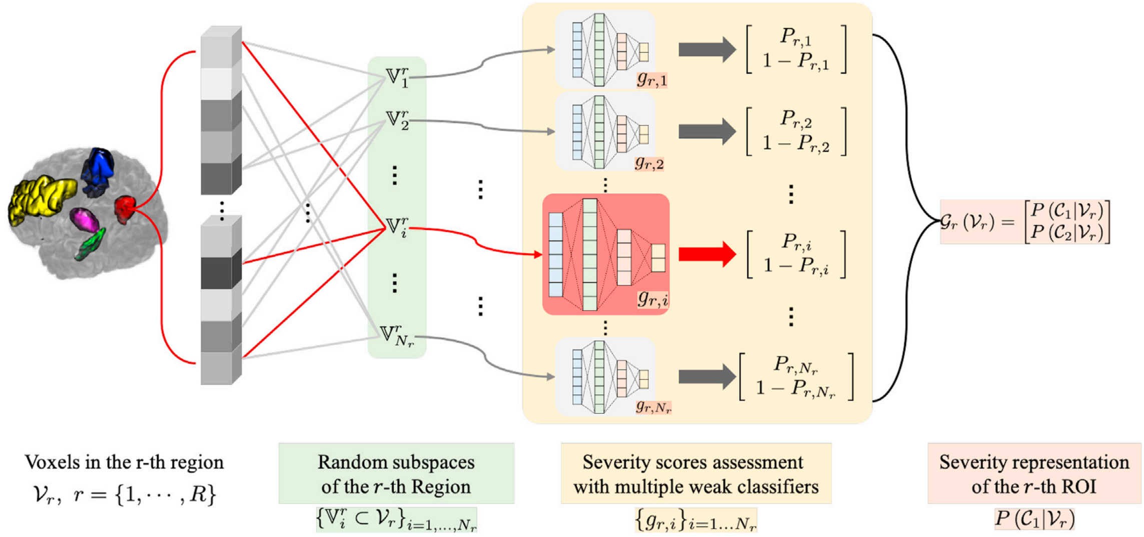

$\mathbb{V}_1^r$
$\mathbb{V}_2^r$
$\mathbb{V}_i^r$
$\mathbb{V}_{N_r}^r$
$g_{r,1}$
$g_{r,2}$
$g_{r,i}$
$g_{r,N_r}$
$\begin{bmatrix} P_{r,1} \\ 1-P_{r,1} \end{bmatrix}$
$\begin{bmatrix} P_{r,2} \\ 1-P_{r,2} \end{bmatrix}$
$\begin{bmatrix} P_{r,i} \\ 1-P_{r,i} \end{bmatrix}$
$\begin{bmatrix} P_{r,N_r} \\ 1-P_{r,N_r} \end{bmatrix}$
$\mathcal{G}_r(\mathcal{V}_r) = \begin{bmatrix} P(\mathcal{C}_1|\mathcal{V}_r) \\ P(\mathcal{C}_2|\mathcal{V}_r) \end{bmatrix}$
Voxels in the r-th region
$\mathcal{V}_r,\ r = \{1, \cdots, R\}$
Random subspaces of the $r$-th Region
$\{\mathbb{V}_i^r \subset \mathcal{V}_r\}_{i=1,\ldots,N_r}$
Severity scores assessment with multiple weak classifiers
$\{g_{r,i}\}_{i=1\ldots N_r}$
Severity representation of the $r$-th ROI
$P(\mathcal{C}_1|\mathcal{V}_r)$

**Fig. 2.** A schematic description of the region-wise abnormality assessment using rDNNs from Fig. 1. "$\mathscr{C}_1$" and "$\mathscr{C}_2$" denote the clinical labels of the later and earlier stages in the AD progression spectrum for a target task, respectively. For other notations, refer to the main text.

### 3.1. Regional abnormality assessment

Here, we describe the method for assessing regional abnormality that is illustrated in Fig. 2. Without loss of generality, we assume that a brain can be parcellated into either disjoint or overlapping $R$ regions, denoted $\{\mathscr{V}^r\}_{r=\{1,\ldots,R\}}$, where $\mathscr{V}^r$ denotes the voxel values (*e.g.*, GM densities) within the $r$-th region. Although we parcellate the brain into a set of regions for dimension reduction purposes, the spaces for some large regions are still too large for training with a limited number of samples. Therefore, we further divide the space of each region into multiple subspaces by randomly selecting voxels within the respective regions.

Let $\mathbb{V}_i^r \subset \mathscr{V}^r$ be a randomly sampled voxel subspace from the $r$-th region's voxel space. We construct a set of random subspaces $\{\mathbb{V}_i^r\}_{i=1,\ldots,N^r}$ through an independent random subsampling procedure. Note that as we sample with replacement, the subspaces of $\mathbb{V}_i^r$ and $\mathbb{V}_j^r$ can be overlapping. By constructing a large number $N^r$ of random subspaces $\{\mathbb{V}_i^r\}_{i=1,\ldots,N^r}$ whose dimensions are manageable based on our dataset for training the model described below, the constructed set of random subspaces can effectively cover the original spaces in the respective regions. This procedure is repeated for every region $r \in \{1,\ldots,R\}$ by setting $N^r$ proportional to the total number of voxels in the $r$-th region, which is denoted as $|\mathscr{V}^r|$.

Based on the constructed set of random subspaces, we wish to learn complex nonlinear relationships between voxels. To this end, for each subspace $\mathbb{V}_i^r$, we construct a DNN (Bengio et al., 2009), which has excellent ability for discovering nonlinear relationships between input variables. Because the input variables for our DNNs are randomly selected from the regional input space $\mathscr{V}^r$, we refer to our network as rDNN.[6] Let $g_i^r$ be the rDNN for a random subspace $\mathbb{V}_i^r$ of the region $r$. Then, we train a set of rDNNs $\{g_i^r\}_{i=1,\ldots,N^r}$ such that each rDNN outputs the probability of the clinical labels for a target task (*e.g.*, AD vs. CN, MCI vs. CN, or pMCI vs. sMCI) by taking the volumetric measurements of voxels in each random subspace as inputs.

$$g_i^r(\mathbb{V}_i^r) = \begin{bmatrix} P(\mathscr{C}_1|\mathbb{V}_i^r) \\ P(\mathscr{C}_2|\mathbb{V}_i^r) \end{bmatrix} \tag{1}$$

where '$\mathscr{C}_1$' and '$\mathscr{C}_2$' denote the clinical labels of the later and earlier stages of the AD progression spectrum for a target task, respectively (*i.e.*, AD ($\mathscr{C}_1$) vs. CN ($\mathscr{C}_2$), MCI ($\mathscr{C}_1$) vs. CN ($\mathscr{C}_2$), or pMCI ($\mathscr{C}_1$) vs. sMCI ($\mathscr{C}_2$)). It is straightforward to extend this representation to a multi-class task, as shown in our experiments (*e.g.*, AD ($\mathscr{C}_1$) vs. MCI ($\mathscr{C}_2$) vs. CN ($\mathscr{C}_3$)). Regarding the size of a random subspace $|\mathbb{V}_i^r|$ (*i.e.*, number of input voxels or variables in an rDNN), we set the size empirically[7] and apply the same size for all rDNNs, regardless of region. With a fixed input size for an rDNN, we compensate for the potentially higher complexity of large regions by constructing a large number of rDNNs with $N^r$ proportional to the size of regions. Eq. (1) describes the class likelihoods or probabilities estimated for the subspace $\mathbb{V}_i^r$.

Because each rDNN outputs class likelihoods based on a limited

[6] The details of our network architecture is presented in Section 4.

[7] We set the size 200 in our experiments.

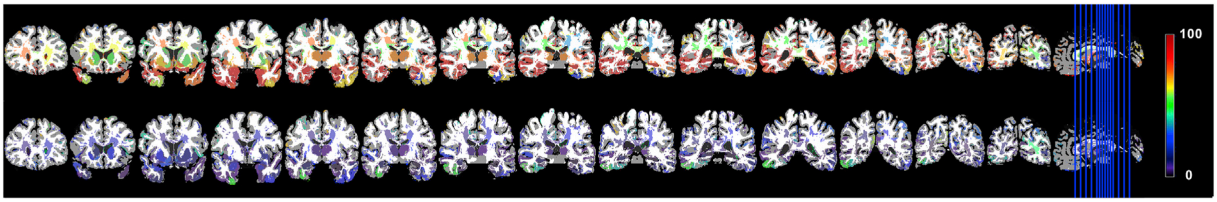


**Fig. 3.** Example regional abnormality maps from a subject with AD (top) and cognitively normal subject (bottom). We colored the voxels based on their *p*-values for a group comparison between AD and CN.

amount of information (*i.e.*, a small subset of the regional voxels), it is necessary to integrate the distributed information of a region to derive regional representative values. Therefore, from a set of voxel sets $\{\mathbb{V}_i^r\}_{i=1,\ldots,N^r}$ and their respective rDNNs $\{g_i^r(\mathbb{V}_i^r)\}_{i=1,\ldots,N^r}$, we derive a consensus vector for the *r*-th region by taking the average of the output class likelihoods as follows:

$$G^r(\mathscr{V}^r) = \frac{1}{N^r}\sum_{i=1}^{N^r} g_i^r(\mathbb{V}_i^r) = \begin{bmatrix} P(\mathscr{C}_1|\mathscr{V}^r) \\ P(\mathscr{C}_2|\mathscr{V}^r) \end{bmatrix}. \tag{2}$$

Finally, we consider one of the elements in the consensus vector as the relative abnormality measurement $a(r)$ of the *r*-th region for the target task. Specifically, in different classification scenarios, we regard the likelihood of the clinical stage residing in a late position on the AD progression spectrum as the abnormality representation. In other words, in AD vs. CN, MCI vs. CN, and pMCI vs. sMCI classification, the probability of AD, MCI, or pMCI, respectively, is regarded as the abnormality representation (see Fig. 3).[8]

### 3.2. *Brain-wise feature extraction and classifier learning*

As we trained one DNN for each individual ROI in a supervised manner, when patterns of an ROI are not separable between classes, *i.e.*, mostly not affected by the disease, the corresponding DNN would produce similar output values or class probabilities for the target classes, representing high ambiguity. Note that we have a second-phase classifier that jointly considers abnormal representations of all ROIs, assigning different weights by possibly ignoring those non-separable ROIs from the final decision.

To make a clinical decision for the entire brain, we must integrate the regional features. Because the probabilities estimated by Eq. (2) can be regarded as the likelihoods of belonging to the respective clinical labels, we consider these estimates as a high-level representation for a brain-wise pattern classification. Therefore, for information integration, we simply concatenate the region-based abnormality representations into a vector as follows:

$$\mathbf{a} = [a(1)\cdots a(r)\cdots a(R)]^\top \in \mathbb{R}^R. \tag{3}$$

Considering that the regional abnormality measurement implicitly represents a nonlinear relationship between different voxel-level information, the brain-wise classifier focuses on learning relationships between regions based on our high-level representation (*i.e.*, regional abnormality representation).

For $M$ number of training samples (*i.e.*, subjects in this case), we extract the feature vectors $\{\mathbf{a}_s\}_{s=1,\ldots,M}$ described in Section 3.1 and Eq. (3). From the pairs of extracted imaging features and clinical labels of subjects (*i.e.*, $\{(\mathbf{a}_s, l_s)\}_{s=1,\ldots,M}$), we train a brain-wise classifier to make a decision regarding the subject's clinical status. For a classifier, we use a linear support vector machine (SVM), which has been widely used in the literature (Liu et al., 2012a, 2018), although any classifier could be used without loss of generality.[9]

Overall, from a methodological perspective, our method can be understood as a combination of machine learning techniques as follows. (1) Each rDNN $g_i^r$ can be regarded as a "weak classifier" for the target task because it outputs class probabilities by exploiting a limited amount of information in a subspace of the original brain space. (2) The set of rDNNs $\{g_i^r\}_{i=1,\ldots,N^r}$ can be thought of as a feature descriptor for the *r*-th region with a mixture of experts by considering each of the rDNNs as an expert. (3) When considering each voxel set in $\mathbb{V}_i^r$ as an instance of the *r*-th region, we use $N^r$ number of instances for the *r*-th region in a single MRI image, which represents a type of multi-instance learning (Herrera et al., 2016).

## 4. Experimental settings and results

In our experiments, we considered three binary classification problems (*i.e.*, AD vs. CN, MCI vs. CN, and pMCI vs. sMCI) and one three-class classification problem (*i.e.*, AD vs. MCI vs. CN). For MCI vs. CN classification, we labeled both pMCI and sMCI as MCI. Because of the limited number of samples, we applied a 10-fold cross-validation technique for each classification problem and took the average of the 10 cross-validation results for performance comparisons[10]. To evaluate the effect of each main component in our proposed framework, we also conducted additional experiments accordingly.

### 4.1. *Experimental settings*

#### 4.1.1. *Random subspace construction*

Prior to exploiting the random subspace method, we first preselected voxels with statistical significance (*i.e.*, *p*-value) through a group comparison based on a training set for the target task (*i.e.*, AD vs. CN, MCI vs. CN, and pMCI vs. sMCI). Instead of using the conventional small values of 0.001 or 0.005 for a statistical significance thresholds, to better exploit potential multivariate relationships, where each variable shows a low group difference, but multiple variables may show a high group difference when considered jointly, we set the threshold conservatively as $p = 0.05$. Voxels of small $p < 0.05$ in group comparison were then considered in constructing random subspaces $\{\mathbb{V}_i^r\}$ in each region $r \in \{1, \ldots, R\}$.

We determined the optimal size of a random subspace $|\mathbb{V}_i^r|$ to be 200 based on preliminary experiments by varying the size of the random subspaces in $\{100, 200, 300, 400\}$ for all regions. The performance changes based on the different sizes of random subspaces are provided in the appendix (Table B1). As described in Section 3.2, we set the number of random subspaces $N^r$ to $Round(|\mathscr{V}^r|/|\mathbb{V}_i^r|) \times 3$, which is proportional to the total voxel size in the *r*-th region. Therefore, for different regions, we generated different numbers of random subspaces by considering the

[8] Since the output probabilities in deep neural networks sum to one, it is enough to consider only one value in a binary classification task.

[9] We also conducted with a multi-layer perceptron for comparison in our experiments.

[10] To see the applicability of our method to independent subject groups, we also conducted experiments by separating the study subjects into two independent groups. The results are presented in Appendix C.

difference in size between regions.

#### 4.1.2. Constructing and training rDNNs and a classifier

For each random voxel set, we constructed a three-layer DNN with an architecture of 200 nodes (input layer), 300 nodes (hidden layer), 60 nodes (hidden layer), and two nodes (output layer).[11] To improve network training, we pre-trained our network using greedy layer-wise pre-training by constructing stacked denoising auto-encoders (SDAEs) (Vincent et al., 2010). We used an encoder with an architecture of 200 nodes (input layer), 300 nodes (hidden layer), and 60 nodes (hidden layer), and a decoder with an architecture of 60 nodes (hidden layer), 300 nodes (hidden layer), and 200 nodes (output layer). The network is trained with an epoch of 120, a mini-batch size of 50, and weight decay of $10^{-4}$ to minimize the $\ell_2$ loss function. Note that an SDAE extracts robust and hierarchical feature representations from data in an unsupervised manner. The hierarchical properties of the SDAE are known to be helpful for representing highly nonlinear and complicated patterns in neuroimaging data (Suk et al., 2015). We then took the encoder portion of an SDAE and added a classification layer with a softmax function on top to create an rDNN. The rDNNs were then fine-tuned via stochastic gradient descent with a tanh activation function for all hidden layers, dropout rate of 0.5 for hidden layers, learning rate of 0.003, momentum of 0.9, and epoch length of 120. Regarding the brain-wise classifier with a linear SVM, we performed a 5-fold nested cross-validation for hyper-parameter setting (*i.e.*, the soft-margin parameter $C \in \{10^{-5}, 10^{-4}, \cdots, 10^{5}\}$). We used the DeepLearnToolbox[12] and libsvm[13] public packages for our experiments.

#### 4.1.3. Comparative methods

To determine the efficacy and validity of our method, we conducted various experiments by comparing our method to existing methods from the literature. Specifically, we considered three different methods characterized as follows:

- Regional mean volume (RMV) (Zhang et al., 2011): This method is directly comparable to our method in terms of feature representation. For each region, this method uses the low-level mean volume as a feature, whereas our method uses the high-level abnormality representation obtained by our rDNNs.
- Hierarchical feature fusion (HFF) (Liu et al., 2014a): This method gradually integrates features from a number of cubical local patches extracted from a GM density map. Based on the original work, we resized the GM density maps to 64 ×64 × 64 voxels for computational efficiency and extracted patches of 11 ×11 × 11 voxels in size, which were fed into a set of SVM classifiers.
- Regional abnormality representation with random forest (RF-RAR): To evaluate the effects of a neural network as a weak classifier in the proposed method, a random forest classifier was used as a weak classifier in this comparative method. In this method, an ensemble of random forest classifiers exploits corresponding random subspaces and the ensemble result is used for regional abnormality representation, similar to our method. To reflect the differences in size between regions, we set the number of trees proportional to the size of the regions (*i.e.*, $N^r = Round(|\mathscr{V}^r|/|\mathbb{V}_i^r|) \times 100$), meaning it uses about 33 times the number of randomized classifiers compared to our number of rDNNs per region. The maximum depth of a tree was set to 12 and a Gini index was used as an impurity function. We used the scikit-learn software[14] to train the random forest classifiers.

For all compared methods, we used a linear SVM for brain-wise classification.

### 4.2. Performance comparison

For performance comparison, we considered four different metrics (*i.e.*, accuracy, sensitivity, specificity, and area under the receiver operating curve (AUC)) that are commonly considered in AD/MCI diagnosis. The results for the four competing methods are listed in Table 3 and Table C2. It is remarkable that our method achieved very promising results on all three tasks. Specifically, for the tasks of MCI vs. CN and pMCI vs. sMCI, which are of high importance for proper clinical treatment, our method obtained accuracies of 89.22±4.13% (MCI vs. CN) and 88.52±5.65% (pMCI vs. sMCI), with AUCs of 0.9573 (MCI vs. CN) and 0.9568 (pMCI vs. sMCI). To the best of our knowledge, these are the best performance results reported in the literature (Rathore et al., 2017) for these classification tasks over the ADNI dataset. Our proposed method clearly outperformed the competing methods in all three binary classification tasks by large margins.

When comparing our method to the RMV method (Zhang et al., 2011), the superior performance of our method stems from the regional abnormality representation, which is the high-level abstract representation obtained by our rDNNs. Such representations have seen many successful applications (LeCun et al., 2015), where the abstract feature representations learned by deep learning methods help to enhance classification performance. Regarding RF-RAR and our method, both methods exploit the random subspace technique and region-based abnormality representations for AD/MCI identification. However, while RF-RAR represents regional abnormality based on the low-level voxel intensities (*i.e.*, volumetric measure), our method uses the complex nonlinear relationships between voxels (*i.e.*, high-level information) learned by DNNs. Based on the performance in Table 3 and Table C2, one can see the positive effect of using DNNs, which are powerful for learning hierarchical nonlinear relationships between input variables. Finally, when comparing our method to the HFF method (Liu et al., 2014a), which uses cubical patches for local feature representation and gradually combines the informative patches to form atypical patches over an entire brain, our method achieved superior performance in terms of all metrics. Furthermore, because our method defines patches based on anatomical knowledge, it is advantageous for interpreting regional statuses based on relationships between any symptomatic observations occurring in a subject.

It is also noteworthy that while the competing methods achieved reduced performance as the target task became more challenging, *i.e.* (AD vs. CN), → (MCI vs. CN) → (AD vs. MCI) → (pMCI vs. sMCI), the proposed method maintained reasonably high performance for all the tasks. Particularly, for the MCI vs. CN and pMCI vs. sMCI tasks, because these tasks focus on subtle structural changes, it is necessary to use more sophisticated feature representations. We believe that our method is able to discover such representations based on the power of DNNs.

To test the generalizability of our framework, we also performed a three-class classification task (*i.e.*, AD vs. MCI vs. CN) and compared the results to those of the competing methods.[15] In our method, we used both a linear SVM and multilayer perceptron (MLP) as brain-wise classifiers. The results, which are summarized in Table 4 and Table C4, reveal that our method combined with an MLP achieved the highest classification accuracy of 72.09±8.60% and our method with a linear SVM achieved the second-highest accuracy of 71.18±7.32%. However, there was no statistically significance difference ($p$-value = 0.7188) between the linear SVM and MLP. Our method outperformed the other methods by large margins.

[11] For details on our strategy to tune network architectures, refer to Appendix A.

[12] ‘https://github.com/rasmusbergpalm/DeepLearnToolbox.

[13] ‘https://www.csie.ntu.edu.tw/~cjlin/libsvm/

[14] http://scikit-learn.org/stable/index.html.

[15] Since the HFF method (Liu et al., 2014a) is applicable for binary classification only, we excluded it in this experiment.

**Table 3**
Performance comparison on four binary classification tasks: regional mean volume (RMV), hierarchical feature fusion (HFF), regional abnormality representation with random forest (RF-RAR).

| Tasks | Methods | Accuracy (%) | Sensitivity (%) | Specificity (%) | AUC |
|---|---|---|---|---|---|
| . AD vs. CN | RMV (Zhang et al., 2011) | 86.19±5.67 | 82.29±6.91 | 89.55±6.53 | 0.9215 |
| | HFF (Liu et al., 2014a) | 85.56±7.13 | 82.25±15.51 | 88.37±5.07 | 0.9220 |
| | RF-RAR (Lebedev et al., 2014) | 85.01±5.14 | 81.23±9.97 | 88.22±4.75 | 0.8516 |
| | Ours | **92.75±6.06** | **91.89±10.88** | **93.47±4.32** | **0.9804** |
| MCI vs. CN | RMV (Zhang et al., 2011) | 64.99±5.15 | 73.74±7.47 | 50.61±8.98 | 0.7023 |
| | HFF (Liu et al., 2014a) | 75.65±3.41 | 77.67±4.79 | 72.36±6.22 | 0.8638 |
| | RF-RAR (Lebedev et al., 2014) | 65.51±3.66 | 90.92±5.47 | 23.99±9.48 | 0.6687 |
| | Ours | **89.22±4.13** | **93.33±4.36** | **82.55±8.17** | **0.9573** |
| AD vs. MCI | RMV (Zhang et al., 2011) | 70.97±4.54 | 44.95±11.77 | 84.76±4.17 | 0.7202 |
| | HFF (Liu et al., 2014a) | 76.20±6.53 | 64.03±15.55 | 82.60±5.13 | 0.8389 |
| | RF-RAR (Lebedev et al., 2014) | 68.89±3.58 | 20.76±7.77 | 94.40±3.67 | 0.6927 |
| | Ours | **81.46±5.41** | **68.66±13.70** | **88.24±6.11** | **0.8954** |
| pMCI vs. sMCI | RMV (Zhang et al., 2011) | 61.41±10.82 | 51.25±15.25 | 68.96±12.48 | 0.6618 |
| | HFF (Liu et al., 2014a) | 63.40±9.25 | 46.13±9.22 | 76.32±15.07 | 0.6631 |
| | RF-RAR (Lebedev et al., 2014) | 67.30±3.66 | 90.91±5.47 | 22.71±9.48 | 0.6274 |
| | Ours | **88.52±5.65** | **87.50±9.77** | **89.22±7.47** | **0.9568** |

**Table 4**
Performance comparison on a three-class classification task: regional mean volume (RMV), regional abnormality representation with random forest (RF-RAR).

| Tasks | Methods | Accuracy (%) | Recall (%) | | Precision (%) | |
|---|---|---|---|---|---|---|
| | | | Macro | Micro | Macro | Micro |
| AD vs. MCI vs. CN | RMV (Zhang et al., 2011) | 53.69±6.08 | 55.88±7.39 | 37.095.47 | 53.69±6.08 | 36.91±5.81 |
| | RF-RAR (Lebedev et al., 2014) | 39.21±5.56 | **57.26±13.96** | 35.76±2.30 | 39.21±5.56 | 24.52±4.34 |
| | Ours | **71.18±7.32** | 54.97±8.65 | **55.70±8.73** | **72.00±6.93** | **71.18±7.32** |

## 5. Discussions

### 5.1. Effect of constructing random subspaces

One of the factors that differentiates our method from other region-based methods is that the proposed method divides the regional voxel spaces into multiple random subspaces. Considering the fact that some of the regional voxel spaces still contain large numbers of voxels (*i.e.*, variables) compared to the number of samples available for training, a computational method for modeling the relationships between all the voxels is highly likely to suffer from overfitting. In particular, the model is likely to focus only on the easy-to-learn features. However, in our random subspace construction approach, each classifier covers randomly selected sub-voxels from the original space, meaning the correlations between individual weak classifiers are reduced. This approach is also referred to as the feature/attribute bagging method (Barandiaran, 1998; Bryll et al., 2003). In the ensemble of individual weak classifiers, which were trained using the voxel intensities from their respective random voxel spaces, the combination of all learned classifiers enhances overall model stability and improves target task performance. This is a very useful property for medical image analysis, where feature dimensionality is typically high and training samples are very limited in number.

Our region-based abnormality representation is comparable to the regional mean volume that is commonly used in previous methods (Rathore et al., 2017). Although both representations use single scalar values for individual regions, from an information perspective, the regional mean volume is low-level information, whereas our abnormality representation is high-level information. Therefore, it is believed that multivariate analysis of a brain using our high-level information is more useful for model generalization, which improves clinical accuracy.

### 5.2. MLP as a brain-wise classifier

As stated in Section 4.1.2, in our framework, it is possible to use any classifier for a brain-wise classification with the regional abnormality representation as an input. Here, we present our experimental results obtained by replacing the linear SVM with an MLP in our framework. The main difference between a linear SVM and an MLP lies in the manner of defining a decision boundary in a feature space (*i.e.*, regional abnormality features in our work). While the linear SVM defines a separating hyperplane using a linear function, the MLP finds such a hyperplane non-linearly by considering potential nonlinear relationships between the input values. A performance comparison between proposed framework with SVM and MLP as a brain-wise classifier is provided in Table 5 and Fig. 4. For different tasks, the best performance was achieved by different classifiers. A statistical significance test[16] of differences in terms of accuracy yielded *p*-values of 0.9375 (AD vs. CN), 0.3926 (MCI vs. CN), 0.2188 (pMCI vs. sMCI), and 0.7188 (AD vs. MCI vs. CN). Therefore, for all tasks, there were no statistically significant differences in terms of accuracy. Based on this result, we believe that the regional feature representations provided to the brain-wise classifiers already include high-level information discovered by the non-linear operations in the rDNNs, meaning there was relatively little space for an MLP to outperform a linear SVM by considering abstract representations. Methodologically, the proposed framework with an MLP as a brain-wise classifier has the potential to take advantage of end-to-end learning and superior computational efficiency for inference. However, the linear SVM allows us to interpret and analyze underlying patterns, which is useful for AD/

**Table 5**
Performance comparison between a linear SVM and MLP in our framework for brain-wise classification.

| Tasks | Classifier | Accuracy (%) | Sensitivity (%) | Specificity (%) | AUC |
|---|---|---|---|---|---|
| AD vs. CN | SVM | 92.75±6.06 | 91.89±10.88 | 93.47±4.32 | 0.9804 |
| | MLP | 92.76± 6.05 | 92.95±7.55 | 92.61±6.50 | 0.9766 |
| MCI vs. CN | SVM | 89.22±4.13 | 93.33±4.36 | 82.55±8.17 | 0.9573 |
| | MLP | 88.39± 2.36 | 89.29±2.47 | 87.52±6.50 | 0.9499 |
| AD vs. MCI | SVM | 81.46±5.41 | 68.66±13.70 | 88.24±6.11 | 0.8954 |
| | MLP | 81.12±4.47 | 66.13±9.86 | 89.03±5.00 | 0.8917 |
| pMCI vs. sMCI | SVM | 88.52±5.65 | 87.50±9.77 | 89.22±7.47 | 0.9568 |
| | MLP | 89.84±4.71 | 87.21±8.59 | 92.89±3.85 | 0.9492 |

[16] We conducted a Wilcoxon signed-rank test.

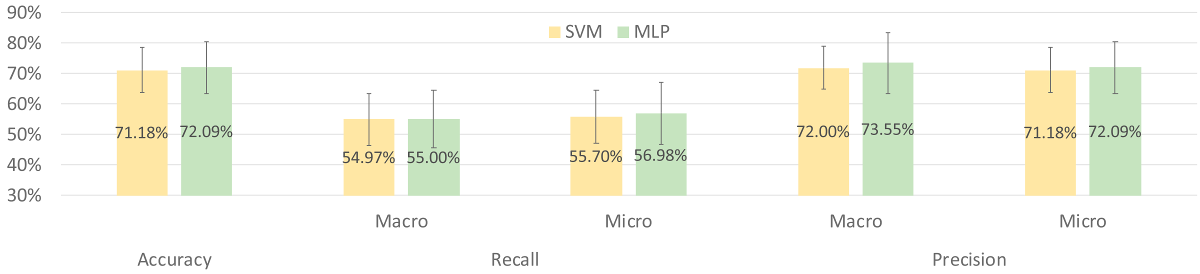


**Fig. 4.** Performance comparison between an SVM and MLP as brain-wise classifiers in a multi-class classification task.

MCI identification, based on its weight coefficients, whereas the MLP suffers from interpretation issues because it functions as a black box model.

### 5.3. Regional abnormality map

One of the main advantages of our regional abnormality representation is that it provides a means of interpreting or understanding the status of a subject's brain from the perspective of neurodegenerative pathology in the AD progression spectrum. This is because it is possible to map and visualize regional abnormalities in the brain space, as shown in Fig. 5. This map is what we refer to as the '*regional abnormality map*'. This map makes it straightforward to interpret regional statuses based on the probability that a region represents the later stages of the AD progression spectrum for a target task, as well as to draw potential relationships between symptomatic observations.

Fig. 5 presents regional abnormality maps of subjects from different groups (*i.e.*, AD, CN, pMCI, and sMCI). First, a comparison between the subjects from two groups for the tasks of AD vs. CN and pMCI vs. sMCI reveals that subjects in the later stages of the AD progression spectrum (*i.e.*, AD and pMCI for the respective tasks) have many regions with high abnormality scores (Fig. 5(a) and (c)) compared to the subjects from the counterpart groups (Fig. 5(b) and (d)). Second, it is noteworthy that our regional abnormality map reflects individual differences. In other words, there are different abnormality patterns between subjects with the same clinical label. Our regional abnormality map also indicates which parts of a brain seem to be at risk based on abnormality scores. For example, in Fig. 5(b), although all the subjects were identified as CN, the regional abnormality map of the subject on the top reveals that their temporal and subcortical areas (around the putamen) show similar patterns of those of subjects with AD, indicating that they may require further testing or care.

We also applied the rDNNs trained with samples from AD and CN subjects to subjects with MCI, including both pMCI and sMCI, and obtained their regional abnormality maps. Fig. 6 illustrates the averaged regional abnormality maps of the four groups in the AD spectrum based on the rDNNs for AD vs. CN. It is noteworthy that the regional abnormality scores over a brain show nearly linear relationships to the clinical status in the AD spectrum. That is to say, based on the averaged abnormality scores, the groups considered in this work can be sorted in the order of AD, pMCI, sMCI, and CN, which empirically validates the effectiveness of our regional abnormality representation.

### 5.4. Regional relevance to AD/MCI identification

Based on our experimental results, we demonstrated the superiority of our method to competing methods in terms of performance. However, such a comparison reveals little regarding the regional importance or relevance to a decision of our method (*i.e.*, no understanding of the learned model). Therefore, we investigated our trained classifier (*i.e.*, a linear SVM) to derive further insights regarding the output clinical decisions. Note that in terms of a decision-making process, the weight coefficients in a linear SVM carry the relevant information. By following Haufe et al.'s method (Haufe et al., 2014), we considered an activation pattern map that can be estimated by jointly considering the covariance of the input features $\Sigma_{\text{input}}$, (co)variance of the predicted values $\Sigma_{\text{prediction}}$, and learned weight coefficients $\mathbf{W}$ as follows:

$$\mathbf{A} \equiv \Sigma_{\text{input}} \mathbf{W}^{-1} \Sigma_{\text{prediction}}. \tag{4}$$

Fig. 7 presents the estimated activation pattern maps for each task, where the colors indicate the regional relevance to the later states in the AD spectrum. First, for all three classification tasks, our classifiers found that the temporal and subcortical regions are highly relevant to the later stages in the AD spectrum (*i.e.*, (a) AD, (b) MCI, and (c) pMCI) for each task. Second, depending on the task, the activation patterns change to reflect the pathological progress related to the AD spectrum stages. While most of the brain regions are positively related to AD in the case of AD vs. CN, the sensorimotor cortex and temporal lobules as well as the subcortical regions, are more closely related to MCI and pMCI for the classification tasks of MCI vs. CN and pMCI vs. sMCI.

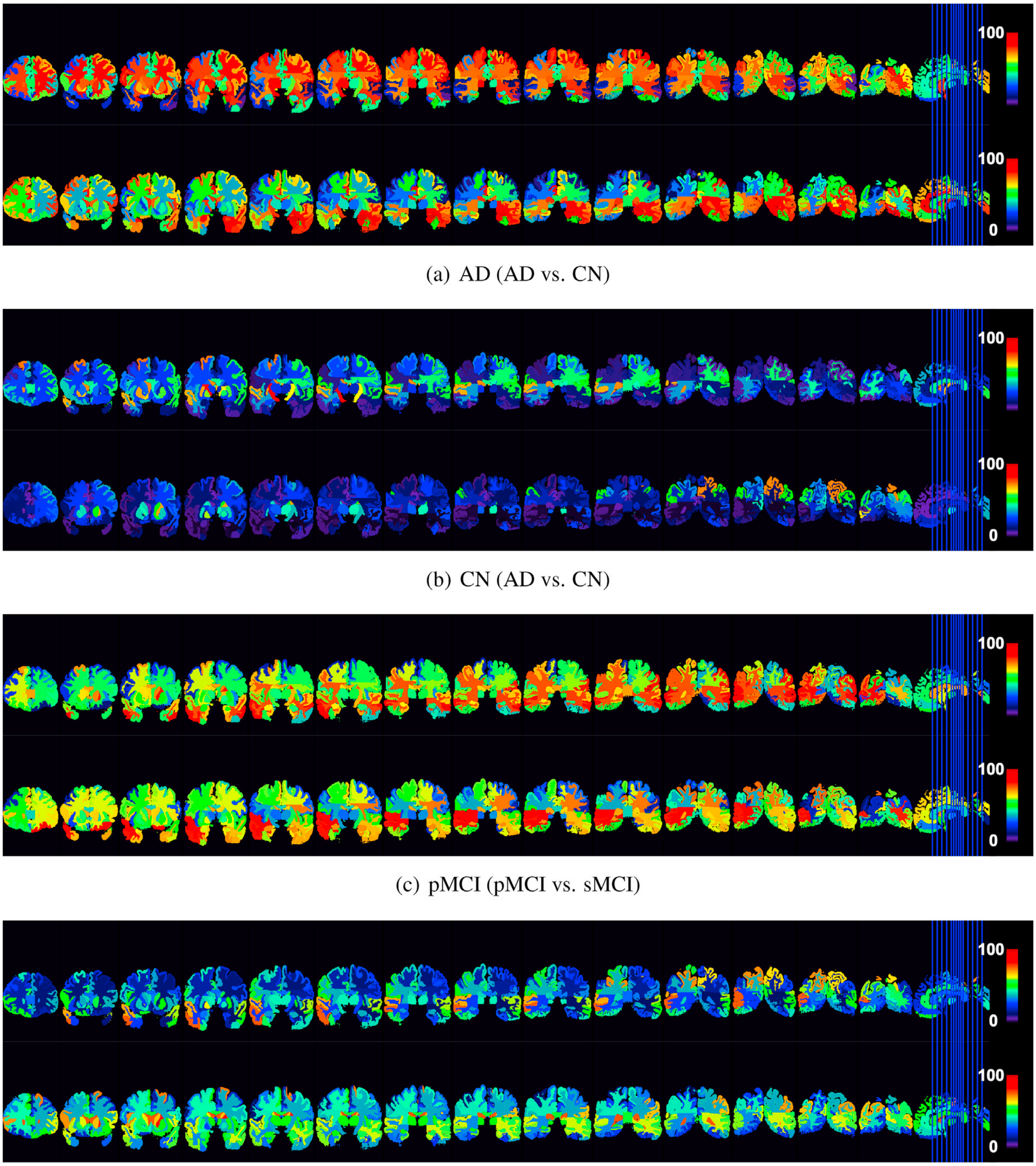
### 5.5. Single-region-based diagnosis

As described in Section 3.1, our method assesses regional abnormalities based on the probability of the clinical stage residing in a late position along the AD progression spectrum for each task. Therefore, regional abnormality can be used directly to make clinical decisions without any further processing. This corresponds to single-region-based diagnosis, which is comparable to diagnosis based on the entire brain. In Fig. 8, we present the averaged regional classification accuracies for different tasks in the brain space. In this figure, one can see that the white matter regions show higher accuracies than the GM regions. Our interpretation of this phenomenon is that because a single white matter region is anatomically adjacent to many GM regions, it includes the volumetric changes in all adjacent GM regions simultaneously.[17] It is also observable that the subcortical and temporal areas generally show high classification accuracies. However, these accuracies are still much lower than that of diagnosis based on the entire brain, as shown in Table 3 and Table C2.

### 5.6. Transferring knowledge of AD vs. CN to pMCI vs. sMCI

In the AD progression spectrum, the classes of pMCI and sMCI are considered to be close to those of AD and CN, respectively. Motivated by this assumption, we performed an additional experiment to discriminate pMCI from sMCI in the models (*i.e.*, rDNNs for regional abnormality representation and a linear SVM for classification) trained using AD and CN samples. By means of knowledge transfer (*i.e.*, use of the knowledge regarding AD and CN classification for pMCI and sMCI classification), we validated the usefulness of our regional abnormality representation. Fig. 9 and Fig. 10 present the regional abnormality maps of individual

[17] due to imperfectness of the registration method applied in our framework, the voxels lying along the boundaries between white and gray matter regions are more likely to be affected, especially regions with neural atrophy.



(a) AD (AD vs. CN)

(b) CN (AD vs. CN)

(c) pMCI (pMCI vs. sMCI)

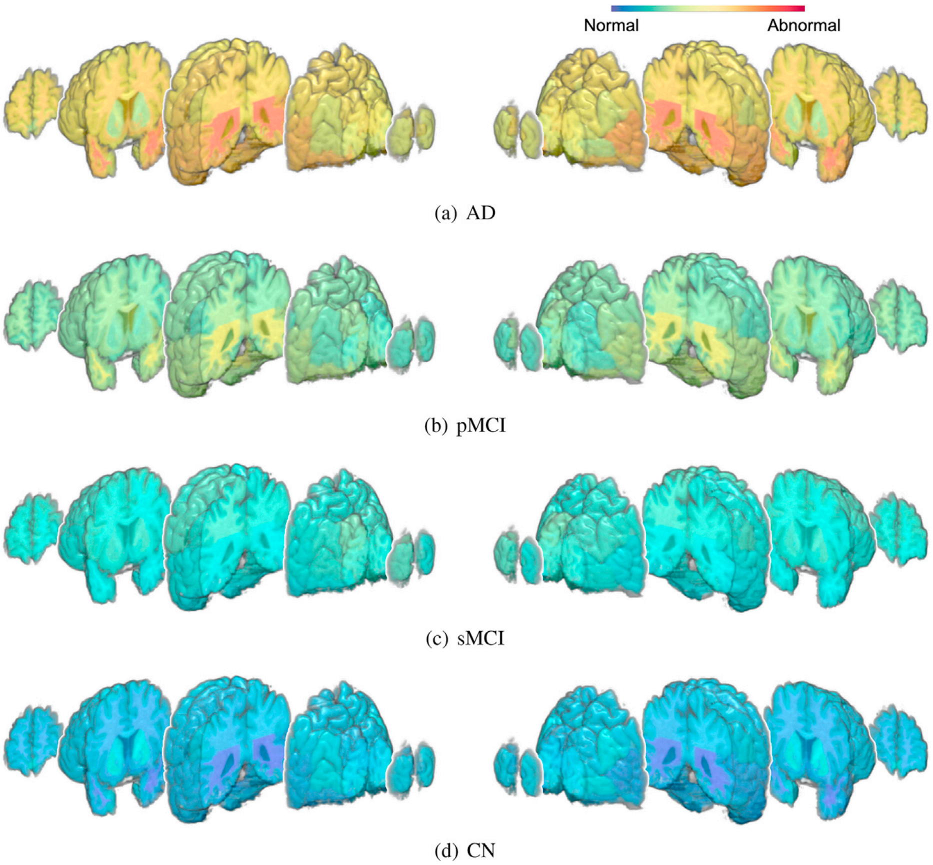
(d) sMCI (pMCI vs. sMCI)

**Fig. 5.** Examples of the regional abnormality map of individual subjects, estimated with the rDNNs from the tasks of (AD vs. CN) and (pMCI vs. sMCI), respectively.

subjects with pMCI and sMCI, and the averaged regional classification accuracy maps for pMCI and sMCI, respectively. With the regional abnormality representation and linear SVM classifier, we achieved an accuracy of 65.40%, sensitivity of 61.12%, sensitivity of 68.59%, and AUC of 0.7050. When comparing these results to those of the RMV (Zhang et al., 2011), HFF (Liu et al., 2014a), and RF-RAR (Lebedev et al., 2014) methods in Table 3 and Table C2, one can see that the knowledge transfer process was very effective because it resulted in the highest AUC, even though there was no training for this particular task.

### 5.7. *Use of the AAL template*

= We have also conducted experiments with the AAL template atlas, widely used as well in the AD literature. Basically, we have repeated all the experiments described above by randomly partitioning the dataset into two independent subject groups, *i.e.*, the one for training/validation and the other for testing. The results are summarized in Table 6 and Table 7. When comparing with the results obtained with Kabani et al.'s template, one noticeable thing is the performance degradation in all tasks except for AD vs. CN. Our understanding for the performance



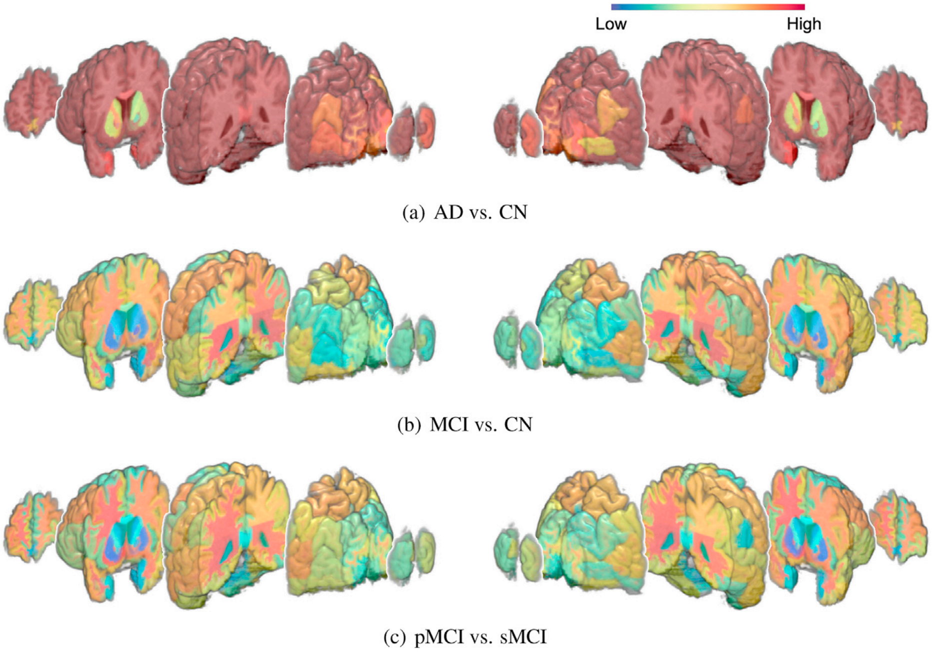
(a) AD

(b) pMCI

(c) sMCI

(d) CN

**Fig. 6.** Averaged regional abnormality maps of the groups as estimated by rDNNs trained on AD and CN samples from the AD progression spectrum. (left: left-view, right: right-view).



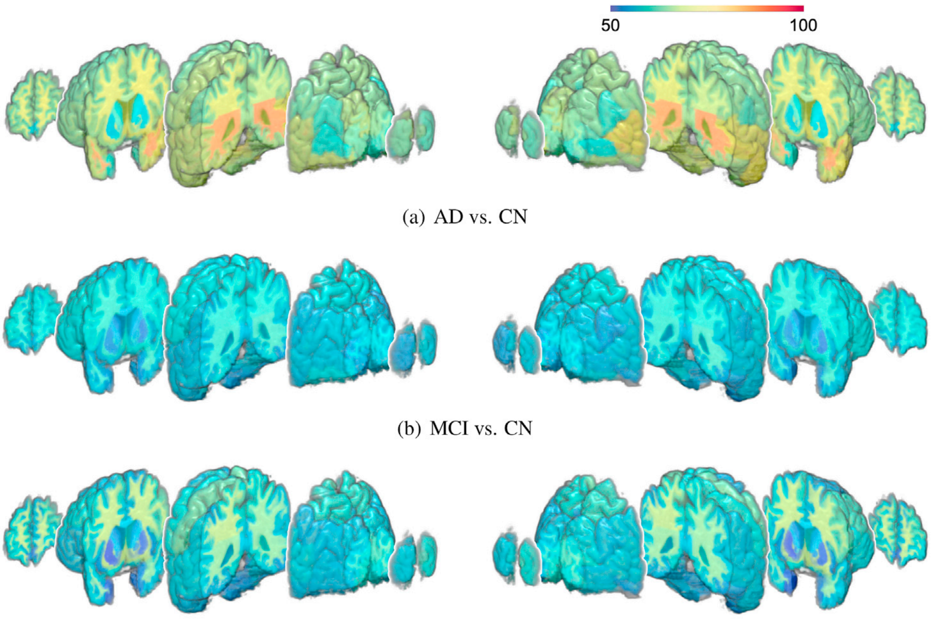
(a) AD vs. CN

(b) MCI vs. CN

(c) pMCI vs. sMCI

**Fig. 7.** Activation pattern maps estimated by the SVM classifiers trained for each classification task. The higher the values, the greater the relevance to the pathological class in the AD progression spectrum (left: left-view, right: right-view).

degradation is due to no use of the regions in the white matter. Specifically, the AAL template parcellates regions of the gray matter only, while the Kabani et al.'s template defines regions in both the gray and white matters. As visualized in Fig. 8 and explained in Section 5.5, the regions in white matter carry useful information in AD/MCI identification. We believe that this AAL-template based experiments further support the

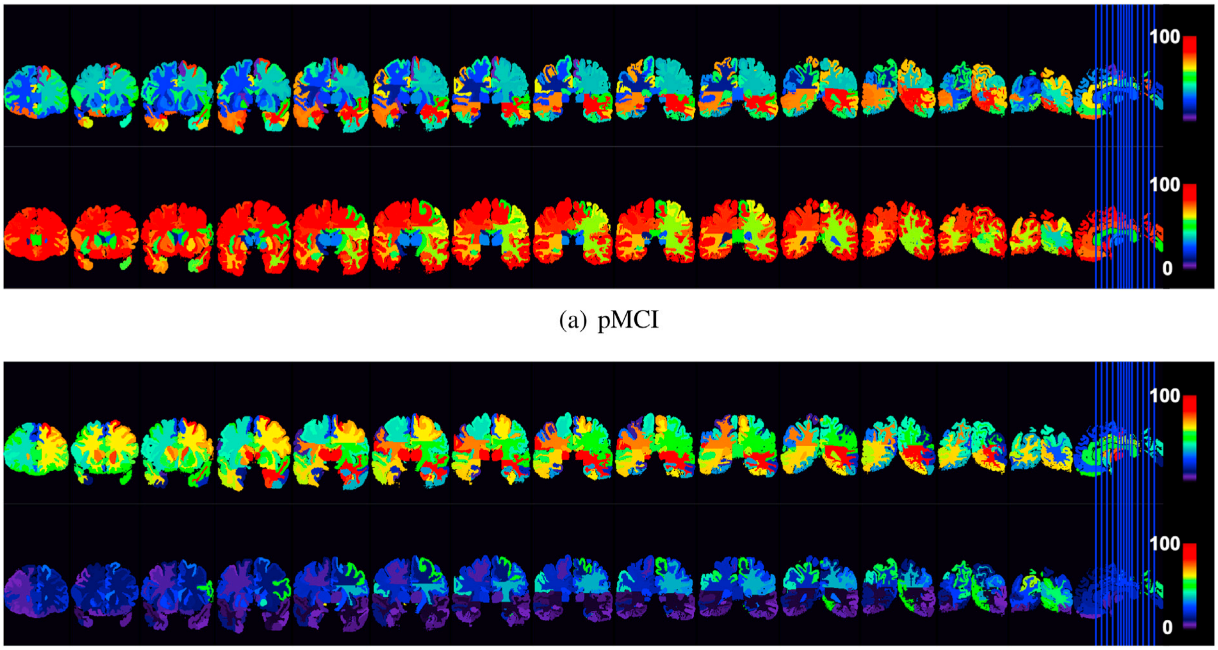


(a) AD vs. CN

(b) MCI vs. CN

(c) pMCI vs. sMCI

**Fig. 8.** Averaged regional classification accuracy maps for binary classification tasks. (left: left-view, right: right-view).



(a) pMCI

(b) sMCI

**Fig. 9.** Examples of the regional abnormality maps from individual pMCI and sMCI subjects, estimated by the rDNNs trained on AD vs. CN samples.

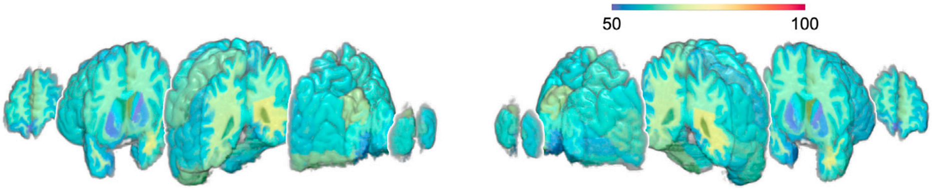


**Fig. 10.** Averaged regional classification accuracy map for pMCI and sMCI classification with rDNNs transferred from an AD and CN classification task (left: left-view, right: right-view).

**Table 6**
Performance of four binary classification tasks with images registered to the AAL template. Numbers in the parentheses denote the number of testing subjects for each class.

| Tasks | Methods | Accuracy (%) | Sensitivity (%) | Specificity (%) | AUC |
|---|---|---|---|---|---|
| AD (19) vs. CN (34) | RMV | 77.36 | 78.95 | 76.47 | 0.8576 |
| | Ours + SVM | 100 | 100 | 100 | 1 |
| AD (24) vs. MCI (47) | RMV | 70.17 | 30.00 | 91.89 | 0.6743 |
| | Ours + SVM | 70.18 | 25.00 | 94.59 | 0.7946 |
| MCI (45) vs. CN (30) | RMV | 57.38 | 68.42 | 39.13 | 0.5584 |
| | Ours + SVM | 63.93 | 60.53 | 69.57 | 0.7128 |
| pMCI (28) vs. sMCI (30) | RMV | 67.57 | 56.25 | 76.19 | 0.7649 |
| | Ours + SVM | 67.17 | 56.21 | 75.97 | 0.7627 |

**Table 7**
Performance of a three-class classification task with images registered to the AAL template. Numbers in the parentheses denote the number of testing subjects for each class.

| Tasks | Classifier | Accuracy (%) | Recall (%) | | Precision (%) | |
|---|---|---|---|---|---|---|
| | | | Macro | Micro | Macro | Micro |
| AD (19) vs. MCI (37) vs. CN (23) | RMV | 44.30 | 28.52 | 28.46 | 43.89 | 44.30 |
| | Ours + SVM | 48.10 | 31.43 | 31.67 | 41.01 | 48.10 |

importance of information from white matter regions. Meanwhile, it is noteworthy that our method still outperforms the competing method of RMV.

## 6. Conclusion

AD or MCI identification based on structural MRI has been a long-standing research issue and various types of related analysis methods have been studied. In the paper, we proposed a novel method that systematically integrates voxel-based, region-based, and patch-based approaches into a unified framework. From a machine learning perspective, our method exploits a random subspace method, nonlinear feature representation with DNNs, and ensemble method to enhance classification performance. By comparing the proposed method to existing methods through experiments on the ADNI dataset, we validated the effectiveness of the proposed method, which achieved superior performance by large margins. Additionally, based on thorough analysis of the region-based abnormality map produced during the inference step in the proposed framework, we determined that our region-based abnormality method is visualizable and interpretable for the sake of further analysis of clinical diagnosis in terms of the AD progression spectrum. Finally, it is also noteworthy that we estimated the activation pattern maps for each task by combining the regional abnormality representations, learned SVM weight coefficients, and label predictions. From the activation pattern maps, we could identify potential imaging biomarker regions, which are positively related to clinical states in different tasks.

## Acknowledgement

This research was supported by the Bio & Medical Technology Development Program of the National Research Foundation (NRF) funded by the Korean government (MSIT) (NRF - 2016M3A9E9941946), and Institute for Information and Communications Technology Planning and Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2017-0-01779, A machine learning and statistical inference framework for explainable artificial intelligence).

## Appendix A. Tuning Regional Deep Neural Network Architectures

Since ROIs have different number of voxels, it is hard to tune hyperparameters related to a network architecture for all ROIs separately. In our work, instead of setting network hyperparameters for each ROI separately, we took a streamlined strategy, obtained empirically from the preliminary experiments with ROIs of hippocampus, entorhinal cortex, and thalamus, for example. First, we confined the input dimension to be relatively low in $\{100, 200, 300, 400\}$ for all ROIs, instead we compensated such a constraint by constructing a number of the same architecture DNNs proportional to the size of each ROI, but having each DNN with different randomly selected input variables. Second, given an input dimension, the number of units in the first hidden layer was set to be approximately 50% larger than the input dimension and the number of units in the second hidden layer was set to be approximately 30% of the input dimension. With this streamlined strategy, we had a manageable number of hyperparameter value sets to apply and choose from.

## Appendix B. Effect of the Random Subspace Size

**Table B.1**
Performance analysis of three binary classification tasks with different sizes of random subspaces in our framework. A linear SVM classifier was used to make a decisions.

| Tasks | Methods | Accuracy (%) | Sensitivity (%) | Specificity (%) | AUC |
|---|---|---|---|---|---|
| AD vs. CN | Ours (RS Size = 100) | 86.49±7.59 | 88.89±9.35 | 88.98±9.39 | 0.9521 |
| | Ours (RS Size = 200) | **92.75±6.06** | **91.89±10.88** | **93.47±4.32** | **0.9804** |
| | Ours (RS Size = 300) | 86.66±4.39 | 84.37±8.68 | 88.65±5.48 | 0.9419 |
| | Ours (RS Size = 400) | 86.18±4.01 | 85.37±7.60 | 86.90±5.80 | 0.9309 |
| MCI vs. CN | Ours (RS Size = 100) | 88.23±2.52 | 92.23±2.05 | 81.66±7.03 | 0.9535 |
| | Ours (RS Size = 200) | 89.22±4.13 | 93.33±4.36 | 82.55±8.17 | 0.9573 |
| | Ours (RS Size = 300) | 88.04±5.35 | 90.63±5.38 | 83.81±9.25 | 0.9473 |
| | Ours (RS Size = 400) | **90.05±2.83** | **94.64±3.13** | **82.55±5.75** | **0.9579** |
| pMCI vs. sMCI | Ours (RS Size = 100) | 86.37±3.43 | 83.12±8.36 | 88.83±6.56 | 0.9432 |
| | Ours (RS Size = 200) | **88.52±5.65** | **87.50±9.77** | 89.22±7.47 | **0.9568** |
| | Ours (RS Size = 300) | 86.36±3.43 | 83.12±8.36 | 88.83±6.56 | 0.9432 |
| | Ours (RS Size = 400) | 87.72±3.53 | 81.25±7.10 | **93.48±3.22** | 0.9460 |

## Appendix C. Experiments with Two Independent Subject Groups

**Table C.2:**
Performance comparison on four binary classification tasks: regional mean volume (RMV), hierarchical feature fusion (HFF), regional abnormality representation with random forest (RF-RAR). Numbers in the parentheses denote the number of testing subjects for each class.

| Tasks | Methods | Accuracy (%) | Sensitivity (%) | Specificity (%) | AUC |
|---|---|---|---|---|---|
| AD (30) vs. CN (24) | RMV (Zhang et al., 2011) | 81.13 | 73.91 | 86.67 | 0.8913 |
| | HFF (Liu et al., 2014a) | 98.11 | 95.65 | 100 | 0.9667 |
| | RF-RAR (Lebedev et al., 2014) | 79.25 | 73.91 | 83.33 | 78.62 |
| | Ours | **100** | **100** | **100** | **1** |
| AD (24) vs. MCI (47) | RMV (Zhang et al., 2011) | 76.06 | 50.00 | 89.36 | 0.8076 |
| | HFF (Liu et al., 2014a) | 71.83 | 50.00 | 82.98 | 0.7881 |
| | RF-RAR (Lebedev et al., 2014) | 70.42 | 25.00 | 93.62 | 0.7097 |
| | Ours | **87.32** | **66.67** | **97.87** | **0.9202** |
| MCI (45) vs. CN (30) | RMV (Zhang et al., 2011) | 68.00 | 83.33 | 40.74 | 0.7631 |
| | HFF (Liu et al., 2014a) | 84.00 | 85.42 | 81.48 | 0.8958 |
| | RF-RAR (Lebedev et al., 2014) | 69.84 | 42.86 | 91.43 | 0.6667 |
| | Ours | **100** | **100** | **100** | **1** |
| pMCI (28) vs. sMCI (30) | RMV (Zhang et al., 2011) | 73.01 | 82.14 | 65.71 | 0.7520 |
| | HFF (Liu et al., 2014a) | 60.32 | 28.57 | 85.71 | 0.6490 |
| | RF-RAR (Lebedev et al., 2014) | 63.49 | 39.29 | 82.86 | 0.6304 |
| | Ours | **95.24** | **89.29** | **100** | **0.9949** |

**Table C.3**
Performance comparison between a linear SVM and MLP on a three-class classification task. Numbers in the parentheses denote the number of testing subjects for each class.

| Tasks | Classifier | Accuracy (%) | Recall (%) | | Precision (%) | |
|---|---|---|---|---|---|---|
| | | | Macro | Micro | Macro | Micro |
| AD (22) vs. MCI (47) vs. CN (31) | SVM | 77.00 | 60.87 | 62.60 | 78.45 | 77.00 |
| | MLP | 76.00 | 58.96 | 61.29 | 78.16 | 76.00 |

**Table C.4**
Performance comparison on a three-class classification task: regional mean volume (RMV), regional abnormality representation with random forest (RF-RAR). Numbers in the parentheses denote the number of testing subjects for each class.

| Tasks | Methods | Accuracy (%) | Recall (%) | | Precision (%) | |
|---|---|---|---|---|---|---|
| | | | Macro | Micro | Macro | Micro |
| AD (22) vs. MCI (47) vs. CN (31) | RMV (Zhang et al., 2011) | 52.00 | 35.30 | 35.14 | 52.50 | 52.00 |
| | RF-RAR (Lebedev et al., 2014) | 41.00 | 22.41 | 21.08 | 28.55 | 43.21 |
| | Ours + SVM | 77.00 | 60.87 | 62.60 | 78.45 | 77.00 |

## Appendix D. List of the ROIs

**Table D.5**
List of the ROIs in Kabani et al.'s atlas.

| ID | ROI Name | ID | ROI Name |
|---|---|---|---|
| 1 | medial front-orbital gyrus right | 48 | middle temporal gyrus left |
| 2 | middle frontal gyrus right | 49 | lingual gyrus left |
| 3 | lateral ventricle left | 50 | superior frontal gyrus left |
| 4 | insula right | 51 | nucleus accumbens left |
| 5 | precentral gyrus right | 52 | occipital lobe WM left |
| 6 | lateral front-orbital gyrus right | 53 | postcentral gyrus left |
| 7 | cingulate region right | 54 | inferior frontal gyrus right |
| 8 | lateral ventricle right | 55 | precentral gyrus left |
| 9 | medial frontal gyrus left | 56 | temporal lobe WM left |
| 10 | superior frontal gyrus right | 57 | medial front-orbital gyrus left |
| 11 | globus palladus right | 58 | perirhinal cortex right |
| 12 | globus palladus left | 59 | superior parietal lobule right |
| 13 | putamen left | 60 | lateral front-orbital gyrus left |
| 14 | inferior frontal gyrus left | 61 | perirhinal cortex left |
| 15 | putamen right | 62 | inferior temporal gyrus left |
| 16 | frontal lobe WM right | 63 | temporal pole left |
| 17 | parahippocampal gyrus left | 64 | entorhinal cortex left |
| 18 | angular gyrus right | 65 | inferior occipital gyrus right |
| 19 | temporal pole right | 66 | superior occipital gyrus left |
| 20 | subthalamic nucleus right | 67 | lateral occipitotemporal gyrus right |
| 21 | nucleus accumbens right | 68 | entorhinal cortex right |

(continued on next column)

**Table D.5** (*continued*)

| ID | ROI Name | ID | ROI Name |
|---|---|---|---|
| 22 | uncus right | 69 | hippocampal formation left |
| 23 | cingulate region left | 70 | thalamus left |
| 24 | fornix left | 71 | parietal lobe WM right |
| 25 | frontal lobe WM left | 72 | insula left |
| 26 | precuneus right | 73 | postcentral gyrus right |
| 27 | subthalamic nucleus left | 74 | lingual gyrus right |
| 28 | posterior limb of internal capsule inc. cerebral peduncle left | 75 | medial frontal gyrus right |
| 29 | posterior limb of internal capsule inc. cerebral peduncle right | 76 | amygdala left |
| 30 | hippocampal formation right | 77 | medial occipitotemporal gyrus left |
| 31 | inferior occipital gyrus left | 78 | parahippocampal gyrus right |
| 32 | superior occipital gyrus right | 79 | anterior limb of internal capsule right |
| 33 | caudate nucleus left | 80 | middle temporal gyrus right |
| 34 | suramarginal gyrus left | 81 | occipital pole right |
| 35 | anterior limb of internal capsule left | 82 | corpus callosum |
| 36 | occipital lobe WM right | 83 | amygdala right |
| 37 | middle frontal gyrus left | 84 | inferior temporal gyrus right |
| 38 | superior parietal lobule left | 85 | superior temporal gyrus |
| 39 | caudate nucleus right | 86 | middle occipital gyrus left |
| 40 | cuneus left | 87 | angular gyrus left |
| 41 | precuneus left | 88 | medial occipitotemporal gyrus right |
| 42 | parietal lobe WM left | 89 | cuneus right |
| 43 | temporal lobe WM right | 90 | lateral occipitotemporal gyrus |
| 44 | suramarginal gyrus right | 91 | thalamus right |
| 45 | superior temporal gyrus left | 92 | occipital pole left |
| 46 | uncus left | 93 | fornix right |
| 47 | middle occipital gyrus right | | |

## References

Adeli, E., Thung, K.-H., An, L., Wu, G., Shi, F., Wang, T., Shen, D., 2019. Semi-supervised discriminative classification robust to sample-outliers and feature-noises. IEEE Trans. Pattern Anal. Mach. Intell. 41 (2), 515–522.

Albert, M.S., DeKosky, S.T., Dickson, D., Dubois, B., Feldman, H.H., Fox, N.C., Gamst, A., Holtzman, D.M., Jagust, W.J., Petersen, R.C., et al., 2011. The diagnosis of mild cognitive impairment due to Azheimer's disease: recommendations from the National Institute on Aging - Alzheimer's Association workgroups on diagnostic guidelines for Azheimer's disease. Alzheimer's Dementia 7 (3), 270–279.

An, L., Adeli, E., Liu, M., Zhang, J., Lee, S.-W., Shen, D., 2017. A hierarchical feature and sample selection framework and its application for Alzheimer's disease diagnosis. Sci. Rep. 7, 45269.

Association, A., et al., 2017. 2017 Alzheimer's disease facts and figures. Alzheimer's Dementia 13 (4), 325–373.

Barandiaran, I., 1998. The random subspace method for constructing decision forests. IEEE Trans. Pattern Anal. Mach. Intell. 20 (8).

Barker, W.W., Luis, C.A., Kashuba, A., Luis, M., Harwood, D.G., Loewenstein, D., Waters, C., Jimison, P., Shepherd, E., Sevush, S., et al., 2002. Relative frequencies of Azheimer's disease, lewy body, vascular and frontotemporal dementia, and hippocampal sclerosis in the state of Florida brain bank. Alzheimers Dis. Assoc. Disord. 16 (4), 203–212.

Bengio, Y., et al., 2009. Learning deep architectures for AI. Found. Trends Mach. Learn. 2 (1), 1–127.

Blennow, K., de Leon, M.J., Zetterberg, H., 2006. Alzheimer's disease. The Lancet 368, 387–403.

Bryll, R., Gutierrez-Osuna, R., Quek, F., 2003. Attribute bagging: improving accuracy of classifier ensembles by using random feature subsets. Pattern Recognit. 36 (6), 1291–1302.

Cormen, T.H., Leiserson, C.E., Rivest, R.L., Stein, C., 2009. Introduction to Algorithms. MIT press.

Davatzikos, C., Genc, A., Xu, D., Resnick, S.M., 2001. Voxel-based morphometry using the RAVENS maps: methods and validation using simulated longitudinal atrophy. Neuroimage 14 (6), 1361–1369.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L., 2009. ImageNet: a large-scale hierarchical image database. In: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, Ieee, pp. 248–255.

Esmaeilzadeh, S., Belivanis, D.I., Pohl, K.M., Adeli, E., 2018. End-to-end Alzheimer's disease diagnosis and biomarker identification. In: International Workshop on Machine Learning in Medical Imaging. Springer, pp. 337–345.

Fan, Y., Shen, D., Gur, R.C., Gur, R.E., Davatzikos, C., 2007. COMPARE: classification of morphological patterns using adaptive regional elements. IEEE Trans. Med. Imaging 26 (1), 93–105.

Gauthier, S., Reisberg, B., Zaudig, M., Petersen, R.C., Ritchie, K., Broich, K., Belleville, S., Brodaty, H., Bennett, D., Chertkow, H., et al., 2006. Mild cognitive impairment. The Lancet 367 (9518), 1262–1270.

Good, C.D., Johnsrude, I.S., Ashburner, J., Henson, R.N., Friston, K.J., Frackowiak, R.S., 2001. A voxel-based morphometric study of ageing in 465 normal adult human brains. Neuroimage 14 (1), 21–36.

Haufe, S., Meinecke, F., Görgen, K., Dähne, S., Haynes, J.-D., Blankertz, B., Bießmann, F., 2014. On the interpretation of weight vectors of linear models in multivariate neuroimaging. Neuroimage 87, 96–110.

He, W., Goodkind, D., Kowal, P.R., 2016. An Aging World: 2015. United States Census Bureau, Washington, DC.

Herrera, F., Ventura, S., Bello, R., Cornelis, C., Zafra, A., Sánchez-Tarragó, D., Vluymans, S., 2016. Multiple instance learning. In: Multiple Instance Learning. Springer, pp. 17–33.

Hirata, Y., Matsuda, H., Nemoto, K., Ohnishi, T., Hirao, K., Yamashita, F., Asada, T., Iwabuchi, S., Samejima, H., 2005. Voxel-based morphometry to discriminate early Azheimer's disease from controls. Neurosci. Lett. 382 (3), 269–274.

Ithapu, V.K., Singh, V., Okonkwo, O.C., Chappell, R.J., Dowling, N.M., Johnson, S.C., Initiative, A.D.N., et al., 2015. Imaging-based enrichment criteria using deep learning algorithms for efficient clinical trials in mild cognitive impairment. Alzheimer's Dementia 11 (12), 1489–1499.

Jack, C.R., Bernstein, M.A., Fox, N.C., Thompson, P., Alexander, G., Harvey, D., Borowski, B., Britson, P.J., L Whitwell, J., Ward, C., et al., 2008. The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods. J. Magn. Reson. Imaging 27 (4), 685–691.

Kabani, N.J., 1998. 3D anatomical atlas of the human brain. Neuroimage 7, 0717.

Kantarci, K., Weigand, S., Przybelski, S., Shiung, M., Whitwell, J., Negash, S., Knopman, D., Boeve, B., O'Brien, P., Petersen, R., et al., 2009. Risk of dementia in MCI combined effect of cerebrovascular disease, volumetric MRI, and 1H. MRS. Neurol. 72 (17), 1519–1525.

Klöppel, S., Stonnington, C.M., Chu, C., Draganski, B., Scahill, R.I., Rohrer, J.D., Fox, N.C., Jack Jr., C.R., Ashburner, J., Frackowiak, R.S., 2008. Automatic classification of MR scans in Azheimer's disease. Brain 131 (3), 681–689.

Lebedev, A., Westman, E., Van Westen, G., Kramberger, M., Lundervold, A., Aarsland, D., Soininen, H., Kłoszewska, I., Mecocci, P., Tsolaki, M., et al., 2014. Random forest ensembles for detection and prediction of Azheimer's disease with a good between-cohort robustness. Neuroimage: Clinical 6, 115–125.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521 (7553), 436.

Lin, W., Tong, T., Gao, Q., Guo, D., Du, X., Yang, Y., Guo, G., Xiao, M., Du, M., Qu, X., et al., 2018. Convolutional neural networks-based MRI image analysis for the Alzheimer's disease prediction from mild cognitive impairment. Front. Neurosci. 12.

Liu, M., Zhang, D., Shen, D., 2012a. Ensemble sparse classification of Alzheimer's disease. Neuroimage 60 (2), 1106–1116.

Liu, M., Zhang, D., Shen, D., 2014a. Hierarchical fusion of features and classifier decisions for Alzheimer's disease diagnosis. Hum. Brain Mapp. 35 (4), 1305–1319.

Liu, M., Zhang, D., Shen, D., Initiative, A.D.N., et al., 2012b. Ensemble sparse classification of Azheimer's disease. Neuroimage 60 (2), 1106–1116.

Liu, M., Zhang, J., Adeli, E., Shen, D., 2018. Landmark-based deep multi-instance learning for brain disease diagnosis. Med. Image Anal. 43, 157–168.

Liu, S., Liu, S., Cai, W., Che, H., Pujol, S., Kikinis, R., Feng, D., Fulham, M.J., et al., 2015. Multimodal neuroimaging feature learning for multiclass diagnosis of Azheimer's disease. IEEE (Inst. Electr. Electron. Eng.) Trans. Biomed. Eng. 62 (4), 1132–1140.

Liu, S., Liu, S., Cai, W., Pujol, S., Kikinis, R., Feng, D., 2014b. Early diagnosis of Azheimer's disease with deep learning. In: 2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI). IEEE, pp. 1015–1018.

McKhann, G.M., Knopman, D.S., Chertkow, H., Hyman, B.T., Jack Jr., C.R., Kawas, C.H., Klunk, W.E., Koroshetz, W.J., Manly, J.J., Mayeux, R., et al., 2011. The diagnosis of dementia due to Azheimer's disease: recommendations from the National Institute on

Aging - Azheimer's Association workgroups on diagnostic guidelines for Azheimer's disease. Alzheimer's Dementia 7 (3), 263–269.

Mitchell, A.J., Shiri-Feshki, M., 2009. Rate of progression of mild cognitive impairment to dementia – meta-analysis of 41 robust inception cohort studies. Acta Psychiatr. Scand. 119 (4), 252–265.

Moradi, E., Pepe, A., Gaser, C., Huttunen, H., Tohka, J., Initiative, A.D.N., et al., 2015. Machine learning framework for early MRI-based Azheimer's conversion prediction in MCI subjects. Neuroimage 104, 398–412.

Rathore, S., Habes, M., Iftikhar, M.A., Shacklett, A., Davatzikos, C., 2017. A review on neuroimaging-based classification studies and associated feature extraction methods for Azheimer's disease and its prodromal stages. Neuroimage 155, 530–548.

Reiman, E.M., Quiroz, Y.T., Fleisher, A.S., Chen, K., Velez-Pardo, C., Jimenez-Del-Rio, M., Fagan, A.M., Shah, A.R., Alvarez, S., Arbelaez, A., et al., 2012. Brain imaging and fluid biomarker analysis in young adults at genetic risk for autosomal dominant Azheimer's disease in the presenilin 1 e280a kindred: a case-control study. Lancet Neurol. 11 (12), 1048–1056.

Shen, D., Davatzikos, C., 2002. Hammer: hierarchical attribute matching mechanism for elastic registration. IEEE Trans. Med. Imaging 21 (11), 1421–1439.

Shen, D., Wee, C.-Y., Zhang, D., Zhou, L., Yap, P.-T., 2014. Machine learning techniques for AD/MCI diagnosis and prognosis. In: Machine Learning in Healthcare Informatics. Springer, pp. 147–179.

Shen, D., Wu, G., Suk, H.-I., 2017. Deep learning in medical image analysis. Annu. Rev. Biomed. Eng. 19, 221–248.

Sled, J.G., Zijdenbos, A.P., Evans, A.C., 1998. A nonparametric method for automatic correction of intensity nonuniformity in MRI data. IEEE Trans. Med. Imaging 17 (1), 87–97.

Sperling, R.A., Aisen, P.S., Beckett, L.A., Bennett, D.A., Craft, S., Fagan, A.M., Iwatsubo, T., Jack Jr., C.R., Kaye, J., Montine, T.J., et al., 2011. Toward defining the preclinical stages of Azheimer's disease: recommendations from the National Institute on the Aging - Alzheimer's Association workgroups on diagnostic guidelines for Azheimer's disease. Alzheimer's Dementia 7 (3), 280–292.

Suk, H.-I., Lee, S.-W., Shen, D., Initiative, A.D.N., et al., 2014. Hierarchical feature representation and multimodal fusion with deep learning for AD/MCI diagnosis. Neuroimage 101, 569–582.

Suk, H.-I., Lee, S.-W., Shen, D., Initiative, A.D.N., et al., 2015. Latent feature representation with stacked auto-encoder for AD/MCI diagnosis. Brain Struct. Funct. 220 (2), 841–859.

Suk, H.-I., Lee, S.-W., Shen, D., Initiative, A.D.N., et al., 2017. Deep ensemble learning of sparse regression models for brain disease diagnosis. Med. Image Anal. 37, 101–113.

Tong, T., Wolz, R., Gao, Q., Guerrero, R., Hajnal, J.V., Rueckert, D., 2014. Multiple instance learning for classification of dementia in brain MRI. Med. Image Anal. 18 (5), 808–818.

Van Essen, D.C., Smith, S.M., Barch, D.M., Behrens, T.E., Yacoub, E., Ugurbil, K., Consortium, W.-M.H., et al., 2013. The Wu-Minn human connectome project: an overview. Neuroimage 80, 62–79.

Villemagne, V.L., Burnham, S., Bourgeat, P., Brown, B., Ellis, K.A., Salvado, O., Szoeke, C., Macaulay, S.L., Martins, R., Maruff, P., et al., 2013. Amyloid $\beta$ deposition, neurodegeneration, and cognitive decline in sporadic Azheimer's disease: a prospective cohort study. Lancet Neurol. 12 (4), 357–367.

Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., Manzagol, P.-A., 2010. Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion. J. Mach. Learn. Res. 11 (12), 3371–3408.

Wang, Y., Nie, J., Yap, P.-T., Shi, F., Guo, L., Shen, D., 2011. Robust deformable-surface-based skull-stripping for large-scale studies. In: Medical Image Computing and Computer-Assisted Intervention (MICCAI). Vol. 6893 of Lecture Notes in Computer Science, pp. 635–642.

Ward, A., Tardiff, S., Dye, C., Arrighi, H.M., 2013. Rate of conversion from prodromal Azheimer's disease to Alzheimer's dementia: a systematic review of the literature. Dement. Geriatr. Cognit. Disord. Extra 3 (1), 320–332.

Zarow, C., Vinters, H.V., Ellis, W.G., Weiner, M.W., Mungas, D., White, L., Chui, H.C., 2005. Correlates of hippocampal neuron number in Azheimer's disease and ischemic vascular dementia. Ann. Neurol.: Off. J. Am. Neurol. Assoc. Child Neurol. Soc. 57 (6), 896–903.

Zhang, D., Wang, Y., Zhou, L., Yuan, H., Shen, D., Initiative, A.D.N., et al., 2011. Multimodal classification of Alzheimer's disease and mild cognitive impairment. Neuroimage 55 (3), 856–867.

Zhang, Y., Brady, M., Smith, S., 2001. Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. IEEE Trans. Med. Imaging 20 (1), 45–57.