Hierarchical feature representation and multimodal fusion with deep learning for AD/MCI diagnosis

Heung-Il Suk [a], Seong-Whan Lee [b], Dinggang Shen [a,b,*], the Alzheimer's Disease Neuroimaging Initiative [1]

[a] Department of Radiology and Biomedical Research Imaging Center (BRIC), University of North Carolina at Chapel Hill, NC, USA
[b] Department of Brain and Cognitive Engineering, Korea University, Seoul, Republic of Korea

ARTICLE INFO

ABSTRACT

For the last decade, it has been shown that neuroimaging can be a potential tool for the diagnosis of Alzheimer's Disease (AD) and its prodromal stage, Mild Cognitive Impairment (MCI), and also fusion of different modalities can further provide the complementary information to enhance diagnostic accuracy. Here, we focus on the problems of both feature representation and fusion of multimodal information from Magnetic Resonance Imaging (MRI) and Positron Emission Tomography (PET). To our best knowledge, the previous methods in the literature mostly used hand-crafted features such as cortical thickness, gray matter densities from MRI, or voxel intensities from PET, and then combined these multimodal features by simply concatenating into a long vector or transforming into a higher-dimensional kernel space. In this paper, we propose a novel method for a high-level latent and shared feature representation from neuroimaging modalities via deep learning. Specifically, we use Deep Boltzmann Machine (DBM)[2], a deep network with a restricted Boltzmann machine as a building block, to find a latent hierarchical feature representation from a 3D patch, and then devise a systematic method for a joint feature representation from the paired patches of MRI and PET with a multimodal DBM. To validate the effectiveness of the proposed method, we performed experiments on ADNI dataset and compared with the state-of-the-art methods. In three binary classification problems of AD vs. healthy Normal Control (NC), MCI vs. NC, and MCI converter vs. MCI non-converter, we obtained the maximal accuracies of 95.35%, 85.67%, and 74.58%, respectively, outperforming the competing methods. By visual inspection of the trained model, we observed that the proposed method could hierarchically discover the complex latent patterns inherent in both MRI and PET.

© 2014 Elsevier Inc. All rights reserved.

Introduction

Alzheimer's Disease (AD), characterized by progressive impairment of cognitive and memory functions, is the most prevalent cause of dementia in elderly subjects. According to a recent report by Alzheimer's Association, the number of subjects with AD is significantly increasing every year, and 10 to 20% of people aged 65 or older have Mild Cognitive Impairment (MCI), known as a prodromal stage of AD (Alzheimer's Association, 2012). However, due to the limited period for which the symptomatic treatments could be effective, it has been of great importance for early diagnosis and prognosis of AD/MCI in the clinic.

To this end, many researchers have devoted their efforts to find biomarkers and develop a computer-aided system, with which we can effectively predict or diagnose the diseases. Recent studies have shown that the neuroimaging such as Magnetic Resonance Imaging (MRI) (Cuingnet et al., 2011; Davatzikos et al., 2011; Li et al., 2012; Wee et al., 2011; Zhang et al., 2012; Zhou et al., 2011), Positron Emission Tomography (PET) (Nordberg et al., 2010), and functional MRI (fMRI) (Greicius et al., 2004; Suk et al., 2013), can be nice tools for diagnosis or prognosis of AD/MCI. Furthermore, fusing the complementary information from multiple modalities helps enhance the diagnostic accuracy (Cui et al., 2011; Fan et al., 2007a; Hinrichs et al., 2011; Kohannim et al., 2010; Perrin et al., 2009; Suk and Shen, 2013; Walhovd et al., 2010; Wee et al., 2012; Westman et al., 2012; Yuan et al., 2012; Zhang and Shen, 2012; Zhang et al., 2011).

Various types of features or patterns extracted from neuroimaging modalities have been considered for brain disease diagnosis with machine learning methods. Here, we divide the previous feature extraction approaches into three categories: voxel-based approach, Region Of Interest (ROI)-based approach, and patch-based approach. A voxel-based approach is the most simple and direct way that uses the voxel intensities as features in classification (Baron et al., 2001; Ishii et al., 2005). Although it is simple and intuitive in terms of interpretation of

the results, its main limitations are the high-dimensionality of feature vectors and also the ignorance of regional information. ROI-based approach considers the structurally or functionally predefined brain regions and extracts representative features from each region (Cuingnet et al., 2011; Davatzikos et al., 2011; Kohannim et al., 2010; Nordberg et al., 2010; Suk and Shen, 2013; Walhovd et al., 2010; Zhang and Shen, 2012). Thanks to the relatively low feature dimensionality and the whole brain coverage, it is widely used in the literature. However, the features extracted from ROIs are very coarse in the sense that they cannot reflect small or subtle changes involved in the brain diseases. Note that the disease-related structural/functional changes occur in multiple brain regions. Furthermore, since the abnormal regions affected by neurodegenerative diseases can be part of ROIs or span over multiple ROIs, the simple voxel- or ROI-based approach may not effectively capture the diseased-related pathologies. To tackle these limitations, recently, Liu et al. proposed a patch-based method that first dissected brain areas into small 3D patches, extracted features from each selected patch individually, and then combined the features hierarchically in a classifier level (Liu et al., 2012, 2013).

As for the fusion of multiple modalities including MRI, PET, biological and neurological data for discriminating AD/MCI patients from healthy Normal Control (NC), Kohannim et al. (2010) concatenated features from modalities into a vector and used a Support Vector Machine (SVM) classifier. Walhovd et al. (2010) applied multi-method stepwise logistic regression analyses, and Westman et al. (2012) exploited a hierarchical modeling of orthogonal partial least squares to latent structures. Hinrichs et al. (2011), Suk and Shen (2013), and Zhang et al. (2011), independently, utilized a kernel-based machine learning technique.

In this paper, we consider the problems of both feature representation and multimodal data fusion for computer-aided AD/MCI diagnosis. Specifically, for feature representation, we exploit a patch-based approach since it can be considered as an intermediate level between voxel-based approach and ROI-based approach, thus efficiently handling the concerns of the high feature dimension and also the sensitivity to small change. Furthermore, from a clinical perspective, neurologists or radiologists examine brain images by searching local distinctive regions and then combine the interpretations with neighboring ones and ultimately with the whole brain. In these regards, we believe that the patch-based approach can effectively handle the region-wide pathologies, which may not be limited to specific ROIs, and accords with the neurologists or radiologists' perspective in terms of examining images, i.e., investigating local patterns and then combining local information distributed in the whole brain for making a clinical decision. In this way, we can also extract richer information that helps enhance diagnostic accuracy.

However, unlike Liu et al.'s method that directly used the gray matter density values in each patch as features, we propose to use a latent high-level feature representation. Meanwhile, in the fusion of multimodal information, the previous methods often applied either simple concatenation of features extracted from multiple modalities or kernel methods to combine them in a high-dimensional kernel space. However, the feature extraction and feature combination were often performed independently. In this work, we propose a novel method of extracting a shared feature representation from multiple modalities, i.e., MRI and PET. As investigated in the previous studies (Catana et al., 2012; Pichler et al., 2010), there exist the inherent relations between modalities of MRI and PET. Thus, finding the shared feature representation, which combines the complementary information from modalities, is helpful to enhance performance on AD/MCI diagnosis.

From a feature representation perspective, it is noteworthy that unlike the previous approaches (Hinrichs et al., 2011; Kohannim et al., 2010; Liu et al., 2012, 2013; Walhovd et al., 2010; Westman et al., 2012; Zhang and Shen, 2012; Zhang et al., 2011) that considered simple low-level features, which are often vulnerable to noises, we propose to consider high-level or abstract features for improving the robustness to

noises. For obtaining the latent high-level feature representations inherent in a patch observation such as correlations among voxels that cover different brain regions, we exploit a deep learning strategy (Bengio, 2009; LeCun et al., 1998), which has been successfully applied to medical imaging analysis (Ciresan et al., 2013; Hjelm et al., 2014; Liao et al., 2013; Shin et al., 2013; Suk and Shen, 2013). Among various deep models, we use a Deep Boltzmann Machine (DBM) (Salakhutdinov and Hinton, 2009) that can hierarchically find feature representations in a probabilistic manner. Rather than using the noisy voxel intensities as features as Liu et al. (2013) did, the high-level representation obtained via DBM is more robust to noises and thus helps enhance diagnostic performances. Meanwhile, from a multimodal data fusion perspective, unlike the conventional multimodal feature combination methods that first extract modality-specific features and then fuse their complementary information during classifier learning, the proposed multimodal DBM fuses the complementary information from different modalities during a feature representation step. Note that once we extract features from each modality, we may already lose some good correlation information between modalities. Therefore, it is important to discover a shared representation by fully utilizing the original information in each modality during feature representation procedure. In our multimodal data fusion method, thanks to the methodological characteristic of the DBM (i.e., undirected graphical model), it allows the bidirectional information flow from one modality (e.g., MRI) to the other modality (e.g., PET) and vice versa. Therefore, we can distribute feature representations over different layers in the path between modalities and thus efficiently discover a shared representation while still utilizing the full information in the observations.

Materials and image processing

Subjects

In this work, we use the ADNI dataset publicly available on the web,[3] but consider only the baseline MRI and 18-Fluoro-DeoxyGlucose PET (FDG-PET) data acquired from 93 AD subjects, 204 MCI subjects including 76 MCI converters (MCI-C) and 128 MCI non-converters (MCI-NC), and 101 NC subjects.[4] The demographics of the subjects are detailed in Table 1.

With regard to the general eligibility criteria in ADNI, subjects were in the age of between 55 and 90 with a study partner, who could provide an independent evaluation of functioning. General inclusion/exclusion criteria[5] are as follows: 1) NC subjects: MMSE scores between 24 and 30 (inclusive), a Clinical Dementia Rating (CDR) of 0, non-depressed, non-MCI, and non-demented; 2) MCI subjects: MMSE scores between 24 and 30 (inclusive), a memory complaint, objective memory loss measured by education adjusted scores on Wechsler Memory Scale Logical Memory II, a CDR of 0.5, absence of significant levels of impairment in other cognitive domains, essentially preserved activities of daily living, and an absence of dementia; and 3) mild AD: MMSE scores between 20 and 26 (inclusive), CDR of 0.5 or 1.0, and meets the National Institute of Neurological and Communicative Disorders and Stroke and the Alzheimer's Disease and Related Disorders Association (NINCDS/ADRDA) criteria for probable AD.

MRI/PET scanning and image processing

The structural MR images were acquired from 1.5 T scanners. We downloaded data in the Neuroimaging Informatics Technology Initiative (NIfTI) format, which had been pre-processed for spatial distortion correction caused by gradient nonlinearity and B1 field inhomogeneity.

---

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

[4] Although there exist in total more than 800 subjects in ADNI database, only 398 subjects have the baseline data including the modalities of both MRI and FDG-PET.

[5] Refer to 'http://www.adniinfo.org' for the details.

Table 1
Demographic and clinical information of the subjects. (SD: standard deviation).

	AD (93)	MCI (204)	NC (101)
Female/male	36/57	68/136	39/62
Age (mean ± SD) [min–max]	75.49 ± 7.4 [55–88]	74.97 ± 7.2 [55–89]	75.93 ± 4.8 [62–87]
Education (mean ± SD) [min–max]	14.66 ± 3.2 [4–20]	15.75 ± 2.9 [7–20]	15.83 ± 3.2 [7–20]
MMSE (mean ± SD) [min–max]	23.45 ± 2.1 [18–27]	27.18 ± 1.7 [24–30]	28.93 ± 1.1 [25–30]
CDR (mean ± SD) [min–max]	0.8 ± 0.25 [0.5–1]	0.5 ± 0.03 [0–0.5]	0 ± 0 [0–0]

The FDG-PET images were acquired 30–60 min post-injection, averaged, spatially aligned, interpolated to a standard voxel size, normalized in intensity, and smoothed to a common resolution of 8 mm full width at half maximum.

The MR images were preprocessed by applying the typical procedures of Anterior Commissure (AC)–Posterior Commissure (PC) correction, skull-stripping, and cerebellum removal. Specifically, we used MIPAV software[6] for AC–PC correction, resampled images to $256 \times 256 \times 256$, and applied N3 algorithm (Sled et al., 1998) to correct non-uniform tissue intensities. After skull stripping (Wang et al., 2014) and cerebellum removal, we manually checked the skull-stripped images to ensure the clean and dura removal. Then, FAST in FSL package[7] (Zhang et al., 2001) was used to segment the structural MR images into three tissue types of Gray Matter (GM), White Matter (WM), and CerebroSpinal Fluid (CSF). Finally, all the three tissues of MR image were spatially normalized onto a standard space, for which in this work we used a brain atlas already aligned with the MNI coordinate space (Kabani et al., 1998), via HAMMER (Shen and Davatzikos, 2002), although other advanced registration methods can also be applied for this process (Friston, 1995; Jia et al., 2010; Tang et al., 2009; Xue et al., 2006; Yang et al., 2008). Then, the regional volumetric maps, called RAVENS maps, were generated by a tissue preserving image warping method (Davatzikos et al., 2001). It is noteworthy that the values of RAVENS maps are proportional to the amount of original tissue volume for each region, giving a quantitative representation of the spatial distribution of tissue types. Due to its relatively high relatedness to AD/MCI compared to WM and CSF (Liu et al., 2012), in this work, we considered only the spatially normalized GM volumes, called GM tissue densities, for classification. Regarding FDG-PET images, they were rigidly aligned to the respective MR images. The GM density maps and the PET images were further smoothed using a Gaussian kernel (with unit standard deviation) to improve the signal-to-noise ratio. We downsampled both the GM density maps and PET images to $64 \times 64 \times 64$ voxels[8] according to Liu et al.'s (2013) work, which saved the computational time and memory cost, but without sacrificing the classification accuracy.

Method

In Fig. 1, we illustrate a schematic diagram of our framework for AD/MCI diagnosis. Given a pair of MRI and PET images, we first select class-discriminative patches by means of a statistical significance test between classes. Using the tissue densities of a MRI patch and the voxel intensities of a PET patch as observations, we build a patch-level feature learning model, called a MultiModal DBM (MM-DBM), that finds a shared feature representation from the paired patches. Here, instead of using the original real-valued tissue densities of MRI and the voxel intensities of PET as inputs to MM-DBM, we first train a Gaussian Restricted Boltzmann Machine (RBM) and use it as a preprocessor to transform the real-valued observations into binary vectors, which become the input to MM-DBM. After finding latent and shared feature representations of the paired patches from the trained MM-DBM, we construct an image-level classifier by fusing multiple classifiers in a

hierarchical manner, i.e., patch-level classifier learning, mega-patch construction, and a final ensemble classification.

Patch extraction

For the class-discriminative patch extraction, we exploit statistical significance for voxels in each patch, i.e., p-values, following Liu et al.'s (2013) work. It is noteworthy that in this step, we take advantage of a group-wise analysis via voxel-wise statistical test. That is, by first performing group comparison, e.g., AD and NC, we can find the statistically significant voxels, which can provide useful information for brain disease diagnosis. Based on these voxels, we can then define the class-discriminative patches to further utilize local regional information. By considering only the selected discriminative patches rather than all patches in an image, we can obtain both performance improvement in classification and reduction in computational cost. Throughout this paper, a patch is defined as a three-dimensional cube with a size of $w \times w \times w$ in a brain image, i.e., MRI or PET. Given a set of training images, we first perform two-sample t-test on each voxel, and then select voxels with the p-value smaller than the predefined threshold.[9] For each of the selected voxels, by taking each of them as a center, we extract patches with a size of $w \times w \times w$, and then compute a mean p-value by averaging the p-values of all voxels within a patch. Finally, by scanning all the extracted patches, we select class-discriminative patches in a greedy manner with the following rules:

- The candidate patch should be overlapped less than 50% with any of the selected patches.
- Among the candidate patches that satisfy the rule above, we select patches whose mean p-values are smaller than the average p-value of all candidate patches.

For the multimodal case, i.e., MRI and PET in our work, we apply the steps of testing the statistical significance, extracting patches, and computing the mean p-values as explained above, for each modality independently. But for the last step of selecting class-discriminative patches, we consider multiple modalities together. That is, regarding the second rule, the mean p-value of a candidate patch should be smaller than that of all candidate patches of all the modalities. Once a patch location is determined from one modality, a patch of the same location in the other modality is paired for multimodal joint feature representation, which is described in the following section.

### Patch-level deep feature learning

Recently, Liu et al. (2013) presented a hierarchical framework that gradually integrated features from a number of local patches extracted from a GM density map. Although they showed the efficacy of their method for AD/MCI diagnosis, it is well-known that the structural or functional images are susceptible to acquisition noise, intensity inhomogeneity, artifacts, etc. Furthermore, the raw voxel density or intensity values in a patch can be considered as low-level features that do not efficiently capture more informative high-level features. To this end, in this paper, we propose a deep learning based high-

---

[6] Available at 'http://mipav.cit.nih.gov/clickwrap.php'.
[7] Available at 'http://fsl.fmrib.ox.ac.uk/fsl/fslwiki/'.
[8] The final voxel size is $4 \times 4 \times 4$ mm^3.

[9] In this work, we set the threshold to 0.05.
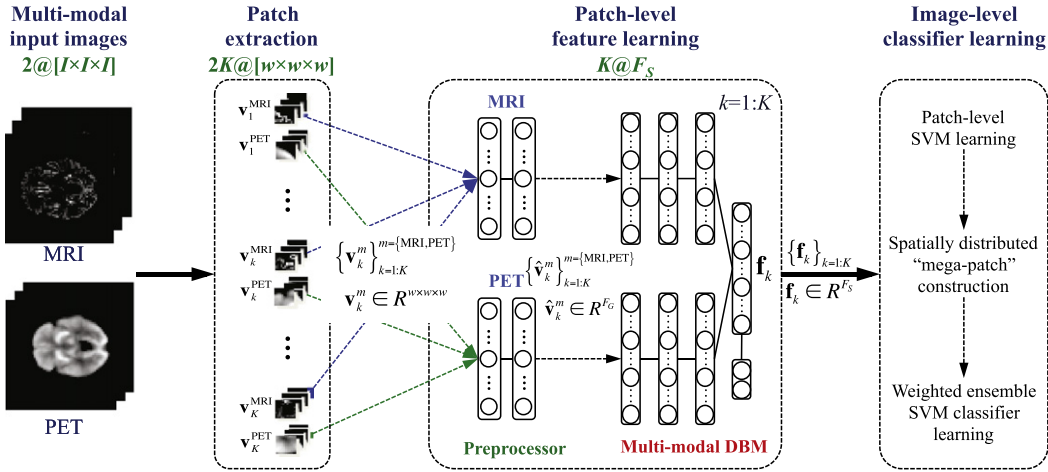
Fig. 1. Schematic illustration of the proposed method in hierarchical feature representation and multimodal fusion with deep learning for AD/MCI diagnosis. (I: image size, w: patch size, K: # of the selected patches, m: modality index, F_G: # of hidden units in a Gaussian restricted Boltzmann machine, i.e., preprocessor, F_S: # of hidden units in the top layer of a multimodal Deep Boltzmann Machine (DBM)).

level structural and functional feature representation from MRI and PET, respectively, for AD/MCI classification.
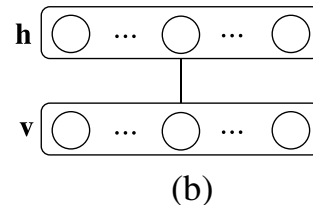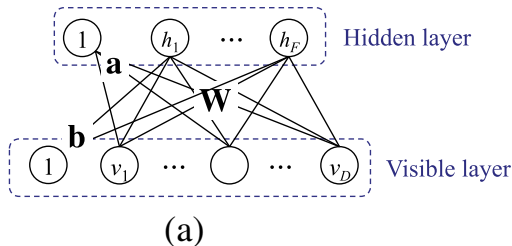
In the following, we first introduce an RBM, which has recently become a prominent tool for feature learning with applications in a wide variety of machine learning fields. Then, we describe a DBM, a network of stacking multiple RBMs, with which we discover a latent hierarchical feature representation from a patch. We finally explain a systemic method to find a joint feature representation from multimodal neuroimaging data, such as MRI and PET.

Restricted Boltzmann Machine

An RBM is a two-layer undirected graphical model with visible and hidden units or variables in each layer (Fig. 2). Hereafter, we use units and variables interchangeably. It assumes a symmetric connectivity **W** between the visible layer and the hidden layer, but no connections within the layers, and each layer has a bias term, **a** and **b**, respectively. In Fig. 2, the units of the visible layer $\mathbf{v} = [v_i]$, $i = \{1, \cdots, D\}$, correspond to the observations while the units of the hidden layer $\mathbf{h} = [h_j]$, $j = \{1, \cdots, F\}$, models the structures or dependencies over visible variables, where D and F denote, respectively, the numbers of visible and hidden units. In our work, the voxel intensities of a patch become the values of the visible units, and the hidden units represent the complex relations of the input units, i.e., voxels in a patch, that can be captured by the symmetric matrix **W**. It is worth noting that because of the symmetricity of the matrix **W**, we can also reconstruct the input observations, i.e., a patch, from the hidden representations. Therefore, an RBM is also considered as an auto-encoder (Hinton and Salakhutdinov, 2006). This favorable characteristic is also used in RBM parameter learning (Hinton et al., 2006).

In RBM, a joint probability of $(\mathbf{v}, \mathbf{h})$ is given by:

$$P(\mathbf{v}, \mathbf{h}; \Theta) = \frac{1}{\mathbf{Z}(\Theta)} exp[-E(\mathbf{v}, \mathbf{h}; \Theta)] \quad (1)$$

where $\Theta = \{\mathbf{W} = [W_{ij}] \in R^{D \times F}, \mathbf{a} = [a_i] \in R^D, \mathbf{b} = [b_j] \in R^F\}$, $E(\mathbf{v}, \mathbf{h}; \Theta)$ is an energy function, and $Z(\Theta)$ is a partition function that can be obtained by summing over all possible pairs of $\mathbf{v}$ and $\mathbf{h}$. For the sake of simplicity, by assuming binary visible and hidden units, the energy function $E(\mathbf{v}, \mathbf{h}; \Theta)$ is defined by

$$E(\mathbf{v}, \mathbf{h}; \Theta) = -\mathbf{h}^\top \mathbf{W} \mathbf{v} - \mathbf{a}^\top \mathbf{v} - \mathbf{b}^\top \mathbf{h}$$
$$= -\sum_{i=1}^{D} \sum_{j=1}^{D} v_i W_{ij} h_j - \sum_{i=1}^{D} a_i v_i - \sum_{j=1}^{F} b_j h_j. \quad (2)$$

The conditional distribution of the hidden variables given the visible variables and also the conditional distribution of the visible variables given the hidden variables are, respectively, computed as follows:

$$P(h_j = 1 | \mathbf{v}; \Theta) = sigm\left(b_j + \sum_{i=1}^{D} W_{ij} v_i\right) \quad (3)$$

$$P(v_i = 1 | \mathbf{h}; \Theta) = sigm\left(a_i + \sum_{j=1}^{F} W_{ij} h_j\right) \quad (4)$$

where $sigm(\cdot) = \frac{exp[\cdot]}{1 + exp[\cdot]}$ is a logistic sigmoid function. Due to the unobservable hidden variables, the objective function is defined as the marginal distribution of the visible variables as follows:

$$P(\mathbf{v}; \Theta) = \frac{1}{Z(\Theta)} \sum_{\mathbf{h}} exp(-E(\mathbf{v}, \mathbf{h}; \Theta)). \quad (5)$$

In our work, the observed patch values from MRI and PET are real-valued $\mathbf{v} \in R^D$. For this case, it is common to use a Gaussian RBM

Fig. 2. An architecture of a restricted Boltzmann machine (a) and its simplified representation (b).

(Hinton and Salakhutdinov, 2006), in which the energy function is given by

$$E(\mathbf{v}, \mathbf{h}; \Theta) = \sum_{i=1}^{D} \frac{(v_i - a_i)^2}{2\sigma_i^2} - \sum_{i=1}^{D} \sum_{j=1}^{F} \frac{v_i}{\sigma_i} W_{ij} h_j - \sum_{j=1}^{F} b_j h_j \qquad (6)$$

where σ_i denotes a standard deviation of the i-th visible variable and $\Theta = \{\mathbf{W}, \mathbf{a}, \mathbf{b}, \boldsymbol{\sigma} = [\sigma_i] \in R^D\}$. This variation leads to the following conditional distribution of visible variables given the binary hidden variables

$$p(v_i | \mathbf{h}; \Theta) = \frac{1}{\sqrt{2\pi}\sigma_{\mathbf{i}}} exp\left(-\frac{\left(v_i - a_i - \sum_{j=1}^{F} h_j W_{ij}\right)^2}{2\sigma_i^2}\right). \qquad (7)$$

Deep Boltzmann Machine

A DBM is an undirected graphical model, structured by stacking multiple RBMs in a hierarchical manner. That is, a DBM contains a visible layer $\mathbf{v}$ and a series of hidden layers $\mathbf{h}^1 \in \{0, 1\}^{F_1}, \cdots, \mathbf{h}^i \in \{0, 1\}^{F_i}, \cdots, \mathbf{h}^L \in \{0, 1\}^{F_L}$, where F_i denotes the number of units in the i-th hidden layer and L is the number of hidden layers. We should note that, hereafter, for simplicity, we omit bias terms and assume that the visible and hidden variables are binary[10] or probability, and the following description on DBM is based on Salakhutdinov and Hinton's (2012) work.

Thanks to the hierarchical nature in the deep network, one of the most important characteristics of the DBM is to capture highly non-linear and complicated patterns or statistics such as the relations among input values. Another important feature of the DBM is that the hierarchical latent feature representation can be learned directly from the data without human intervention. In other words, unlike the previous methods that mostly considered hand-crafted/predefined features (Fan et al., 2007b; Liu et al., 2013; Zhang et al., 2011) or outputs from the predefined functions (Dinov et al., 2005; Hackmack et al., 2012), we assign the role of determining feature representations to a DBM and find them autonomously from the training samples. Utilizing its representational and self-taught learning properties, we can find a latent representation of the original GM tissue intensities and/or PET voxel intensities in a patch. When an input patch is presented to a DBM, the different layers of the network represent different levels of information. That is, the lower the layer in the network, the simpler patterns (e.g., linear relations of input variables); the higher the layer, the more complicated or abstract patterns inherent in the input values (e.g., non-linear relations among input variables).

The rationale of using DBM for feature representation is as follows: (1) it can learn internal latent representations that capture non-linear complicated patterns and/or statistical structures in a hierarchical manner (Bengio, 2009; Bengio et al., 2007; Hinton et al., 2006; Mohamed et al., 2012). However, unlike many other deep network models such as deep belief network (Hinton and Salakhutdinov, 2006), and stacked auto-encoder (Shin et al., 2013), the approximate inference procedure after the initial bottom-up pass incorporates top-down feedback, which allows DBM to use higher-level knowledge to resolve uncertainty about intermediate-level features, thus creating better data-dependent representations and statistics for learning (Salakhutdinov and Hinton, 2012). Thanks to this two-way dependencies, i.e., bottom-up and top-down, it was shown that DBMs outperform the other deep learning methods in computer vision (Montavon et al., 2012; Salakhutdinov and Hinton, 2009; Srivastava and Salakhutdinov, 2012). To this end, we use a DBM to discover hierarchical feature representation from neuroimaging, e.g., MRI and PET, in our work.

Fig. 3(a) shows an example of the three-layer DBM. The energy of the state $(\mathbf{v}, \mathbf{h}^1, \mathbf{h}^2)$ in the DBM is given by

$$E\left(\mathbf{v}, \mathbf{h}^1, \mathbf{h}^2; \Theta\right) = -\mathbf{v}^{\top}\mathbf{W}^1\mathbf{h}^1 - \left(\mathbf{h}^1\right)^{\top}\mathbf{W}^2\mathbf{h}^2 \qquad (8)$$

where $\mathbf{W}^1 = \left[W_{ij}^1\right] \in R^{D \times F_1}$ and $\mathbf{W}^2 = \left[W_{jk}^2\right] \in R^{F_1 \times F_2}$ are, respectively, symmetric connections of $(\mathbf{v}, \mathbf{h}^1)$ and $(\mathbf{h}^1, \mathbf{h}^2)$, and $\Theta = \{\mathbf{W}^1, \mathbf{W}^2\}$. Then the probability that the model assigns to a visible vector $\mathbf{v}$ is given by:

$$P(\mathbf{v}; \Theta) = \frac{1}{Z(\Theta)} \sum_{\mathbf{h}^1, \mathbf{h}^2} exp\left(-E\left(\mathbf{v}, \mathbf{h}^1, \mathbf{h}^2; \Theta\right)\right) \qquad (9)$$

where $Z(\Theta)$ is a normalizing factor. Given the values of the units in the neighboring layer(s), the probability of the binary visible or binary hidden units being set to 1 is computed as follows:

$$P\left(h_j^1 = 1 \middle| \mathbf{v}, \mathbf{h}^2\right) = sigm\left(\sum_{i=1}^{D} W_{ij}^1 v_i + \sum_{k=1}^{F_2} W_{jk}^2 h_k^2\right) \qquad (10)$$

$$P\left(h_k^2 = 1 \middle| \mathbf{h}^1\right) = sigm\left(\sum_{j=1}^{F_1} W_{jk}^2 h_j^1\right) \qquad (11)$$

$$P\left(v_i = 1 | \mathbf{h}^1\right) = sigm\left(\sum_{j=1}^{F_1} W_{ij}^1 h_j^1\right). \qquad (12)$$

Note that in the computation of the probability of the hidden units $\mathbf{h}^1$, we incorporate both the lower visible layer $\mathbf{v}$ and the higher hidden layer $\mathbf{h}^2$, and this makes DBM differentiated from other deep learning models and also more robust to noisy observations (Salakhutdinov and Hinton, 2009; Srivastava and Salakhutdinov, 2012).

Unlike the conventional generative DBM, in this work, we consider a discriminative DBM, by injecting a discriminative RBM (Larochelle and Bengio, 2008) at the top hidden layer. That is, the top hidden layer is connected to both the lower hidden layer and the additional label layer (Fig. 3(b)), which indicates the label of the input $\mathbf{v}$. In this way, we can train DBM to discover hierarchical and discriminative feature representations by integrating the process of discovering features of inputs with their use in classification (Larochelle and Bengio, 2008). Our model does not require an additional fine-tuning step for classification as done in Ngiam et al. (2011) and Salakhutdinov and Hinton (2009). With the inclusion of the additional label layer, the energy of the state $(\mathbf{v}, \mathbf{h}^1, \mathbf{h}^2, \mathbf{o})$ in the modified DBM is given by

$$E\left(\mathbf{v}, \mathbf{h}^1, \mathbf{h}^2, \mathbf{o}; \Theta\right) = -\mathbf{v}^{\top}\mathbf{W}^1\mathbf{h}^1 - \left(\mathbf{h}^1\right)^{\top}\mathbf{W}^2\mathbf{h}^2 - \left(\mathbf{h}^2\right)^{\top}\mathbf{U}\mathbf{o} \qquad (13)$$

where $\mathbf{U} = [U_{lk}] \in R^{C \times F_2}$ and $\mathbf{o} = [o_l] \in \{0, 1\}^C$ denote, respectively, a connectivity between the top hidden layer and the label layer and a class-label indicator vector, C is the number of classes, and $\Theta = \{\mathbf{W}_1, \mathbf{W}_2, \mathbf{U}\}$. The probability of an observation $(\mathbf{v}, \mathbf{o})$ is computed by

$$P(\mathbf{v}, \mathbf{o}; \Theta) = \frac{1}{Z(\Theta)} \sum_{\mathbf{h}^1, \mathbf{h}^2} exp\left(-E\left(\mathbf{v}, \mathbf{h}^1, \mathbf{h}^2, \mathbf{o}; \Theta\right)\right). \qquad (14)$$

The conditional probability of the top hidden units being set to 1 is given by

$$P\left(h_k^2 = 1 \middle| \mathbf{h}^1, \mathbf{o}\right) = sigm\left(\sum_{j=1}^{F_1} W_{jk}^2 h_j^1 + \sum_{l=1}^{C} U_{lk} o_l\right). \qquad (15)$$

---

[10] In our experiments, we trained a Gaussian RBM by fixing the standard deviations to 1 and transformed the observed real-values from MRI and PET to binary vectors using it as a preprocessor, following Nair and Hinton's (2008) work.
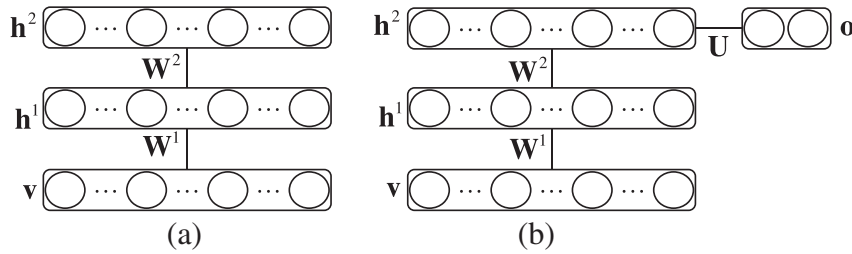
Fig. 3. An architecture of (a) a conventional Deep Boltzmann Machine and (b) its discriminative version with label information at the top layer.

For the label layer, we use a logistic function

$$P(o_l = 1|\mathbf{h}^2) = \frac{exp\left[\sum_{k=1}^{F_2} U_{lk} h_k^2\right]}{\sum_{l'=1}^{C} exp\left[\sum_{k=1}^{F_2} U_{l'k} h_k^2\right]}. \quad (16)$$

In this way, the hidden units capture class-predictive information about the input vector. Here, we should note that the label layer connected to the top hidden layer is considered during only the training phase of finding the class-discriminative parameters.

From a feature learning perspective, in the low layer of our model, basic image features such as spots and edges are captured from the input data. The learned low-level features are further fed into the high-level of the network, which encodes more abstract and higher level semantic information inherent in the input data. But, here, we should note that the output layer linked to the top hidden layer imposes the learned features to be discriminative between classes.

In order to learn the parameters $\Theta = \{\mathbf{W}^1, \mathbf{W}^2, \mathbf{U}\}$, we maximize the log-likelihood of the observed data $(\mathbf{v}, \mathbf{o})$. The derivative of the log-likelihood of the observed data with respect to the model parameters takes the simple form of

$$\frac{\partial logP(\mathbf{v}, \mathbf{o})}{\partial \mathbf{W}^i} = \left\langle \mathbf{h}^{i-1}\left(\mathbf{h}^i\right)^\top \right\rangle_{data} - \left\langle \mathbf{h}^{i-1}\left(\mathbf{h}^i\right)^\top \right\rangle_{model} \quad (17)$$

$$\frac{\partial logP(\mathbf{v}, \mathbf{o})}{\partial \mathbf{U}} = \left\langle \mathbf{h}^2 \mathbf{o}^\top \right\rangle_{data} - \left\langle \mathbf{h}^2 \mathbf{o}^\top \right\rangle_{model} \quad (18)$$

where $<\cdot>_{data}$ denotes the data-dependent statistics obtained by sampling the model conditioned on the visible units $\mathbf{v}(\equiv \mathbf{h}^0)$ and the label units $\mathbf{o}$ clamped to the observation and the corresponding label, respectively, and $<\cdot>_{model}$ denotes the data-independent statistics obtained by sampling from the model. When the model approximates the data distribution well, it can be reached for the equilibrium of data-dependent and data-independent statistics. In parameter learning, we use a gradient-based optimization strategy. In Eqs. (17) and (18), we need to compute the data-dependent and the data-independent statistics. First, because of the two-way dependency in DBM, it is not tractable for the data-dependent statistics. Fortunately, variational mean-field approximation works well for estimating the data-dependent statistics. For the details of computing the data-dependent statistics, please refer to Appendix A and Salakhutdinov and Hinton (2012).

Here, we should note that due to the large number of parameters involved in the DBM, it generally requires a huge number of training samples for generalization, which is not valid in practice, especially for the neuroimaging studies. However, Hinton et al. recently introduced a greedy layer-wise learning algorithm and successfully applied to learn a deep belief network (Hinton and Salakhutdinov, 2006). Since the pioneering work, many research groups have used this approach to initialize the parameters in deep learning and called it 'pre-training' (Bengio, 2009; Hinton et al., 2006). We apply the same procedure to provide a good initial configuration of the parameters, which helps the learning procedure converge much faster than random initialization.

The key idea in a greedy layer-wise learning is to train one layer at a time by maximizing the variational lower bound. That is, we first train the 1st hidden layer with the training data as input, and then train the 2nd hidden layer with the outputs from the 1st hidden layer as input, and so on. That is, the representation of the l-th hidden layer is used as input for the $(l + 1)$-th hidden layer and this pairwise model becomes an RBM. Here, it should be mentioned that, unlike the other deep networks, because the DBM integrates both bottom-up and top-down information, the first and last RBMs in the network need modification by using weights twice as big as in one direction. Since the detailed explanation on this issue is out of domain of our work, please refer to Salakhutdinov and Hinton (2012) for details.

In a nutshell, the learning proceeds by two steps: (1) a greedy-layer-wise pre-training for a good initial setup of the modal parameters, and (2) iterative alternation of variational mean-field approximation to estimate the posterior probabilities of hidden units and stochastic approximation to update model parameters (refer to Appendix A). After learning the parameters, we can then obtain a latent feature representation for an input sample, by inferring the probabilities of the hidden units in the trained DBM.[11]

Multimodal deep feature fusion

There are increasing evidences that biomarkers from different modalities can provide complementary information in AD/MCI diagnosis (Hinrichs et al., 2011; Kohannim et al., 2010; Perrin et al., 2009; Suk and Shen, 2013; Zhang et al., 2011). Unlike the previous methods that either simply concatenated features from multiple modalities into a long vector (Kohannim et al., 2010) or fused the modality-dependent features in a kernel space (Hinrichs et al., 2011; Suk and Shen, 2013; Zhang et al., 2011), in this work, we propose a systematic method of extracting multimodal feature representations in a probabilistic manner.

Different modalities will have different statistical properties, thus making it difficult to jointly model them using a shallow architecture. Therefore, simple concatenation of the features of multiple modalities can cause strong connections among the variables of individual modality, but few units across modalities (Ngiam et al., 2011). In order to tackle this problem, Srivastava and Salakhutdinov (2012) proposed a Multi-Modal DBM (MM-DBM) to combine images and texts for information retrieval. Motivated by their work, in this paper, we devise a modified MM-DBM, in which the top hidden layer has multiple entries of the lower hidden layers and the label layer, to extract a shared feature representation by fusing neuroimaging information of MRI and PET. Fig. 4 presents a multimodal deep network in which one path represents the statistical properties of MRI and the other path represents those of PET, and the top shared hidden layer finally discovers the shared properties of the modalities in a supervised manner. We argue that this joint feature representation discriminates our method from the previous multimodal

---

[11] Instead of the standard mean-field approximation, inspired by Montavon et al.'s (2012) work, in this work, we traverse the trained DBM in a feed-forward manner, i.e., $\mathbf{f} = sigm(\mathbf{W}^2 \cdot sigm(\mathbf{W}^1\mathbf{v}))$, for feature representations. The same strategy is applied for the multimodal DBM described in the Multimodal deep feature fusion section.
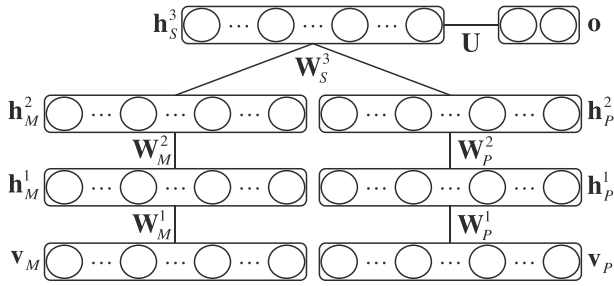
Fig. 4. An architecture of a multimodal Deep Boltzmann Machine for neuroimaging data fusion.

methods (Hinrichs et al., 2011; Suk and Shen, 2013; Zhang et al., 2011), which first extracted features from each modality independently, and then combined them through kernel machines.

The joint distribution over the multimodal inputs of MRI and PET can be estimated as follows:

$$P(\mathbf{v}_M, \mathbf{v}_P, \mathbf{o}; \Theta) = \sum_{\mathbf{h}_M^2, \mathbf{h}_P^2, \mathbf{h}_S} P\left(\mathbf{h}_M^2, \mathbf{h}_P^2, \mathbf{h}_S^3, \mathbf{o}\right)$$
$$\left(\sum_{\mathbf{h}_M^1} P\left(\mathbf{v}_M, \mathbf{h}_M^1, \mathbf{h}_M^2\right)\right)\left(\sum_{\mathbf{h}_P^1} P\left(\mathbf{v}_P, \mathbf{h}_P^1, \mathbf{h}_P^2\right)\right) \quad (19)$$

where the subscripts M, P, and S denote, respectively, units of the MRI path, the PET path, and the shared hidden layer. Regarding parameter learning for MM-DBM, the same strategy with the unimodal DBM learning can be applied. For details, please refer to Appendix B.

Image-level hierarchical classifier learning

In order to combine the distributed patch information over an image and build an image-level classifier, we use a hierarchical classifier learning scheme, proposed by Liu et al. (2013). That is, we first build a classifier for each patch, independently, and then combine them in a hierarchical manner by feeding the outputs from the lower-level classifiers to the upper-level classifier. Specifically, we build a three-level classifier for decision: patch-level, mega-patch-level, and image-level. For the patch-level classification, a linear Support Vector Machine (SVM) is trained for each patch location independently with the (MM-)DBM-learned feature representations as input. The output from a patch-level SVM, measured by the relative distance from the decision hyperplane, is then converted to a probability via a softmax function. Here, we should note that in patch-level classifier learning, we randomly partition the training data into a training set and a validation set.[12] The patch-level classifier is trained on the training set, and then the classification accuracy is obtained with the validation set.

In the following hierarchy, instead of considering all patch-level classifiers' output simultaneously, we agglomerate the information of the locally distributed patches by constructing spatially distributed 'mega-patches' under the consideration that the disease-related brain areas are distributed over some distant brain regions with arbitrary shape and size (Liu et al., 2013). Similar to the patch extraction described in Section 3.1, we construct mega-patches and the respective classifiers in a greedy manner. Concretely, we first sort the patches in a descending order based on the classification accuracy obtained with the validation set in patch-level classifier learning. Starting with the patch with the highest classification accuracy as a new mega-patch, we greedily merge the neighboring patches into the mega-patch. The merging condition is that, if and only if, a mega-patch classifier, which

is trained with the patches already included in the current mega-patch and also the candidate patch under consideration, produces a better classification accuracy. The process is repeated until all the patches are visited. The size of the constructed mega-patches and their component-patches are determined by a cross-validation. We also constrain that none of the final mega-patches overlap each other larger than the half of the respective mega-patches' size. Note that, in the step of the mega-patch classifier learning, some patches that are not informative in classification are discarded, and each mega-patch classifier covers different regions of the brain with a different size.

Finally, we build an image-level classifier by fusing the mega-patch classifiers. We select an optimal subset of mega-patches in a forward greedy search strategy for a final fusion. However, since the mega-patch selection is performed on the training data, the resulting image-level classifier may not be optimal for the testing data. To this end, we divide the training data into multiple subsets, and train an image-level classifier in each subset individually.[13] In this way, we can build multiple image-level classifiers, each of which selects possibly a different subset of mega-patches. By counting the selected frequency of mega-patches in each image-level classifier, we can finally compute the relative importance of the mega-patches. After normalizing the frequencies, we use them as weights of the respective mega-patches. The final decision in image-level classifiers is made by a weighted combination of the mega-patch classifiers' outputs.

Experimental results and discussions

In this section, we evaluate the effectiveness of the proposed method for (1) a latent feature representation with DBM and (2) a shared feature representation between MRI and PET with an MM-DBM, by considering three binary classification problems: AD vs. NC, MCI vs. NC, MCI converter (MCI-C) vs. MCI non-converter (MCI-NC). Due to the limited number of data, we applied a 10-fold cross validation technique. Specifically, we randomly partitioned the dataset into 10 subsets, each of which included 10% of the total data. We repeated experiments for each classification problem 10 times, by using 9 out of 10 subsets for training and the remaining one for testing at each time. It is worth noting that, for each classification problem, during a training phase, we performed patch selection, (MM-)DBM and SVM model learning only using the 9 training subsets. Based on the selected patches and also the trained (MM-)DBM and SVM models, we finally evaluated the performance on the left-out testing subset. We compare the proposed method with Liu et al.'s (2013) method, using the same training and testing set in each experiment for a fair comparison.

Experimental setup

As for the patch size w, we set it to 11 by following Liu et al.'s (2013) work. During mega-patch construction, the size of a mega-patch was allowed in the range of $w \times$ [1.2, 1.4, 1.6, 1.8, 2] and the optimal size for each mega-patch was determined by cross-validation as explained in the Image-level hierarchical classifier learning section.

In building (MM-)DBM of GM patches and/or PET patches, we can use Gaussian visible units for the input patches by considering the voxels as continuous variables. However, learning (MM-)DBMs with Gaussian visible units is very slow and requires a huge number of parameter updates, compared with the binary visible units. To this end, we first trained RBM with $1,331 (= 11^3)$ Gaussian visible units and also 500 binary hidden units by using contrastive divergence learning (Hinton et al., 2006) for 1000 epochs.[14] After training a Gaussian RBM for each modality, we used it as a preprocessor, following Nair and Hinton's (2008) work that effectively converts GM tissue densities or PET voxel intensities into

---

[12] In our work, we set 80% of the entire training data as a training set and the rest for a validation set.

[13] In this work, we set the number of subsets to 10.

[14] The input data were first normalized and whitened by zero component analysis, and the standard deviation was fixed to 1 during the parameter updates.

500-dimensional binary vectors. We then used the binary vectors as 'preprocessed data' to train our (MM-)DBMs. We should note that the Gaussian RBMs were not updated during (MM-)DBM learning.

We structured a three-layer DBM for MRI (MRI-DBM) and PET (PET-DBM), respectively, and a four-layer DBM for MRI + PET (MM-DBM). For all these models, we used binary visible and binary hidden units. Both the MRI-DBM and the PET-DBM were structured with 500(visible)–500(hidden)–500(hidden), and the MM-DBM was structured with 500(visible)–500(hidden)–500(hidden) for a MRI pathway, 500(visible)–500(hidden)–500(hidden) for a PET pathway, and finally 1000 hidden units for the shared hidden layer. In (MM-)DBM learning, we updated the parameters, i.e., weights and biases, with a learning rate of 10^{-3} and a momentum of 0.5 with an increment gradually up to 0.9 for 500 epochs. We used the trained parameters of MRI-DBM and PET-DBM as the initial setup of the MRI and PET pathways in MM-DBM learning. We implemented the DBM method based on Salakhutdinov's codes.[15]

We used a linear SVM for the hierarchical classifiers, i.e., patch-level classifier, mega-patch-level classifier, and image-level classifier. An LIBSVM toolbox[16] was used for SVM learning and classification. The free parameter that controls the soft margin was determined by a nested cross-validation.

Extracted patches and trained DBMs

In Fig. 5, we presented the example images overlaid with *p*-values of the voxels, obtained from AD and NC groups, based on which we selected patch locations for AD and NC classification. It is worth noting that, for both modalities, the voxels in the subcortical and medial temporal areas showed low *p*-values, i.e., statistically different between classes, while for other areas, each modality presents slightly different *p*-value distributions, from which we could possibly obtain complementary information for classification. Samples of the selected 3D patches are also presented in Fig. 6, in which one 3D volume is displayed in each row, for each modality. Taking these patches as input data to a Gaussian RBM and then transforming to binary vectors, we trained our feature representation models, i.e., MRI-DBM, PET-DBM, and MM-DBM. Regarding the trained MM-DBM, we visualized the trained weights in Fig. 7 by linearly projecting them to the input space for intuitive interpretation of the feature representations.[17] In the figure, the left images represent the trained weights of our Gaussian RBMs that were used to convert the real-valued patches into binary vectors as a preprocessor, and the right images represent the trained weights of the first-layer hidden units of the respective modality's pathway in our MM-DBM. From the figure, we can regard the hidden units in the Gaussian RBM as simple cells of a human visual cortex that maximally responds to specific spot- or edge-like stimulus patterns within the receptive field, i.e., a patch in our case. In particular, each hidden unit in a Gaussian RBM finds simple volumetric or functional patterns in the input 3D patch by assigning different weights to the corresponding voxels. For example, hidden units of the Gaussian RBM for MRI (left in Fig. 7(a)) focus on different parts of a patch to detect a simple spot- or edge-like pattern in the input 3D GM patch. The hidden units in a Gaussian RBM for PET (left in Fig. 7(b)) can be understood as descriptors that discover local functional relations among voxels within a patch.

Note that the hidden units of a Gaussian RBM for either MRI or PET find, respectively, the structural or functional relations among voxels in a localized way. Meanwhile, the hidden units in our (MM-)DBM served as complex filters of a human visual cortex that combine the outputs from the simple cells and maximally responds to more complex patterns

within the receptive field. For example, the weights of hidden units in the hidden layer of the MRI pathway in an MM-DBM (right in Fig. 7(a)) discover more complicated structural patterns in the input 3D GM patch, such as combination of edges orienting in different directions. With respect to the PET, the weights of hidden units in the hidden layer of the PET pathway in an MM-DBM (right in Fig. 7(b)) discover non-linear functional relations among voxels within a 3D patch. In this way, as it forwards to the higher layer, the MM-DBM finds complex latent features in the input patch, and ultimately in the top hidden layer, the hidden units discover the inter-modality relations in between the pair of MRI and PET patches, each of which comes from the same location in a brain.

Performance evaluation

Let TP, TN, FP, and FN denote, respectively, True Positive, True Negative, False Positive, and False Negative. In this work, we consider the following quantitative measurements and presented the performances of the competing methods in Table 2.

- ACCuracy (ACC) = (TP + TN) / (TP + TN + FP + FN)
- SENsitivity (SEN) = TP / (TP + FN)
- SPECificity (SPEC) = TN / (TN + FP)
- Balanced ACcuracy (BAC) = (SEN + SPEC) / 2
- Positive Predictive Value (PPV) = TP / (TP + FP)
- Negative Predictive Value (NPV) = TN / (TN + FN)
- Area Under the receiver operating characteristic Curve (AUC)

In the classification of AD and NC, the proposed method showed the mean accuracies of 92.38% (MRI), 92.20% (PET), and 95.35% (MRI + PET). Compared to Liu et al.'s method that showed the accuracies of 90.18% (MRI), 89.13% (PET), and 90.27% (MRI + PET),[18] the proposed method improved by 2.2% (MRI), 3.07% (PET), and 5.08% (MRI + PET). That is, the proposed method outperformed Liu et al.'s method in all the cases of MRI, PET, and MRI + PET. In the discrimination of MCI from NC, the proposed method showed the accuracies of 84.24% (MRI), 84.29% (PET), and 85.67% (MRI + PET). Meanwhile, Liu et al.'s method showed the accuracies of 81% (MRI), 81.14% (PET), and 83.90% (MRI + PET). Again, the proposed method outperformed Liu et al.'s method by making performance improvements of 3.24% (MRI), 3.15% (PET), and 1.77% (MRI + PET). In the classification between MCI-C and MCI-NC, which is the most important for early diagnosis and treatment, Liu et al.'s method achieved the accuracies of 64.75% (MRI), 67.17% (PET), and 73.33% (MRI + PET). Compared to these results, the proposed method improved the accuracies by 7.67% (MRI), 3.58% (PET), and 2.59% (MRI + PET), respectively. Concisely, in our three binary classifications, based on the classification accuracy, the proposed method clearly outperformed Liu et al.'s method by achieving the maximal accuracies of 95.35% (AD vs. NC), 85.67% (MCI vs. NC), and 75.92% (MCI-C vs. MCI-NC), respectively.

Regarding sensitivity and specificity, the higher the sensitivity, the lower the chance of mis-diagnosing AD/MCI patients; also the higher the specificity, the lower the chance of mis-diagnosing NC to AD/MCI. Although the proposed method had a lower sensitivity than that of Liu et al.'s method for a couple of cases, e.g., 90.06% (Liu et al.'s method) vs. 88.04% (proposed) with PET in the AD diagnosis, 98.97% (Liu et al.'s method) vs. 95.37% (proposed) with MRI + PET in the MCI diagnosis, and 40.02% (Liu et al.'s method) vs. 25.45% (proposed) with PET in the MCI-C diagnosis, in general, the proposed method showed higher sensitivity and specificity in all three classification problems. Hence, from a clinical point of view, the proposed method is less likely to mis-diagnose subjects with AD/MCI and vice versa, compared to Liu et al.'s method.

---

[15] Available at 'http://www.cs.toronto.edu/rsalakhu/DBM.html'.
[16] Available at 'http://www.csie.ntu.edu.tw/cjlin/libsvm/'.
[17] For the hidden units of the MRI pathway and the PET pathway, their weights were visualized as a weighted linear combination of the weights of the Gaussian RBM, similar to Lee et al.'s (2009) work.
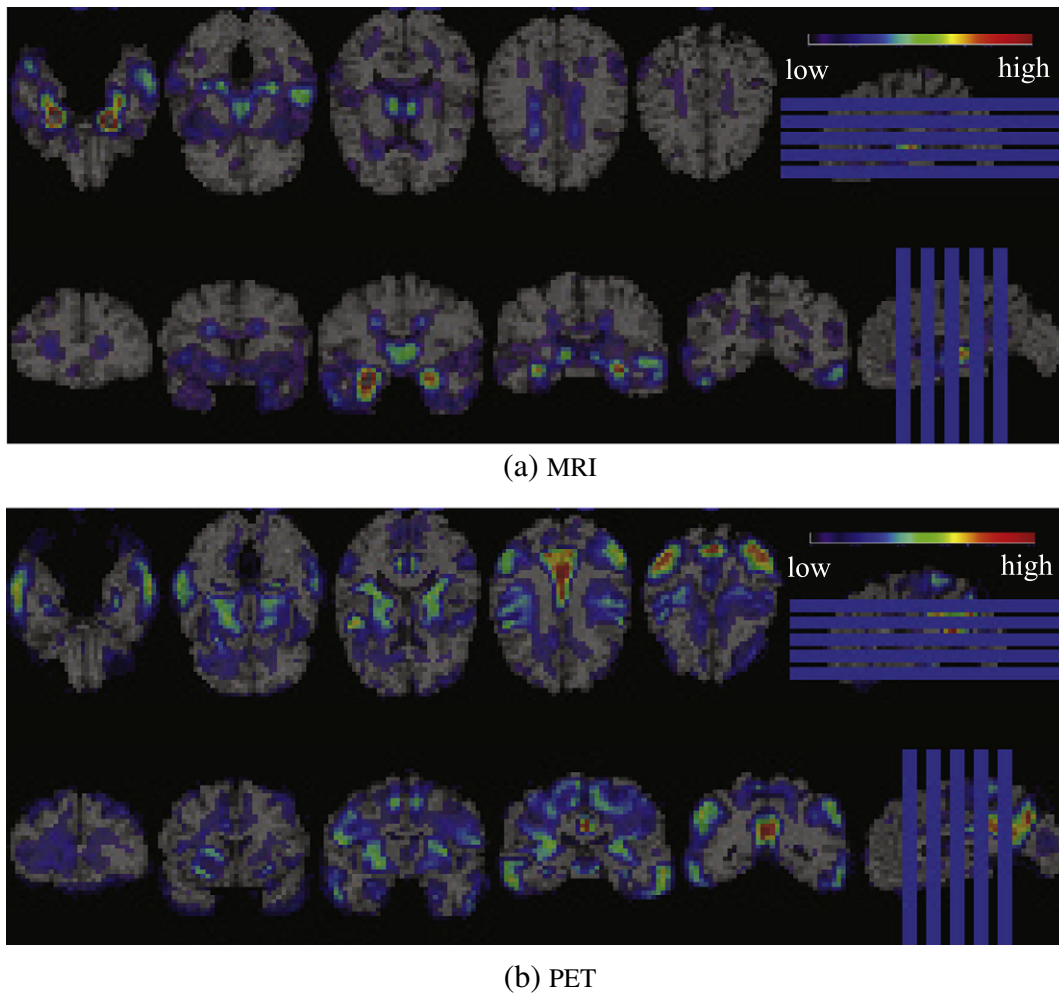
[18] For the multimodal case, we concatenated the patches of modalities into a single vector for Liu et al.'s method.

(a) MRI

(b) PET

Fig. 5. Visualization of the *p*-value distributions used to select the patch locations of MRI and PET in AD and NC classification.

Meanwhile, because of the data imbalance between classes, i.e., AD (93 subjects), MCI (204 subjects; 76 MCI-C and 128 MCI-NC subjects), and NC (101 subjects), we obtained low sensitivity (MCI vs. NC) or specificity (MCI-C vs. MCI-NC). The balanced accuracy, which is calculated by taking the average of sensitivity and specificity, avoids inflated performance estimates on imbalanced datasets. Based on this metric, we clearly see that the proposed method is superior to the competing method. Note that in discrimination between MCI and NC, while the accuracy improvement by the proposed method with MRI + PET was 1.43% and 1.38% compared to the same method with MRI and PET,

(a) MRI   (b) PET
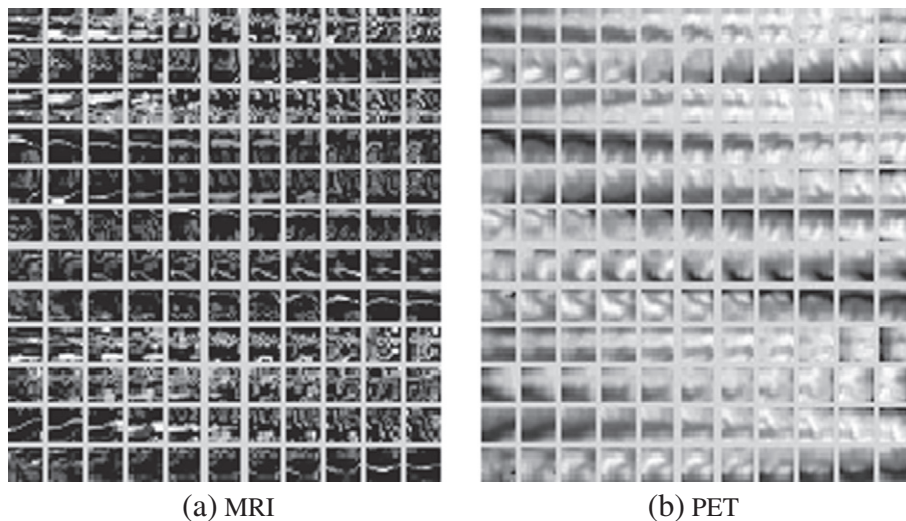
Fig. 6. Samples of the selected patches, whose voxel values are the input to the (MM-)DBM.

(a) MRI

(b) PET
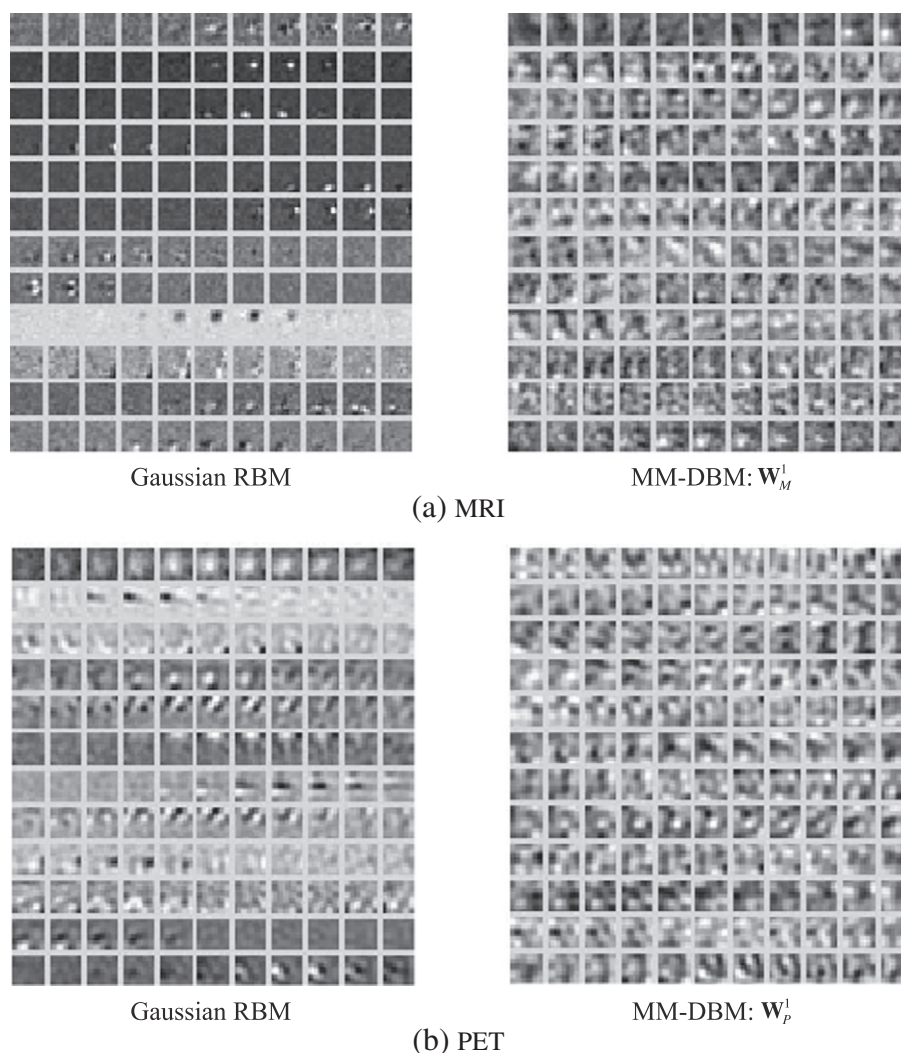
Fig. 7. Visualization of the trained weights of our modality-specific Gaussian RBMs (left) used for data conversion from a real-valued vector to a binary vector, and those of our MM-DBM (right) used for latent feature representations. For the weights of our MM-DBM, they correspond to the first hidden layer in the respective modality's pathway in the model. In each subfigure, one row corresponds to one hidden unit in the respective Gaussian RBM or MM-DBM.

respectively, in terms of the balanced accuracy, the improvements went up to 3.93% (vs. MRI) and 2.95% (vs. PET).

With a further concern on low sensitivity and specificity, especially in classifications of MCI vs. NC and MCI-C vs. MCI-NC, we also computed a Positive Predictive Value (PPV) and a Negative Predictive Value (NPV). Statistically, PPV and NPV measure, respectively, the proportion of subjects with AD, MCI, or MCI-C who are correctly diagnosed as patients, and the proportion of subjects without AD, MCI, or MCI-C who are

Table 2
A summary of the performances of two methods. The boldface denotes the best performance in each metric for each classification task.

	Method	Modality	ACC (%)	SEN (%)	SPEC (%)	BAC (%)	PPV (%)	NPV (%)	AUC (%)
AD/NC	Liu et al.	MRI	90.18 ± 5.25	91.54	90.61	91.08	88.94	90.67	0.9620
		PET	89.13 ± 6.81	90.06	89.36	89.71	88.49	89.26	0.9594
		MRI + PET	90.27 ± 7.02	89.48	92.44	90.96	90.56	88.70	0.9655
	Proposed	MRI	92.38 ± 5.32	91.54	94.56	93.05	92.65	90.84	0.9697
		PET	92.20 ± 6.70	88.04	**96.33**	92.19	95.03	89.66	0.9798
		MRI + PET	**95.35 ± 5.23**	**94.65**	95.22	**94.93**	**96.80**	**95.67**	**0.9877**
MCI/NC	Liu et al.	MRI	81.00 ± 4.98	97.08	48.18	72.63	79.14	88.99	0.8352
		PET	81.14 ± 10.22	96.03	52.59	74.31	80.26	84.16	0.8231
		MRI + PET	83.90 ± 5.80	98.97	52.59	75.78	81.18	97.22	0.8301
	Proposed	MRI	84.24 ± 6.26	**99.58**	53.79	76.69	81.23	**98.75**	0.8478
		PET	84.29 ± 7.22	98.69	56.87	77.78	81.99	94.57	0.8297
		MRI + PET	**85.67 ± 5.22**	95.37	**65.87**	**80.62**	**85.02**	89.00	**0.8808**
MCI-C/MCI-NC	Liu et al.	MRI	64.75 ± 14.83	22.22	89.57	55.90	46.29	77.39	0.6355
		PET	67.17 ± 13.43	40.02	82.61	61.32	64.13	70.31	0.6911
		MRI + PET	73.33 ± 12.47	33.25	**97.52**	65.38	80.00	73.18	0.7159
	Proposed	MRI	72.42 ± 13.09	36.70	90.98	63.84	65.49	**77.84**	0.7342
		PET	70.75 ± 13.23	25.45	96.55	61.00	75.00	70.69	0.7215
		MRI + PET	**75.92 ± 15.37**	**48.04**	95.23	**71.63**	**83.50**	74.33	**0.7466**

correctly diagnosed as cognitive normal. Based on a recent report by Alzheimer's Association (2012), the AD prevalence is projected to be 11 million to 16 million by 2050. For MCI and MCI-C, although there is high variation among reports depending on definitions, the median of the prevalence estimates of MCI or MCI-C in the literature is 26.4% (MCI) and 4.9% (amnestic MCI) (Ward et al., 2012). Regarding the AD prevalence by 2050, the proposed method, which achieved 96.80% of the PPV in the classification of AD and NC, can correctly identify 10.648 million to 15.488 million of subjects with AD while Liu et al.'s method, whose respective PPV was 90.56%, can identify 9.9616 million to 14.4896 million of subjects with AD. Accordingly, our method can correctly identify as many as 0.6864 million to 0.9984 million of subjects more.

The Receiver Operating Characteristic (ROC) curve[19] and the Area Under the ROC Curve (AUC) are also widely used metrics to evaluate the performance of diagnostic tests in brain disease as well as other medical areas. In particular, the AUC can be thought as a measure of the overall performance of a diagnostic test. The proposed method with MRI + PET showed the best AUCs of 0.9877 in AD vs. NC, 0.8808 in MCI vs. NC, and 0.7466 in MCI-C vs. MCI-NC. Compared to Liu et al.'s method with MRI + PET, the proposed multimodal method increased the AUCs by 0.0222 (AD vs. NC), 0.0507 (MCI vs. NC), and 0.0307 (MCI-C vs. MCI-NC). Noticeably, the proposed method with MRI enhanced the AUC as much as 0.0987 than the corresponding AUC of Liu et al.'s method. It is also noteworthy that in the classification of MCI and NC, the proposed method with MRI + PET improved the AUC by 0.0330 (vs. MRI) and 0.0389 (vs. PET), while the improvements in the classifications of AD vs. NC and MCI-C vs. MCI-NC were, respectively, 0.0180/0.0079 (vs. MRI/PET) and 0.0124/0.0251 (vs. MRI/PET).

Based on the quantitative measurements depicted above, the proposed method clearly outperforms Liu et al.'s method. In terms of modalities used for classification, similar to the previous work (Hinrichs et al., 2011; Suk and Shen, 2013; Zhang et al., 2011), we also obtained the best performances with the complementary information from multiple modalities, i.e., MRI + PET.

Comparison with State-of-the-Art Methods

In Table 3, we also compared the classification accuracies of the proposed method with those of the state-of-the-art methods that considered multi-modality in classifications of AD vs. NC, MCI vs. NC, and MCI-C vs. MCI-NC. Note that, due to different datasets and different approaches of extracting features and building classifiers, it is not fair to directly compare the performances among methods. Nonetheless, it is remarkable that the proposed method showed the highest accuracies among the methods in all the binary classification problems. It is also worth noting that our method is the only one that considered the patch-based approach for feature extraction, while the other methods used an ROI-based approach.

Importance of brain areas in classification

For the investigation of the relative importance of different brain areas determined by the proposed method for AD/MCI diagnosis, we visualized the weights of the selected patches in Fig. 8. Specifically, the weight of each patch was calculated by accumulating the selected frequency of mega-patches in final ensemble classifiers over cross-validations. That is, the weight of a patch was determined with the sum of the weights of the mega-patches that included the patch and was used in the final decision. The high weighted patches were in accordance with the previous reports on AD/MCI studies. Those were distributed around a medial temporal lobe (that includes amygdala, hippocampal formation, entorhinal cortex) (Braak and

Braak, 1991; Burton et al., 2009; Desikan et al., 2009; Devanand et al., 2007; Ewers et al., 2012; Lee et al., 2006; Mosconi, 2005; Visser et al., 2002; Walhovd et al., 2010), superior/medial frontal gyrus (Johnson et al., 2005), precentral/postcentral gyrus (Belleville et al., 2011), precuneus (Bokde et al., 2006; Davatzikos et al., 2011; Singh et al., 2006), thalamus, putamen (de Jong et al., 2008), caudate nucleus (Dai et al., 2009), etc.

Limitations

In our experiments, we validated the efficacy of the proposed method in three classification problems by achieving the best performances. However, there still exist some limitations of the proposed method.

First, even though we could visualize the trained weights in our MM-DBMs in Fig. 7, from a clinical perspective, it is difficult to understand or interpret the resulting feature representations. Particularly, with respect to the investigation of brain abnormalities affected by neurodegenerative disease, i.e., AD or MCI, our method cannot provide useful clinical information. In this regard, it could be a good research direction in which we further extend the proposed method to find or detect brain abnormalities in terms of brain regions or areas for easy understanding to clinicians.

Second, in our experiments, we manually determined the number of hidden units in each layer. Furthermore, we used a relatively small data samples (93 AD, 76 MCI-C, 128 MCI-NC, and 101 NC). Therefore, the network structures used to discover high-level feature representations in our experiments were not necessarily optimal. We believe that it needs more intensive studies such as learning the optimal network structure from big data for practical use of deep learning in clinical settings.

Third, as the graphical model illustrated in Fig. 4, the current method only considers bi-modalities of MRI and PET. However, it is generally beneficiary to combine as many modalities as possible to use their richer information. Therefore, it is necessary to build a more systematic model that can efficiently find and use complementary information from genetics, proteomics, imaging, cognition, disease status, and other phenotypic modalities.

Lastly, according to a recent broad spectrum of studies, there are increasing evidences that subjective cognitive complaint is one of the important genetic risk factors, which increases the risk of progression to MCI or AD (Loewenstein et al., 2012; Mark and Sitskoorn, 2013). That is, among the cognitively normal elderly individuals who have subjective cognitive impairments, there exists a high possibility for some of them to be in the stage of 'pre-MCI'. However, in the ADNI dataset, there is no related information. Thus, in our experiments, the NC group could include both genuine controls and those with subjective cognitive complaints.

Conclusions

In this paper, we proposed a method for a shared latent feature representation from MRI and PET in deep learning. Specifically, we used DBM to find a latent feature representation from a volumetric patch and further devised method to systemically discover a joint feature representation from multi-modality. Unlike the previous methods that mostly considered the direct use of the GM tissue densities from MRI and/or voxel intensities from PET and then fused the complementary information in a kernel technique, the proposed method learned high-level features in a self-taught manner via deep learning, and thus could efficiently combine the complimentary information from MRI and PET during feature representation procedure. Experimental results on ADNI dataset showed that the proposed method is superior to the previous methods in terms of various quantitative metrics.

---

[19] A plot of test true positive rate versus its false positive rate.

Table 3
Comparison of classification accuracy with state-of-the-art methods. The numbers in the parentheses denote the number of AD/MCI(MCI-C, MCI-NC)/NC subjects in the dataset used. The boldface denotes the best performance in each classification task.

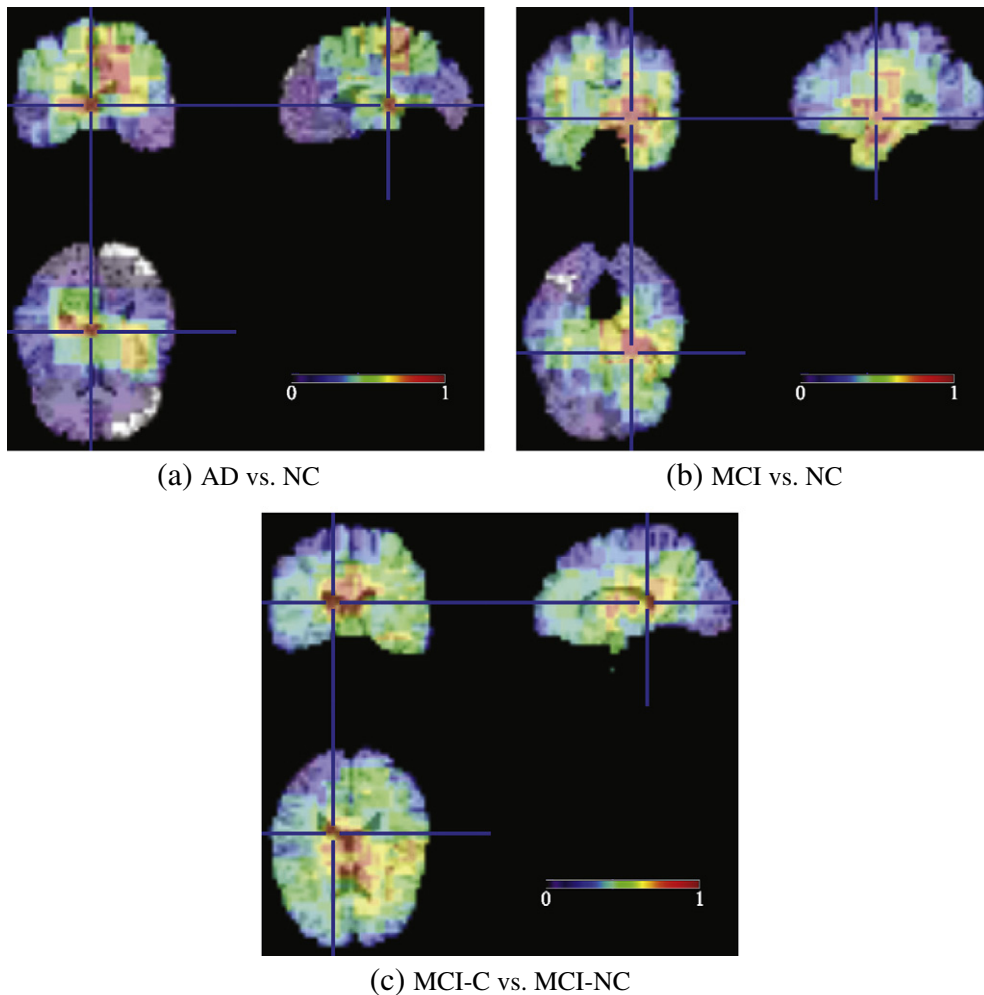| Methods	Dataset	Features	AD vs. NC (%)	MCI vs. NC (%)	MCI-C vs. MCI-NC (%)
Kohannim et al. (2010)	MRI + PET + CSF (40/83(43,40)/43)	ROI	90.7	75.8	n/a
Walhovd et al. (2010)	MRI + CSF (38/73/42)	ROI	88.8	79.1	n/a
Hinrichs et al. (2011)	MRI + PET (48/119(38,81)/66)	ROI	92.4	n/a	72.3
Westman et al. (2012)	MRI + CSF (96/162(81,81)/111)	ROI	91.8	77.6	66.4
Zhang and Shen (2012)	MRI + PET + CSF (45/91(43,48)/50)	ROI	93.3	83.2	73.9
Proposed method	MRI + PET (93/204(76,128)/101)	Patch	**95.35**	**85.67**	**75.92**

Appendix A. Variational approximation for DBM learning

The main idea of applying variational approximation is to assume that the true posterior distribution over latent variables $P(\mathbf{h}^1, \mathbf{h}^2 \mid \mathbf{v}; \Theta)$ for each training vector $\mathbf{v}$ is unimodal and can be replaced by an approximate posterior $Q(\mathbf{h}^1, \mathbf{h}^2 \mid \mathbf{v}; \Omega)$, which can be computed efficiently, and the parameters are updated to maximize the variational lower bound on the log-likelihood

$$logP(\mathbf{v}; \Theta) \geq \sum_{\mathbf{h}^1, \mathbf{h}^2} Q\left(\mathbf{h}^1, \mathbf{h}^2 \middle| \mathbf{v}; \Omega\right) logP\left(\mathbf{v}, \mathbf{h}^1, \mathbf{h}^2; \Theta\right) + H(Q) \qquad (A.1)$$

$$= logP(\mathbf{v}; \Theta) - KL\left[Q\left(\mathbf{h}^1, \mathbf{h}^2 \middle| \mathbf{v}; \Omega\right) \middle\| P\left(\mathbf{h}^1, \mathbf{h}^2 \middle| \mathbf{v}; \Theta\right)\right] \qquad (A.2)$$

(a) AD vs. NC

(b) MCI vs. NC

(c) MCI-C vs. MCI-NC

Fig. 8. Patch weight distributions in classification of AD vs. NC and MCI vs. NC.

where $H(\cdot)$ is the entropy functional, KL$[\cdot\|\cdot]$ denotes Kullback–Leibler divergence, and Ω is a variational parameter set.

For computational simplicity and learning speed, the naïve mean-field approximation, which uses a fully factorized distribution, is generally used in the literature (Tanaka, 1998). That is,

$$Q\left(\mathbf{h}^1, \mathbf{h}^2 \middle| \mathbf{v}; \mathbf{\Omega}\right) = \prod_{j=1}^{F_1} q\left(h_j^1\right) \prod_{k=1}^{F_2} q\left(h_k^2\right) \tag{A.3}$$

where $\mathbf{\Omega} = \{\boldsymbol{\mu}^1, \boldsymbol{\mu}^2\}, \boldsymbol{\mu}^1 = \left(\mu_1^1, \cdots, \mu_{F_1}^1\right), \boldsymbol{\mu}^2 = \left(\mu_1^2, \cdots, \mu_{F_2}^2\right), q(h_j^1 = 1) = \mu_j^1$ $(j \in \{1, \cdots, F_1\})$, and $q(h_k^2 = 1) = \mu_k^2$ $(k \in \{1, \cdots, F_2\})$. It alternatively estimates the state of the hidden units, $\boldsymbol{\mu}^1$ and $\boldsymbol{\mu}^2$, for fixed Θ until convergence:

$$\mu_j^1 \leftarrow sigm\left(\sum_{i=1}^{D} W_{ij}^1 v_i + \sum_{k=1}^{F_2} W_{jk}^2 \mu_k^2\right) \tag{A.4}$$

$$\mu_k^2 \leftarrow sigm\left(\sum_{j=1}^{F_1} W_{jk}^2 \mu_j^1 + \sum_{l=1}^{C} U_{lk} o_l\right). \tag{A.5}$$

Regarding the data-independent statistics, we apply a stochastic approximation procedure to obtain samples, also called particles, of $\widetilde{\mathbf{v}}, \widetilde{\mathbf{h}}^1$, $\widetilde{\mathbf{h}}^2$, and $\widetilde{\mathbf{o}}$ by running repeatedly the alternate Gibbs sampler on a set of particles. Once both the data-dependent and data-independent statistics are computed, we then update parameters as follows:

$$\mathbf{W}^{1,(t+1)} = \mathbf{W}^{1,(t)} + \alpha_t \left(\frac{1}{N} \sum_{n=1}^{N} \mathbf{v}^n \boldsymbol{\mu}^{1,n\top} - \frac{1}{M} \sum_{m=1}^{M} \widetilde{\mathbf{v}}^m \widetilde{\mathbf{h}}^{1,m\top}\right) \tag{A.6}$$

$$\mathbf{W}^{2,(t+1)} = \mathbf{W}^{2,(t)} + \alpha_t \left(\frac{1}{N} \sum_{n=1}^{N} \boldsymbol{\mu}^{1,n} \boldsymbol{\mu}^{2,n\top} - \frac{1}{M} \sum_{m=1}^{M} \widetilde{\mathbf{h}}^{1,m} \widetilde{\mathbf{h}}^{2,m\top}\right) \tag{A.7}$$

$$\mathbf{U}^{(t+1)} = \mathbf{U}^{(t)} + \alpha_t \left(\frac{1}{N} \sum_{n=1}^{N} \boldsymbol{\mu}^{2,n} \mathbf{o}^{n\top} - \frac{1}{M} \sum_{m=1}^{M} \widetilde{\mathbf{h}}^{2,m} \widetilde{\mathbf{o}}^{n\top}\right) \tag{A.8}$$

where α_t is a learning rate, and N and M denote, respectively, the numbers of training data and particles, and superscripts n and m denote, respectively, indices of an observation and a particle.

Appendix B. Learning multimodal DBM parameters

The same approach to the unimodal DBM described in the Deep Boltzmann Machine section can be applied, i.e., iterative alternation of the variational mean-field approximation for data-dependent statistics and the stochastic approximation procedure for data-independent statistics, and parameters update. Let $\mathbf{H} = \{\mathbf{h}_M^1, \mathbf{h}_M^2, \mathbf{h}_P^1, \mathbf{h}_P^2, \mathbf{h}_S^3\}$ and $\mathbf{V} = \{\mathbf{v}_M, \mathbf{v}_P\}$. In variational learning of our MM-DBM, a fully factorized mean-field variational function for approximation of the true posterior distribution $P(\mathbf{H} \mid \mathbf{V}, \mathbf{o}; \Theta = \{\mathbf{W}_M^1, \mathbf{W}_M^2, \mathbf{W}_P^1, \mathbf{W}_P^2, \mathbf{W}_S^3, \mathbf{U}\})$ is defined as follows:

$$Q(\mathbf{H}|\mathbf{V}, \mathbf{o}; \mathbf{\Omega}) = \prod_{i=1}^{F_{M1}} q\left(h_{M,i}^1\right) \prod_{j=1}^{F_{M2}} q\left(h_{M,j}^2\right) \prod_{k=1}^{F_{P1}} q\left(h_{P,k}^1\right) \prod_{l=1}^{F_{P2}} q\left(h_{P,l}^2\right) \prod_{m=1}^{F_S} q\left(h_{S,m}^3\right) \tag{B.1}$$

$$= \prod_{i=1}^{F_{M1}} \mu_{M,i}^1 \prod_{j=1}^{F_{M2}} \mu_{M,j}^2 \prod_{k=1}^{F_{P1}} \mu_{P,k}^1 \prod_{l=1}^{F_{P2}} \mu_{P,l}^2 \prod_{m=1}^{F_S} \mu_{S,m}^3$$

where $\mathbf{\Omega} = \{\boldsymbol{\mu}_M^1, \boldsymbol{\mu}_M^2, \boldsymbol{\mu}_P^1, \boldsymbol{\mu}_P^2, \boldsymbol{\mu}_S^3\}$ is a mean-field parameter set with $\boldsymbol{\mu}_M^1 = \left(\mu_{M,1}^1, \cdots, \mu_{M,F_{M1}}^1\right), \boldsymbol{\mu}_M^2 = \left(\mu_{M,1}^2, \cdots, \mu_{M,F_{M2}}^2\right), \boldsymbol{\mu}_P^1 = \left(\mu_{P,1}^1, \cdots, \mu_{P,F_{P1}}^1\right), \boldsymbol{\mu}_P^2 = \left(\mu_{P,1}^2, \cdots, \mu_{P,F_{P2}}^2\right)$, and $\boldsymbol{\mu}_S^3 = \left(\mu_{S,1}^3, \cdots, \mu_{S,F_S}^3\right)$. Referring Eqs. (A.4) and (A.5), given a fixed model parameter Θ, it is straightforward to estimate the mean-field parameters $\mathbf{\Omega}$.

The learning proceeds by iteratively alternating the variational mean-field inference to find the values of $\mathbf{\Omega}$ for the fixed current model parameters Θ and the stochastic approximation procedure to update model parameters Θ given the variational parameters $\mathbf{\Omega}$. Finally, the shared feature representations can be obtained by inferring the values of the hidden units in the top hidden layer from the trained MM-DBM.

References

Alzheimer's Association, 2012. 2012 Alzheimer's disease facts and figures. Alzheimers Dement. 8, 131–168.

Baron, J., Chtelat, G., Desgranges, B., Perchey, G., Landeau, B., de la Sayette, V., Eustache, F., 2001. In vivo mapping of gray matter loss with voxel-based morphometry in mild Alzheimer's disease. Neuroimage 14, 298–309.

Belleville, S., ClŽment, F., Mellah, S., Gilbert, B., Fontaine, F., Gauthier, S., 2011. Training-related brain plasticity in subjects at risk of developing Alzheimer's disease. Brain 134, 1623–1634.

Bengio, Y., 2009. Learning deep architectures for AI. Foundations and Trends in Machine Learning, 2, pp. 1–127.

Bengio, Y., Lamblin, P., Popovici, D., Larochelle, H., 2007. Greedy layer-wise training of deep networks. In: Schölkopf, B., Platt, J., Hoffman, T. (Eds.), Advances in Neural Information Processing Systems, 19. MIT Press, Cambridge, MA, pp. 153–160.

Bokde, A.L.W., Lopez-Bayo, P., Meindl, T., Pechler, S., Born, C., Faltraco, F., Teipel, S.J., Möller, H.J., Hampel, H., 2006. Functional connectivity of the fusiform gyrus during a face-matching task in subjects with mild cognitive impairment. Brain 129, 1113–1124.

Braak, H., Braak, E., 1991. Neuropathological stageing of Alzheimer-related changes. Acta Neuropathol. 82, 239–259.

Burton, E.J., Barber, R., Mukaetova-Ladinska, E.B., Robson, J., Perry, R.H., Jaros, E., Kalaria, R.N., OBrien, J.T., 2009. Medial temporal lobe atrophy on MRI differentiates Alzheimer's disease from dementia with Lewy bodies and vascular cognitive impairment: a prospective study with pathological verification of diagnosis. Brain 132, 195–203.

Catana, C., Drzezga, A., Heiss, W.D., Rosen, B.R., 2012. PET/MRI for neurologic applications. J. Nucl. Med. 53, 1916–1925.

Ciresan, D.C., Giusti, A., Gambardella, L.M., Schmidhuber, J., 2013. Mitosis detection in breast cancer histology images with deep neural networks. Medical Image Computing and Computer-Assisted Intervention MICCAI 2013, pp. 411–418.

Cui, Y., Liu, B., Luo, S., Zhen, X., Fan, M., Liu, T., Zhu, W., Park, M., Jiang, T., Jin, J.S., the Alzheimer's Disease Neuroimaging Initiative, 2011. Identification of conversion from mild cognitive impairment to Alzheimer's disease using multivariate predictors. PLoS One 6, e21896.

Cuingnet, R., Gerardin, E., Tessieras, J., Auzias, G., Lehéricy, S., Habert, M.O., Chupin, M., Benali, H., Colliot, O., The Alzheimer's Disease Neuroimaging Initiative, 2011. Automatic classification of patients with Alzheimer's disease from structural MRI: a comparison of ten methods using ADNI database. Neuroimage 56, 766–781.

Dai, W., Lopez, O., Carmichael, O., Becker, J., Kuller, L., Gach, H., 2009. Mild cognitive impairment and Alzheimer disease: patterns of altered cerebral blood flow at MR imaging. Radiology 250, 856–866.

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, 1361–1369.

Davatzikos, C., Bhatt, P., Shaw, L.M., Batmanghelich, K.N., Trojanowski, J.Q., 2011. Prediction of MCI to AD conversion, via MRI, CSF biomarkers, and pattern classification. Neurobiol. Aging 32 (2322.e19–2322.e27).

de Jong, L.W., van der Hiele, K., Veer, I.M., Houwing, J.J., Westendorp, R.G.J., Bollen, E.L.E.M., de Bruin, P.W., Middelkoop, H.A.M., van Buchem, M.A., van der Grond, J., 2008. Strongly reduced volumes of putamen and thalamus in Alzheimer's disease: an MRI study. Brain 131, 3277–3285.

Desikan, R., Cabral, H., Hess, C., Dillon, W., Salat, D., Buckner, R., Fischl, B., Initiative, A.D.N., 2009. Automated MRI measures identify individuals with mild cognitive impairment and Alzheimer's disease. Brain 132, 2048–2057.

Devanand, D.P., Pradhaban, G., Liu, X., Khandji, A., De Santi, S., Segal, S., Rusinek, H., Pelton, G.H., Hoing, L.S., Mayeux, R., Stern, Y., Tabert, M.H., de Leon, J.J., 2007. Hippocampal and ento-rhinal atrophy in mild cognitive impairment. Neurology 68, 828–836.

Dinov, I., Boscardin, J., Mega, M., Sowell, E., Toga, A., 2005. A wavelet-based statistical analysis of fMRI data. Neuroinformatics 3, 319–342.

Ewers, M., Walsh, C., Trojanowski, J.Q., Shaw, L.M., Petersen Jr., R.C., C.R.J., Feldman, H.H., Bokde, A.L., Alexander, G.E., Scheltens, P., Vellas, B., Dubois, B., Weiner, M., Hampel, H., 2012. Prediction of conversion from mild cognitive impairment to Alzheimer's disease dementia based upon biomarkers and neuropsychological test performance. Neurobiol. Aging 33, 1203–1214 (e2).

Fan, Y., Rao, H., Hurt, H., Giannetta, J., Korczykowski, M., Shera, D., Avants, B.B., Gee, J.C., Wang, J., Shen, D., 2007a. Multivariate examination of brain abnormality using both structural and functional MRI. Neuroimage 36, 1189–1199.

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

Friston, K.J., 1995. Functional and effective connectivity in neuroimaging: a synthesis. Hum. Brain Mapp. 2, 56–78.

Greicius, M.D., Srivastava, G., Reiss, A.L., Menon, V., 2004. Default-mode network activity distinguishes Alzheimer's disease from healthy aging: evidence from functional MRI. Proc. Natl. Acad. Sci. U. S. A. 101, 4637–4642.

Hackmack, K., Paul, F., Weygandt, M., Allefeld, C., Haynes, J.D., 2012. Multi-scale classification of disease using structural MRI and wavelet transform. Neuroimage 62, 48–58.

Hinrichs, C., Singh, V., Xu, G., Johnson, S.C., 2011. Predictive markers for AD in a multi-modality framework: an analysis of MCI progression in the ADNI population. Neuroimage 55, 574–589.

Hinton, G.E., Salakhutdinov, R.R., 2006. Reducing the dimensionality of data with neural networks. Science 313, 504–507.

Hinton, G.E., Osindero, S., Teh, Y.W., 2006. A fast learning algorithm for deep belief nets. Neural Comput. 18, 1527–1554.

Hjelm, R.D., Calhoun, V.D., Salakhutdinov, R., Allen, E.A., Adali, T., Plis, S.M., 2014. Restricted boltzmann machines for neuroimaging: an application in identifying intrinsic networks. Neuroimage 96, 245–260.

Ishii, K., Kawachi, T., Sasaki, H., Kono, A.K., Fukuda, T., Kojima, Y., Mori, E., 2005. Voxel-based morphometric comparison between early- and late-onset mild Alzheimer's disease and assessment of diagnostic performance of z score images. Am. J. Neuroradiol. 26, 333–340.

Jia, H., Wu, G., Wang, Q., Shen, D., 2010. ABSORB: atlas building by self-organized registration and bundling. Neuroimage 51, 1057–1070.

Johnson, N.A., Jahng, G.H., Weiner, M.W., Miller, B.L., Chui, H.C., Jagust, W.J., Gorno-Tempini, M.L., Schuff, N., 2005. Pattern of cerebral hypoperfusion in Alzheimer disease and mild cognitive impairment measured with arterial spin-labeling MR imaging: initial experience. Radiology 234, 851–859.

Kabani, N., MacDonald, D., Holmes, C., Evans, A., 1998. A 3D atlas of the human brain. Neuroimage 7, S717.

Kohannim, O., Hua, X., Hibar, D.P., Lee, S., Chou, Y.Y., Toga Jr., A.W., C.R.J., Weiner, M.W., Thompson, P.M., 2010. Boosting power for clinical trials using classifiers based on multiple biomarkers. Neurobiol. Aging 31, 1429–1442.

Larochelle, H., Bengio, Y., 2008. Classification using discriminative restricted Boltzmann machines. Proceedings of the 25th International Conference on Machine Learning, pp. 536–543.

LeCun, Y., Bottou, L., Bengio, Y., Haffner, P., 1998. Gradient-based learning applied to document recognition. Proc. IEEE 86, 2278–2324.

Lee, A.C.H., Buckley, M.J., Gaffan, D., Emery, T., Hodges, J.R., Graham, K.S., 2006. Differentiating the roles of the hippocampus and perirhinal cortex in processes beyond long-term declarative memory: a double dissociation in dementia. J. Neurosci. 26, 5198–5203.

Lee, H., Grosse, R., Ranganath, R., Ng, A.Y., 2009. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. Proceedings of the 26th International Conference on Machine Learning, pp. 609–616.

Li, Y., Wang, Y., Wu, G., Shi, F., Zhou, L., Lin, W., Shen, D., 2012. Discriminant analysis of longitudinal cortical thickness changes in Alzheimer's disease using dynamic and network features. Neurobiol. Aging 33 (427.e15–427.e30).

Liao, S., Gao, Y., Oto, A., Shen, D., 2013. Representation learning: a unified deep learning framework for automatic prostate MR segmentation. Medical Image Computing and Computer-Assisted Intervention MICCAI 2013. Lecture Notes in Computer Science, vol. 8150, pp. 254–261.

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

Liu, M., Zhang, D., Shen, D., the Alzheimer's Disease Neuroimaging Initiative, 2013. Hierarchical fusion of features and classifier decisions for Alzheimer's disease diagnosis. Hum. Brain Mapp. 35, 1305–1319.

Loewenstein, D.A., Greig, M.T., Schinka, J.A., Barker, W., Shen, Q., Potter, E., Raj, A., Brooks, L., Varon, D., Schoenberg, M., Banko, J., Potter, H., Duara, R., 2012. An investigation of PreMCI: subtypes and longitudinal outcomes. Alzheimers Dement. 8, 172–179.

Mark, R.E., Sitskoorn, M., 2013. Are subjective cognitive complaints relevant in preclinical Alzheimer's disease? A review and guidelines for healthcare professionals. Rev. Clin. Gerontol. 23, 61–74.

Mohamed, A., Dahl, G.E., Hinton, G.E., 2012. Acoustic modeling using deep belief networks. IEEE Trans. Audio Speech Lang. Process. 20, 14–22.

Montavon, G., Braun, M.L., Möller, K.R., 2012. Deep Boltzmann machines as feed-forward hierarchies. Journal of Machine Learning Research—Proceedings Track, 22, pp. 798–804.

Mosconi, L., 2005. Brain glucose metabolism in the early and specific diagnosis of Alzheimer's disease. Eur. J. Nucl. Med. Mol. Imaging 32, 486–510.

Nair, V., Hinton, G.E., 2008. Implicit mixtures of restricted Boltzmann machines. Advances in Neural Information Processing Systems, pp. 1145–1152.

Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., Ng, A.Y., 2011. Multimodal deep learning. Proceedings of the 28th International Conference on Machine Learning, pp. 689–696.

Nordberg, A., Rinne, J.O., Kadir, A., Langstrom, B., 2010. The use of PET in Alzheimer disease. Nat. Rev. Neurol. 6, 78–87.

Perrin, R.J., Fagan, A.M., Holtzman, D.M., 2009. Multimodal techniques for diagnosis and prognosis of Alzheimer's disease. Nature 461, 916–922.

Pichler, B.J., Kolb, A., Nägele, T., Schlemmer, H.P., 2010. PET/MRI: paving the way for the next generation of clinical multimodality imaging applications. J. Nucl. Med. 51, 333–336.

Salakhutdinov, R., Hinton, G.E., 2009. Deep Boltzmann machines. Proceedings of the International Conference on Artificial Intelligence and, Statistics, pp. 448–455.

Salakhutdinov, R., Hinton, G., 2012. An efficient learning procedure for deep Boltzmann machines. Neural Comput. 24, 1967–2006.

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

Shin, H.C., Orton, M.R., Collins, D.J., Doran, S.J., Leach, M.O., 2013. Stacked autoencoders for unsupervised feature learning and multiple organ detection in a pilot study using 4D patient data. IEEE Trans. Pattern Anal. Mach. Intell. 35, 1930–1943.

Singh, V., Chertkow, H., Lerch, J.P., Evans, A.C., Dorr, A.E., Kabani, N.J., 2006. Spatial patterns of cortical thinning in mild cognitive impairment and Alzheimer's disease. Brain 129, 2885–2893.

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, 87–97.

Srivastava, N., Salakhutdinov, R., 2012. Multimodal learning with deep Boltzmann machines. Advances in Neural Information Processing Systems, 25, pp. 2231–2239.

Suk, H.I., Shen, D., 2013. Deep learning-based feature representation for AD/MCI classification. Medical Image Computing and Computer-Assisted Intervention—MICCAI 2013. Lecture Notes in Computer Science, vol. 8150, pp. 583–590.

Suk, H.I., Wee, C.Y., Shen, D., 2013. Discriminative group sparse representation for mild cognitive impairment classification. Machine Learning in Medical Imaging. Lecture Notes in Computer Science, vol. 8184, pp. 131–138.

Tanaka, T., 1998. A theory of mean field approximation. Advances in Neural Information Processing Systems (NIPS). The MIT Press, pp. 351–360.

Tang, S., Fan, Y., Wu, G., Kim, M., Shen, D., 2009. RABBIT: rapid alignment of brains by building intermediate templates. Neuroimage 47, 1277–1287.

Visser, P.J., Verhey, F.R.J., Hofman, P.A.M., Scheltens, P., Jolles, J., 2002. Medial temporal lobe atrophy predicts Alzheimer's disease in patients with minor cognitive impairment. J. Neurol. Neurosurg. Psychiatry 72, 491–497.

Walhovd, K., Fjell, A., Brewer, J., McEvoy, L., Fennema-Notestine Jr., C., D.H., Jennings, R., Karow, D., Dale, A., the Alzheimer's Disease Neuroimaging Initiative, 2010. Combining MR imaging, positron-emission tomography, and CSF biomarkers in the diagnosis and prognosis of Alzheimer disease. Am. J. Neuroradiol. 31, 347–354.

Wang, Y., Nie, J., Yap, P.T., Li, G., Shi, F., Geng, X., Guo, L., Shen, D., 2014. Knowledge-guided robust MRI brain extraction for diverse large-scale neuroimaging studies on humans and non-human primates. PLoS One 9, e77810.

Ward, A., Arrighi, H.M., Michels, S., Cedarbaum, J.M., 2012. Mild cognitive impairment: disparity of incidence and prevalence estimates. Alzheimers Dement. 8, 14–21.

Wee, C.Y., Yap, P.T., Li, W., Denny, K., Browndyke, J.N., Potter, G.G., Welsh-Bohmer, K.A., Wang, L., Shen, D., 2011. Enriched white matter connectivity networks for accurate identification of MCI patients. Neuroimage 54, 1812–1822.

Wee, C.Y., Yap, P.T., Zhang, D., Denny, K., Browndyke, J.N., Potter, G.G., Welsh-Bohmer, K.A., Wang, L., Shen, D., 2012. Identification of MCI individuals using structural and functional connectivity networks. Neuroimage 59, 2045–2056.

Westman, E., Muehlboeck, J.S., Simmons, A., 2012. Combining MRI and CSF measures for classification of Alzheimer's disease and prediction of mild cognitive impairment conversion. Neuroimage 62, 229–238.

Xue, Z., Shen, D., Davatzikos, C., 2006. Statistical representation of high-dimensional deformation fields with application to statistically constrained 3D warping. Med. Image Anal. 10, 740–751.

Yang, J., Shen, D., Davatzikos, C., Verma, R., 2008. Diffusion tensor image registration using tensor geometry and orientation features. In: Metaxas, D., Axel, L., Fichtinger, G., SzŽkely, G. (Eds.), Medical Image Computing and Computer-assisted Intervention — MICCAI 2008. Springer, Berlin Heidelberg, pp. 905–913.

Yuan, L., Wang, Y., Thompson, P.M., Narayan, V.A., Ye, J., 2012. Multi-source feature learning for joint analysis of incomplete multiple heterogeneous neuroimaging data. Neuroimage 61, 622–632.

Zhang, D., Shen, D., 2012. Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in Alzheimer's disease. Neuroimage 59, 895–907.

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, 45–57.

Zhang, D., Wang, Y., Zhou, L., Yuan, H., Shen, D., 2011. Multimodal classification of Alzheimer's disease and mild cognitive impairment. Neuroimage 55, 856–867.

Zhang, D., Shen, D., Alzheimer's Disease Neuroimaging, I., 2012. Predicting future clinical changes of MCI patients using longitudinal and multimodal biomarkers. PLoS One 7, e33182.

Zhou, L., Wang, Y., Li, Y., Yap, P.T., Shen, D., the Alzheimer's Disease Neuroimaging, I., 2011. Hierarchical anatomical brain networks for MCI prediction: revisiting volumetric measures. PLoS One 6, e21935.