NeuroImage 208 (2020) 116459

Contents lists available at ScienceDirect

NeuroImage

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

A multi-model deep convolutional neural network for automatic hippocampus segmentation and classification in Alzheimer's disease

Manhua Liu [a,b,*], Fan Li [b], Hao Yan [b], Kundong Wang [b], Yixin Ma [b], Alzheimer's Disease Neuroimaging Initiative[1], Li Shen [c], Mingqing Xu [d,e,**]

[a] MoE Key Lab of Artificial Intelligence, Artificial Intelligence Institute, Shanghai Jiao Tong University, Shanghai, China
[b] Department of Instrument Science and Engineering, School of EIEE, Shanghai Jiao Tong University, Shanghai, China
[c] Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, 19104, USA
[d] Bio-X Institutes, Key Laboratory for the Genetics of Developmental and Neuropsychiatric Disorders, Shanghai Jiao Tong University, Shanghai, China
[e] Shanghai Key Laboratory of Psychotic Disorders, Shanghai Mental Health Center, School of Medicine, Shanghai Jiao Tong University, Shanghai, China

ARTICLE INFO

Keywords:
Alzheimer's disease
Hippocampus
Magnetic resonance imaging
Convolutional neural network
Image classification

ABSTRACT

Alzheimer's disease (AD) is a progressive and irreversible brain degenerative disorder. Mild cognitive impairment (MCI) is a clinical precursor of AD. Although some treatments can delay its progression, no effective cures are available for AD. Accurate early-stage diagnosis of AD is vital for the prevention and intervention of the disease progression. Hippocampus is one of the first affected brain regions in AD. To help AD diagnosis, the shape and volume of the hippocampus are often measured using structural magnetic resonance imaging (MRI). However, these features encode limited information and may suffer from segmentation errors. Additionally, the extraction of these features is independent of the classification model, which could result in sub-optimal performance. In this study, we propose a multi-model deep learning framework based on convolutional neural network (CNN) for joint automatic hippocampal segmentation and AD classification using structural MRI data. Firstly, a multi-task deep CNN model is constructed for jointly learning hippocampal segmentation and disease classification. Then, we construct a 3D Densely Connected Convolutional Networks (3D DenseNet) to learn features of the 3D patches extracted based on the hippocampal segmentation results for the classification task. Finally, the learned features from the multi-task CNN and DenseNet models are combined to classify disease status. Our method is evaluated on the baseline T1-weighted structural MRI data collected from 97 AD, 233 MCI, 119 Normal Control (NC) subjects in the Alzheimer's Disease Neuroimaging Initiative (ADNI) database. The proposed method achieves a dice similarity coefficient of 87.0% for hippocampal segmentation. In addition, the proposed method achieves an accuracy of 88.9% and an AUC (area under the ROC curve) of 92.5% for classifying AD vs. NC subjects, and an accuracy of 76.2% and an AUC of 77.5% for classifying MCI vs. NC subjects. Our empirical study also demonstrates that the proposed multi-model method outperforms the single-model methods and several other competing methods.

1. Introduction

Alzheimer's disease (AD) is a progressive and irreversible brain degenerative disorder characterized by memory loss and cognitive impairment. At present, there are around 90 million people diagnosed with AD, and it is estimated that the number of AD patients will reach 300 million by 2050 (Zhan et al., 2015; Zhu et al., 2015). To date, no effective drug treatments are available to cure AD, while the existing AD medicines can only ease symptoms or slow down its progression. Thus, the detection of AD at its early or prodromal stage is important for the

* Corresponding author. Shanghai Jiao Tong University, 800 Dongchuan Road, Shanghai, 200240, China.
** Corresponding author. Bio-X Institutes, Key Laboratory for the Genetics of Developmental and Neuropsychiatric Disorders, Shanghai Jiao Tong University, 1954 Huashan Road, Shanghai 200030, China.
E-mail addresses: mhliu@sjtu.edu.cn (M. Liu), mingqingxu@sjtu.edu.cn (M. Xu).

[1] 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.2019.116459
Received 29 August 2018; Received in revised form 9 December 2019; Accepted 10 December 2019
Available online 16 December 2019

prevention and intervention of its progression. For example, mild cognitive impairment (MCI) is a prodromal stage of AD (Silveira M, 2015). In the past few decades, neuroimaging technologies have been widely used to discover the relevant biomarkers in the human brain for AD and MCI diagnosis. Magnetic resonance imaging (MRI) is a non-invasive imaging technology that can produce detailed 3D anatomical images of internal body structures such as the brain, and has been widely used to help us understand anatomical and functional brain changes related to AD (Herrup, 2011; Jack et al., 2008; Jr et al., 2011). In particular, structural MRI scans provide detailed information about the anatomical structures of the brain, which can help detect and measure brain atrophy patterns in AD.

In recent years, many machine learning models have been employed to analyze the MRI data for identifying biomarkers and deciphering the disease etiology (Cao et al., 2017, 2018; Cheng and Wang, 2017; Hosseini-Asl et al., 2016; Jr et al., 2011; Liu et al., 2018a, b; Ortiz et al., 2016; Suk et al., 2015; Zhang et al., 2016, 2017). One common approach is to partition brain MRI scans into multiple regions of interest (ROIs) and extract ROI features such as volumes and shapes for AD diagnosis (Herrup, 2011; Suk et al., 2015; Zhang et al., 2011). For example, Zhang et al. proposed to extract the volumetric features from 93 ROIs and train a support vector machine (SVM) classifier for AD diagnosis (Zhang et al., 2011). A multi-task method was proposed to select the most discriminative features from 93 ROIs for multimodal classification of AD/MCI (Ye et al., 2016). An approach was proposed to combine the marginal fisher analysis with norm-based multi-kernel learning to achieve the sparsity of ROIs, which could simultaneously select a subset of relevant brain regions and learn a dimensionality transformation (Cao et al., 2017). A deep network with stacked autoencoder was employed on the ROI features to extract the latent high-level features for improving the performance of disease classification using neuroimaging data (Suk et al., 2015). A set of anatomical landmarks were detected from MRI to extract the morphological features for AD diagnosis (Liu et al., 2017, 2018b; Zhang et al., 2016, 2017).

Recently, deep learning networks, including convolutional neural networks (CNNs), have been widely used in image classification and computer vision (Chen et al., 2016; Liu et al., 2014, 2017, 2018b; Ng et al., 2015; Simonyan and Zisserman, 2014; Wang et al., 2017). The deep 3D CNNs were used to extract the features of 3D medical images for classification (Hosseini-Asl et al., 2016). A multi-task deep learning (MDL) method was proposed for joint hippocampal segmentation and clinical score regression using MRI scans (Cao et al., 2018). Given the very high dimensionality of the brain MRI data, it requires huge computational resources and a large dataset to train a deeper CNN with robustness. Since the MRI datasets used for AD diagnosis are typically very small compared with the datasets used in computer vision, it remains a major challenge to train a deeper CNN model with a large number of parameters to be learned (Gray et al., 2011). Recently, a classification scheme with an ensemble of deep learning architectures was proposed for early diagnosis of AD (Ortiz et al., 2016). In this approach, the gray matter (GM) image of each brain was split into 3D patches according to ROIs defined by the Automated Anatomical Labeling (AAL) atlas, and different deep belief networks were trained with the patches of different ROIs and followed by an ensemble with a voting scheme for final prediction. Meanwhile, a landmark-based deep feature learning (LDFL) framework for AD diagnosis was proposed to use a CNN model for extraction of the patch-based representation from a set of anatomical landmarks (Liu et al., 2017, 2018b).

Among all brain ROIs, the hippocampus is one of the first affected regions in AD, and it is an important anatomical region in the AD etiology. In many MRI studies for AD diagnosis, people proposed to compute the shape and volume features from bilateral hippocampi (Amoroso et al., 2018; Beg et al., 2013; Chupin et al., 2009; Gerardin et al., 2009; Ho et al., 2011; Leung et al., 2010; Lindberg et al., 2012; Platero and Tobar, 2016; Shen et al., 2012; Xin et al., 2012). Chupin el al. Proposed a fully automatic method for hippocampal segmentation using probabilistic and anatomical priors and the hippocampal volumes were computed for AD diagnosis (Chupin et al., 2009). A fast multiple-atlas segmentation method was proposed to measure hippocampal volume for discriminating AD/MCI patients from NC using MRI scans (Platero and Tobar, 2016). However, the volumetric analysis only assesses global changes of the hippocampus and is limited by the variety of hippocampal volumes among different individuals. To alleviate this problem, shape analysis was often used to capture detailed hippocampal morphology for AD diagnosis (Beg et al., 2013; Gerardin et al., 2009; Lindberg et al., 2012). It could unveil local atrophy of the hippocampus. In (Lindberg et al., 2012), hippocampal shape analysis was performed by modeling the surface of 3D hippocampal structure using the spherical harmonic shape description (SPHARM).

There remain several problems with the existing hippocampal analysis methods. First, both hippocampal volumetric and shape analyses depend on accurate segmentation of the hippocampus. But it is still a challenging task to accurately segment the hippocampus due to its irregular shape and blurred boundary in MRI scans. Second, the hand-crafted features of hippocampal volumes and shapes may not be optimal for the subsequent analysis, which may affect classification performance in disease diagnosis. Third, using the hippocampus alone may not be sufficient for discriminating MCI from NC subjects. Other regions adjacent to the hippocampus such as the parahippocampus and amygdala are also affected in early stage of AD. Finally, the visual and texture features of MRI scans derived from the hippocampal region can be of great help for AD diagnosis (Ahmed et al., 2015; Xin et al., 2012).

To address the above problems in the computer-aided AD diagnosis, we propose a new deep learning framework with a combination of multiple deep CNN models for simultaneous hippocampal segmentation and disease classification using MRI data. In our framework, we first construct a multi-task deep CNN model for jointly learning hippocampal segmentation and disease classification. Then, a 3D Densely Connected Convolutional Network (3D DenseNet) model is developed with the inputs of 3D patches extracted based on the hippocampal segmentation results to learn rich features. Finally, the learned features by the multi-task deep CNN model and the DenseNet model are combined with a fully connected layer to yield a final classification of disease status. The proposed multi-model framework will be shown to outperform each single model method. Our method is evaluated for both hippocampal segmentation and disease classification, where we use the baseline T1-weighted structural MRI data from the ADNI database including 97 AD, 233 MCI, and 119 NC subjects. In addition, we also test our method on an additional dataset of 135 subjects from the ADNI MRI cohort with the EADC-ADNI Harmonized Protocol (HarP) for manual hippocampal segmentation (Boccardi et al., 2015).

2. Materials and methods

As shown in Fig. 1, our proposed deep learning framework consists of two deep learning models. One model is a multi-task deep CNNs for jointly learning hippocampus segmentation and disease classification, which generates a binary segmentation mask of the hippocampus and learns features for disease classification. However, the learned features by the multi-task model are not sufficient for accurate disease classification. A 3D patch covering the hippocampus is extracted based on the centroid of the segmentation mask and input into a 3D DenseNet model that learns more relevant features for disease classification. Finally, a fully connected layer and a softmax layer are appended to combine the learned features from these models for final disease classification. In this study, two classes are considered in the classification task. The input of this framework is a large image patch covering the hippocampus. The outputs are the hippocampus mask as well as the prediction of disease status.

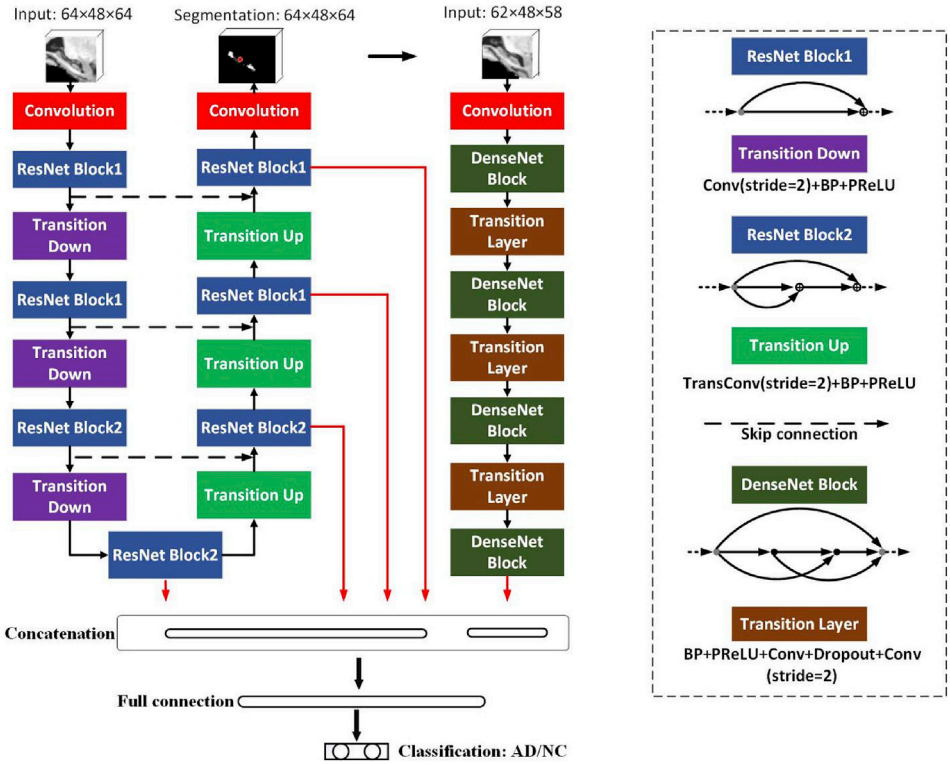

Fig. 1. Flowchart of the proposed deep learning framework for hippocampal segmentation and disease status classification with the integration of two deep CNN models. BN denotes Batch Normalization; Conv denotes Convolution; TransConv denotes Transpose Convolution; BP denotes Back Propagation; and PReLU denotes Parametric Rectified Linear Unit activation. The outputs are hippocampus mask and disease status. The unit of input patch is voxel of $1 \times 1 \times 1$ mm^3.

2.1. Data source and preprocessing

The data were obtained from the ADNI database, which is publicly available on the website (http://www.loni.ucla.edu/ADNI). The ADNI was launched in 2003 by the National Institute on Aging (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the Food and Drug Administration (FDA), private pharmaceutical companies and non-profit organizations, as a $60 million, 5-year public-private partnership. The primary goal of the ADNI was to test whether serial MRI, Positron Emission Tomography (PET), other biological markers, and clinical and neuropsychological assessments could be combined to measure the progression of MCI and early AD. The subjects were recruited from over 50 sites across the U.S. and Canada, gave written informed consent at the time of enrollment for imaging and genetic sample collection and completed questionnaires approved by each participating site's Institutional Review Board (IRB).

We randomly selected 449 participants, including 97 AD, 233 MCI, and 119 NC subjects, from whom the baseline T1-weighted MR imaging data and the pre-processed images are available in ADNI. Table 1 shows the demographic and clinical information of these subjects, where the CDR denotes the Clinical Dementia Rating and MMSE denotes the Mini-Mental State Examination. We used the MR images acquired with 1.5T scanners according to the ADNI acquisition protocol (Jack et al., 2008). More detailed information about the image acquisition procedures is available on the ADNI website. All MR images were resampled to $256 \times 256 \times 256$ voxels of size $1 \times 1 \times 1$ mm^3. They were skull-stripped and cerebellum-removed after correction of intensity inhomogeneity using a nonparametric non-uniform intensity normalization (N3) algorithm (Sled et al., 1998; Wang et al., 2011). With 12 degrees of freedom and the default parameters, we performed the affine registration to align all MR images to a template image using FMRIB Software Library (FSL) 5.0 from https://fsl.fmrib.ox.ac.uk/.

Table 1
Demographic and clinical information (mean ± standard deviation) of the studied ADNI subjects. AD: Alzheimer's disease; MCI: mild cognitive impairment; NC: normal control; M: male; F: female; MMSE: Mini-Mental State Examination; CDR: Clinical Dementia Rating.

Diagnosis	Age	Gender (M/F)	MMSE	Education(year)	CDR
AD	75.9 ± 6.8	48/49	23.2 ± 1.8	15.0 ± 3.0	0.8 ± 0.3
MCI	75.2 ± 7.3	145/88	26.9 ± 1.8	15.7 ± 2.6	0.5 ± 0
NC	75.9 ± 5.0	59/60	29.2 ± 1.0	15.6 ± 2.7	0 ± 0

Since directly labeling the hippocampus from scratch by radiologists would be time-consuming, FIRST from FSL (the FMRIB Software Library) (Morey et al., 2009) was used as a starting tool to obtain a rough segmentation. Three radiologists worked together to check the initial FIRST segmentation for each subject and reached a consensus after a discussion to produce a final segmentation by manual editing. Fig. 2 shows several examples of hippocampal segmentation results before and after correction together with their hippocampal image patches. Table 2 shows quantitative measures of hippocampal volumes (i.e., mean, standard deviation, and range) before and after manual correction for different groups of subjects. We can see that the mean and standard deviation of hippocampal volume are reduced after correction. Fig. 3 shows the scatterplots of left and right hippocampal volumes for the studied AD, MCI and NC subjects.

The hippocampus is a small region located in the medial temporal lobe of the human brain. The number of voxels in the hippocampus is much smaller than those of background, which would result in a severe class-imbalance problem. After image preprocessing and affine registration, we crop each MR image into a 3D image patch with a bounding cube for each hippocampus. The bounding cubes are defined as 3D axes to

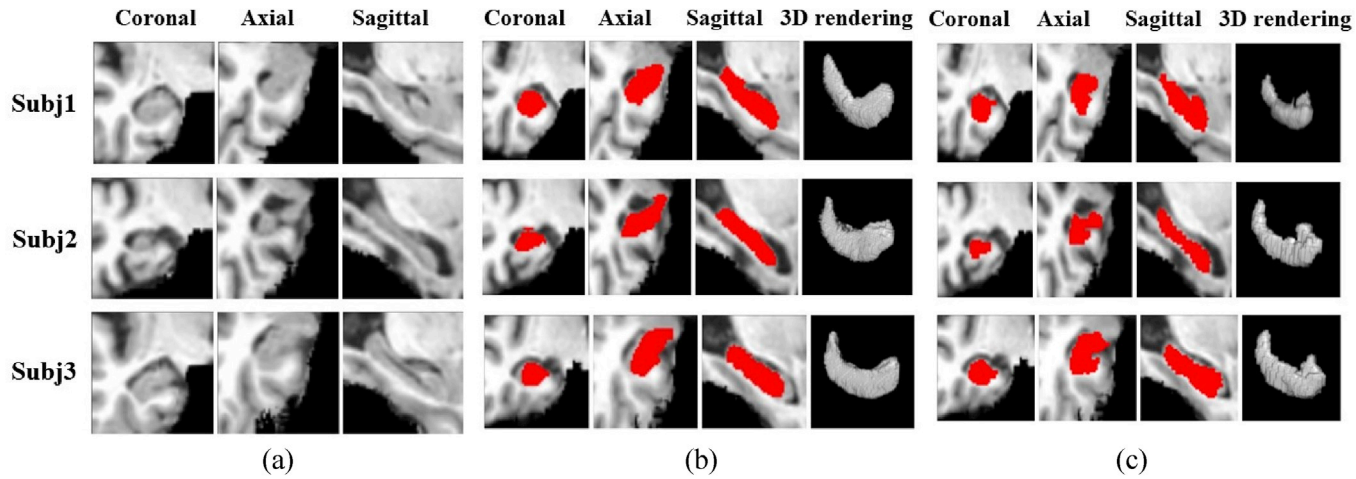

Fig. 2. Examples of hippocampal image patches: (a) patches without segmentation labels, (b) patches overlaid with segmentation labels before manual correction, and (c) patches overlaid with segmentation labels after manual correction. Subj1, Subj2, and Subj3 are three example subjects randomly selected from the AD, MCI and NC groups respectively.

Table 2
Hippocampal volume (mm^3) before and after corrections denoted with Mean ± Standard Deviation (Range). AD: Alzheimer's disease; MCI: mild cognitive impairment; NC: normal control.

Groups	AD		MCI		NC	
	Left	Right	Left	Right	Left	Right
Before correction	3390 ± 753 (1263–5223)	3467 ± 836 (1103–5553)	3670 ± 840 (1080–5742)	3759 ± 774 (1051–5891)	4377 ± 677 (2558–6024)	4442 ± 596 (2582–6018)
After correction	3203 ± 592 (2023–5029)	3310 ± 674 (1928–5283)	3471 ± 562 (1963–5480)	3608 ± 574 (1911–5167)	4139 ± 524 (2735–5462)	4221 ± 536 (2971–5648)

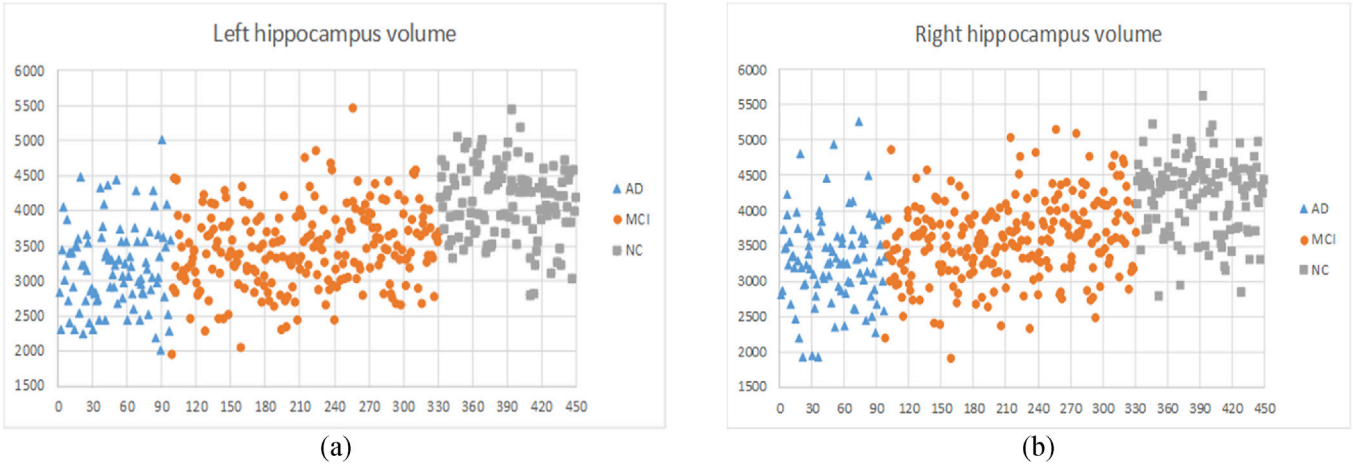

Fig. 3. The scatterplots of (a) left and (b) right hippocampus volumes from AD, MCI and NC subjects.

extract 3D patches from MR images. The size of the bounding cube has some effects on hippocampus segmentation. Using a large bounding cube not only increases the computation cost but also may result in the class-imbalance problem. However, using a small bounding cube may lose important context information for hippocampal segmentation. To balance the tradeoff, we empirically set the optimal size of the bounding cube to $64 \times 48 \times 64$ voxels. Based on the patches, we construct the deep learning model for hippocampal segmentation and disease classification in the following sections.

2.2. Multi-task deep CNN for joint hippocampal segmentation and disease classification

Different from the conventional methods that perform hippocampal segmentation and disease classification separately, we propose a multi-task CNN model for jointly learning hippocampal segmentation and disease classification. CNN is a special type of multi-layer neural network widely used in image classification and object detection (Krizhevsky et al., 2012; Lécun et al., 1998; Zhu et al., 2015). A volumetric and fully

CNN named "V-Net" was proposed for the prostate segmentation in MRIs (Milletari et al., 2016). Motivated by the success of V-Net in prostate segmentation, we exploited a similar network structure to develop a multi-task deep CNN model for joint hippocampal segmentation and disease classification.

A deep CNN has been formulated to learn residual functions at the convolutional stages to achieve fast convergence. We define two residual blocks as "ResNet Block1" and "ResNet Block2", consisting of 3D convolutional, batch normalization (BN), Parametric Rectified Linear Unit (PReLU) activation and dropout layers as shown in Fig. 4. In ResNet Block1, a residual function is learned by a short connection: the input is added to the output of the second convolutional layer. ResNet Block2 consists of two convolutional layers and the input of each block is added to both the outputs of the second and third convolutional layers to learn the residual function. The kernels are trained with supervision from the batches of MRI data. Small kernels have fewer numbers of parameters to train for fast inference. Large kernels learn much more complex patterns and have a stronger expressive power. This effect can be achieved by stacking more convolutional layers of small kernels. Therefore, the kernel size is set to $3 \times 3 \times 3$ for all convolutions. The learned filters are convolved with the input image followed by a non-linear PReLU activation and a feature map is generated for each filter.

As illustrated in Fig. 1, this multi-task deep CNN model consists of two parts: shown on the left is a procedure of compressing feature maps, and shown on the right is a procedure of decompressing feature maps to their original patch size. The left compression part has two blocks of ResNet Block1 and two blocks of ResNet Block2 followed by down-sampling for each block, while the right decompression part comprises of two blocks of ResNet Block1 and one block of ResNet Block2 followed by up-sampling for each block. Table 3 lists the details of relevant parameters for the multi-task deep network model. In the compression part, down-sampling is used to reduce the size of feature maps and increase the receptive field of features in the subsequent layers. It is implemented by convolution with the kernels of size $2 \times 2 \times 2$ and stride 2. In the decompression part, the spatial support of the lower resolution feature maps is expanded to extract features and assemble the necessary information to generate a volumetric segmentation mask. The up-sampling by de-convolution is performed with the kernels of $2 \times 2 \times 2$ and stride 2. Then, a convolutional layer with $1 \times 1 \times 1$ kernel and stride 1 is performed to generate the outputs with the same size as the input patch,

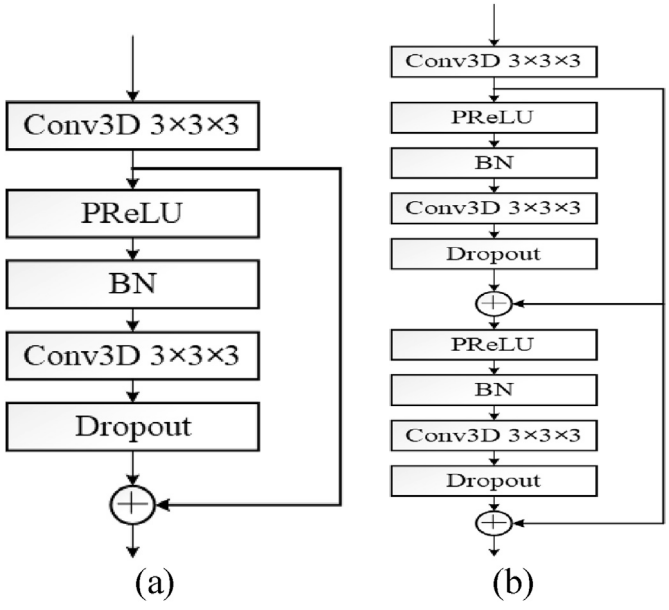

Fig. 4. The network architecture of (a) ResNet Block1 and (b) ResNet Block2, consisting of 3D convolution, PReLU, BN, and dropout layers.

Table 3
The network architecture of the multi-task deep model (the layers with * are used for classification and the unit of input patch is $1 \times 1 \times 1$ mm^3).

Layers	Output Size	Filter size, stride, number
Input Layer	$64 \times 48 \times 64$	–
Convolution	$16, 64 \times 48 \times 64$	$3 \times 3 \times 3, 1, 16$
ResNet Block1	$16, 64 \times 48 \times 64$	$\begin{bmatrix} 3 \times 3 \times 3, & 1, & 16 \\ 3 \times 3 \times 3, & 1, & 16 \end{bmatrix}$
Down-sampling	$32, 32 \times 24 \times 32$	$2 \times 2 \times 2, 2, 32$
ResNet Block1	$32, 32 \times 24 \times 32$	$\begin{bmatrix} 3 \times 3 \times 3, & 1, & 32 \\ 3 \times 3 \times 3, & 1, & 32 \end{bmatrix}$
Down-sampling	$64, 16 \times 12 \times 16$	$2 \times 2 \times 2, 2, 64$
ResNet Block2	$64, 16 \times 12 \times 16$	$\begin{bmatrix} 3 \times 3 \times 3, & 1, & 64 \\ 3 \times 3 \times 3, & 1, & 64 \times 2 \end{bmatrix}$
Down-sampling	$128, 8 \times 6 \times 8$	$2 \times 2 \times 2, 2, 128$
ResNet Block2 (*)	$128, 8 \times 6 \times 8$	$\begin{bmatrix} 3 \times 3 \times 3, & 1, & 128 \\ 3 \times 3 \times 3, & 1, & 128 \times 2 \end{bmatrix}$
Up-sampling	$64, 16 \times 12 \times 16$	$2 \times 2 \times 2, 2, 64$
ResNet Block2 (*)	$64, 16 \times 12 \times 16$	$\begin{bmatrix} 3 \times 3 \times 3, & 1, & 64 \\ 3 \times 3 \times 3, & 1, & 64 \times 2 \end{bmatrix}$
Up-sampling	$32, 32 \times 24 \times 32$	$2 \times 2 \times 2, 2, 32$
ResNet Block1 (*)	$32, 32 \times 24 \times 32$	$\begin{bmatrix} 3 \times 3 \times 3, & 1, 32 \\ 3 \times 3 \times 3, & 1, 32 \end{bmatrix}$
Up-sampling	$16, 64 \times 48 \times 64$	$2 \times 2 \times 2, 2, 16$
ResNet Block1 (*)	$16, 64 \times 48 \times 64$	$\begin{bmatrix} 3 \times 3 \times 3, & 1, & 16 \\ 3 \times 3 \times 3, & 1, & 16 \end{bmatrix}$
Convolution	$64 \times 48 \times 64$	$1 \times 1 \times 1, 1, 1$

which are converted to probabilistic segmentation of the hippocampal regions by applying voxel-wise softmax. Finally, the threshold is set to 0.5 to convert the probabilistic output to a binary mask.

As shown in Fig. 1, the deep network outputs the hippocampal mask and disease status prediction. For hippocampal segmentation of subject m, the optimization objective is to minimize the Dice loss function which evaluates the capability of our model to segment hippocampal voxels from the background:

$$L_s^m = 1 - \frac{2\sum_{i=1}^{N} p_i q_i + \varepsilon}{\sum_{i=1}^{N} p_i^2 + \sum_{i=1}^{N} q_i^2 + \varepsilon} \tag{1}$$

where N is the number of total voxels on the segmentation output; p_i and q_i are the predicted segmentation result and the ground truth label of voxel i, respectively; ε is a small value to prevent denominator from being zero. The Dice loss function can deal with situations where a strong imbalance exists between the numbers of foreground and background voxels (Sled et al., 1998). As to classification, a fully connected layer is stacked to concatenate the outputs of the compression part and each layer of the decompression part, which integrates information from two sources to enhance classification accuracy. The loss function for classification of subject m is the categorical cross-entropy loss to evaluate the difference between the predicted label and the ground truth label as follows:

$$L_C^m = -\left(\widehat{y_m}\log y_m + (1 - \widehat{y_m})\log(1 - y_m)\right) \tag{2}$$

The loss function for the multi-task deep CNN model is a weighted sum of the segmentation loss and classification loss, calculated as follows:

$$L_M = \alpha \cdot L_s + (1-\alpha) \cdot L_C = \frac{1}{M}\prod_{m=1}^{M}\left[\alpha \cdot L_s^m - (1-\alpha) \cdot \left(\widehat{y_m}\log y_m + (1 - \widehat{y_m})\log(1 - y_m)\right)\right] \tag{3}$$

where M is the total number of subjects; $\widehat{y_m}$ and y_m are the ground truth label and the predicted label for subject m. The parameter $\alpha \in [0, \ 1]$ is a weight to adjust for the losses in training the hippocampal segmentation and disease classification. During the training procedure of the multi-task deep CNN model, the segmentation task is more important than classification at the early stage while the classification task takes effect at the late stage when the segmentation objective function value converges. In

our implementation, the value of α evolves throughout the training process. During the initial warm-up phase, α is set to 1 to emphasize on the segmentation task. Then it changes to 0.5 for multi-task training. In the final phase, α is set to 0 to focus on the classification task. The joint optimization of the multi-task network model is performed with the Adam method, and a backpropagation algorithm is used to calculate the network gradients.

2.3. The deep 3D DenseNet model

Traditionally, the shape and volume features are calculated from the hippocampal mask for AD diagnosis. Different from that, we propose to construct a deep 3D DenseNet model based on the hippocampal region to learn the discriminative features for AD diagnosis. Although the features learned by the multi-task deep CNN model can capture some important information for disease classification, they are still not sufficient for accurate diagnosis of the disease. Therefore, a 3D patch of a fixed size centered on the centroid of a segmentation mask is extracted to build the proposed CNN model. To enhance the representation power, CNN becomes increasingly deep to learn features. However, this may cause information loss when the input passes through many layers to reach the end of network. DenseNet is proposed to connect each layer to every other layer in a feed-forward fashion, which increases direct connections between the low and high level layers (Huang et al., 2016). Compared with traditional CNN, DenseNets have several advantages. First, they can alleviate the vanishing-gradient problem since there is a direct connection from the low to high-level layers. Second, feature propagation is strengthened to reuse the low-level features. Third, they can reduce the number of parameters.

Table 4 shows the structure and parameters of our proposed 3D DenseNet model, consisting of a convolutional layer, 4 dense blocks, 3 transition layers, a global average pooling layer, and a softmax layer. First, a convolutional layer was added to the input layer with stride 2, followed by dense blocks. Then, the dense block uses dense connectivity through which the l-th layer receives the feature maps of all preceding layers (Huang et al., 2016) as follows:

$$x_l = H_l([x_0,\ x_1, \ldots,\ x_{l-1}]) \tag{4}$$

where $[x_0,\ x_1,\ \ldots,\ x_{l-1}]$ is the concatenation of feature maps from all previous layers into a single tensor, and H_l denotes a composite nonlinear transformation function of four consecutive operations: BN, PReLU, $3 \times 3 \times 3$ convolution, and voxel-wise dropout. Dense block includes three dense layers with each layer consisting of one $1 \times 1 \times 1$ and one $3 \times 3 \times 3$ convolutional layers, two BN layers and two activation layers. Every dense layer receives the feature maps of all previous dense layers by shortcut connections. A transition layer, which consists of five consecutive operations: BN, PReLU, a $1 \times 1 \times 1$ convolution, voxel-wise dropout and a $3 \times 3 \times 3$ convolution with a $2 \times 2 \times 2$ stride, is set between two dense blocks for feature map reduction. Following the last dense block, an average pooling and a softmax classifier are appended to reduce feature dimension and classify disease status. The subject labels are used through back-propagation for updating the weights of DenseNet. All layers of DenseNet receive direct supervision from loss function through shortcut connections.

Table 4
The network architecture of our deep DenseNet Model (input patch unit is voxel of $1 \times 1 \times 1$ mm^3).

Layers	Output Size	Filter size, stride, number
Input Layer	$62 \times 48 \times 58$	–
Convolution	64, $31 \times 24 \times 29$	$3 \times 3 \times 3$, 2, 64, conv
Dense Block (1)	96, $31 \times 24 \times 29$	$\begin{bmatrix} 1\times1\times1,\ 1,\ 64,\ \text{conv} \\ 3\times3\times3,\ 1,\ 16,\ \text{conv} \end{bmatrix} \times 2$
Transition Layer	48, $16 \times 13 \times 15$	$\begin{bmatrix} 1\times1\times1,\ 1,\ 48,\ \text{conv} \\ 2\times2\times2,\ 2,\ 48,\ \text{conv} \end{bmatrix}$
Dense Block (2)	80, $16 \times 13 \times 15$	$\begin{bmatrix} 1\times1\times1,\ 1,\ 64,\ \text{conv} \\ 3\times3\times3,\ 1,\ 16,\ \text{conv} \end{bmatrix} \times 2$
Transition Layer	40, $9 \times 7 \times 8$	$\begin{bmatrix} 1\times1\times1,\ 1,\ 40,\ \text{conv} \\ 2\times2\times2,\ 2,\ 40,\ \text{conv} \end{bmatrix}$
Dense Block (3)	72, $9 \times 7 \times 8$	$\begin{bmatrix} 1\times1\times1,\ 1,\ 64,\ \text{conv} \\ 3\times3\times3,\ 1,\ 16,\ \text{conv} \end{bmatrix} \times 2$
Transition Layer	36, $6 \times 4 \times 6$	$\begin{bmatrix} 1\times1\times1,\ 1,\ 36,\ \text{conv} \\ 3\times3\times3,\ 1,\ 36,\ \text{conv} \end{bmatrix}$
Dense Block (4)	68, $6 \times 4 \times 6$	$\begin{bmatrix} 1\times1\times1,\ 1,\ 64,\ \text{conv} \\ 3\times3\times3,\ 1,\ 16,\ \text{conv} \end{bmatrix} \times 2$
Global Average Pooling	68, $1 \times 1 \times 1$	–
Softmax Layer	2	–

For training DenseNets, the initial weights for the whole network are uniform and Adam optimizer is adopted with a learning rate of 10–4, The network is stable after iteration of 120 epochs. The batch size is set to 64 and PReLU is used for each neuron of DenseNet. Dropout layers are used to alleviate the overfitting problem. The DenseNet models for the left and right hippocampi have the same structure but are trained individually with different patches for the classification task.

2.4. Final multi-model ensemble classification

The multi-task deep CNN model captures the multi-level features for joint hippocampal segmentation and disease classification, while the deep 3D DenseNet model learns the features from the image patches of the hippocampus for disease classification. To integrate these deep models, we further stack an extra fully connected layer above the concatenation of the learned features from deep models for disease classification. The DenseNet models and multi-task model are individually trained, and a fully connected layer followed by a softmax layer is finely tuned to make the final classification. They are implemented with Keras library in the framework of Tensorflow. We will show that the proposed multi-model deep network framework outperforms the single-model approaches.

3. Results

3.1. Datasets and implementation

The proposed deep learning framework was tested on the structural MRI data from the baseline visits of 449 ADNI participants consisting of 97 AD, 233 MCI, and 119 NC subjects. Both hippocampal segmentation and disease classification tasks were conducted for method evaluation. The classification task was tested to distinguish AD vs. NC and MCI vs. NC. We randomly divided the whole MRI data set into 5 folds and 5-fold cross-validation was used to train and test the proposed method. Each time, one fold of dataset was used for testing, while the other 4 folds were for training. For tuning the parameter of the iteration number in the deep learning model, we randomly select 10% of training data as the validation set while the remaining training data were used to train model. The iteration number was increased gradually one by one for parameter tuning and the optimal values were 110, 94, 96, 125 and 124 for the 5 different testing folds, showing relatively robust results. The testing set was not used for model training and parameter tuning but for general performance evaluation.

To evaluate the performance of hippocampal segmentation, we calculated four measures: Dice similarity coefficient (DSC), sensitivity (SEN_S), positive predicted value (PPV) and volume error (VE):

$$\mathrm{DSC} = \frac{2TP}{2TP + FP + FN},\ PPV = \frac{TP}{TP + FP},\ SEN\_\mathrm{S} = \frac{TP}{TP + FN},\ VE = \frac{2|V_a - V_m|}{V_a + V_m} \tag{5}$$

where TP denotes the true positives, i.e., the predicted hippocampal voxels inside the positive regions of ground-truth; FP denotes the false positives, i.e., the predicted hippocampal voxels outside the positive regions of ground-truth; FN denotes the false negatives as the predicted

background voxels inside the positive regions of ground-truth; V_a and V_m denote the hippocampal volumes by automatic and manual segmentation, respectively. For classification tasks, four performance measures were computed for evaluation, including classification accuracy (ACC), sensitivity (SEN_C), specificity (SPE), receiver operating characteristic (ROC) curve and area under ROC curve (AUC). ACC is the proportion of correctly classified subjects among the whole population. SEN_C is the proportion of AD/MCI patients correctly classified. SPE is the proportion of correctly classified NC subjects. The ROC curve is generated by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings on the prediction scores.

In our implementation, the trade-off parameter α in the loss function of our multi-task deep CNN framework evolved from 1 to 0 throughout the training process. We tested the segmentation and classification tasks by setting α to 1 or 0 for single task learning and setting α to 0.5 for multi-task learning. Shown in Tables 5 and 6 is the comparison of the segmentation and classification results by setting different α′s for single and multi-task learning as well as the proposed adaptive method. From the results, we can see the multi-task learning with fixed α performs better than single-task learning for classification. The proposed adaptive method performs better than the single-task and multi-task learnings with fixed α for both segmentation and classification. From Table 5, we can see better segmentation results on NC subjects than AD subjects even though the training set includes AD subjects. This may be caused by the hippocampal atrophy of AD and its variation among different subjects, which introduces difficulties in both automatic and manual segmentations.

3.2. Results on hippocampal segmentation

In this experiment, we tested the performance of the proposed framework on hippocampus segmentation in terms of DSC, PPV, SEN_S, and VE and compared the results with different parameter settings. We also performed empirical comparation with a few other competing methods. Fig. 5 (a), (b), (c) and (d) show the results of DSC, PPV, SEN_S, and VE, respectively, by the proposed segmentation method in 5 cross-validation trials using box-plots.

The inputs of our multi-task deep CNN framework were 3D image patches obtained by cropping MR images with a bounding cube for each hippocampus. The first experiment was to test the effects of various bounding cube sizes on the segmentation results. We gradually increased the bounding cube size from $48 \times 40 \times 48$, $64 \times 48 \times 64$ to $88 \times 80 \times 88$ to test hippocampal segmentation. The results in Table 7 demonstrate that the segmentation performance is improved by increasing the cube size from $48 \times 40 \times 48$ to $64 \times 48 \times 64$ and is degraded by further increasing the cube size to $88 \times 80 \times 88$. While a small boundary bounding cube might have limited context information, a large bounding cube could include unnecessary background voxels increasing both overfitting risk and computational burden. With this observation, we set the bounding cube to $64 \times 48 \times 64$ in the following experiments.

Second, we compared the segmentation results of our method with other three methods including FSL and two deep learning-based methods (Cao et al., 2018; Thyreau et al., 2018), denoted as "FSL", "Thyreau's method" and "Cao's method", respectively. FSL is a comprehensive library of analysis tools for fMRI, MRI and DTI brain images. The FIRST algorithm in the FSL library was used for hippocampal segmentation on our dataset. However, 27 subjects have large segmentation errors due to internal registration failure. In building the reference segmentation, these errors were manually corrected by the radiologists. For fair comparison, these subjects were removed for performance evaluation. Thyreau's method (Thyreau et al., 2018) was proposed to segment the bilateral hippocampi using a deep-learning appearance model by transferring algorithmic knowledge for large cohort processing. A wide and variable training set from multiple cohorts was used to train the segmentation model. Since the trained model is available online (https://github.com/bthyreau/hippodeep), we downloaded it to test on our data set for comparison. It is worth noting that the differences in the definitions of manual segmentations for training and testing data may bias the evaluation results. Cao's method (Cao et al., 2018) was proposed to construct a multi-task deep learning method for joint hippocampal segmentation and clinical score regression. For fair comparison, we rebuilt the multi-task deep network as in (Cao et al., 2017), which were trained and tested on our data set.

Fig. 6 shows the comparison of the segmentation results using different methods for three example subjects selected from the AD, MCI and NC groups of the test data, respectively. The original images, the segmented results of FSL, Thyreau's method, Cao's method, our method and ground truth are demonstrated in columns from left to right in Fig. 6. From the original images, we notice that it is not easy to distinguish the hippocampal regions from the adjacent tissues due to the small difference between their intensity values. The segmented hippocampal regions obtained by our method appear to be smoother and more accurate than those by other methods. In addition, Table 8 shows the comparison of segmentation results in terms of DSC, PPV, SEN_S and VE on the same training and testing data sets. Our method achieves the best performance compared to other methods. In particular, our method achieves the best PPV, which indicates that it can effectively detect hippocampal regions from the background. Incorporating this accurately segmented hippocampal region into our subsequent DenseNet can provide meaningful anatomical information to help improve disease classification.

Since the same hippocampal label definition was used for training and testing in our experiments, there may be a bias in the comparison of segmentation results because some competitive methods were not trained using this particular label definition. Of note, there are numerous protocols for labeling the hippocampus, e.g. http://www.hippocampal-protocol.net. To address this issue, furthermore, we tested our

Table 6
Comparison of classification results by setting different α for single-task and multi-task learning. AD: Alzheimer's disease; MCI: mild cognitive impairment; NC: normal control; ACC: classification accuracy; SEN_C: classification sensitivity; SPE: specificity; AUC: area under receiver operating characteristic curve.

Performance(%)	α = 0(Single-task)		α = 0.5(Multi-task)		Adaptive α (Multi-task)	
	AD vs. NC	MCI vs. NC	AD vs. NC	MCI vs. NC	AD vs. NC	MCI vs. NC
ACC	75.0	67.8	78.0	69.5	80.1	71.5
SEN_C	74.2	84.9	76.3	77.6	79.9	73.2
SPE	75.6	50.0	79.4	61.6	80.3	69.8
AUC	82.5	71.6	84.9	72.5	86.6	74.4

Table 5
Comparison of segmentation results by setting different α for single-task and multi-task learning. ALL: all studied subjects; AD: Alzheimer's disease; MCI: mild cognitive impairment; NC: normal control; DSC: Dice similarity coefficient; PPV: positive predicted value; SEN_S: segmentation sensitivity; VE: volume error.

Performance(%)	α = 1(Single-task)				α = 0.5(Multi-task)				Adaptive α (Multi-task)			
	ALL	AD	MCI	NC	ALL	AD	MCI	NC	ALL	AD	MCI	NC
DSC	86.7	86.3	86.7	87.2	81.8	81.5	81.7	82.3	87.0	86.4	86.9	87.5
PPV	84.7	84.4	84.9	86.5	80.4	79.6	80.5	81.0	84.6	83.6	84.0	85.1
SEN_S	88.9	88.5	88.9	88.4	84.5	84.2	84.8	84.6	89.7	89.5	90.0	89.7
VE	6.1	6.6	6.3	5.7	10.5	10.7	10.5	10.3	6.0	6.6	6.1	5.6

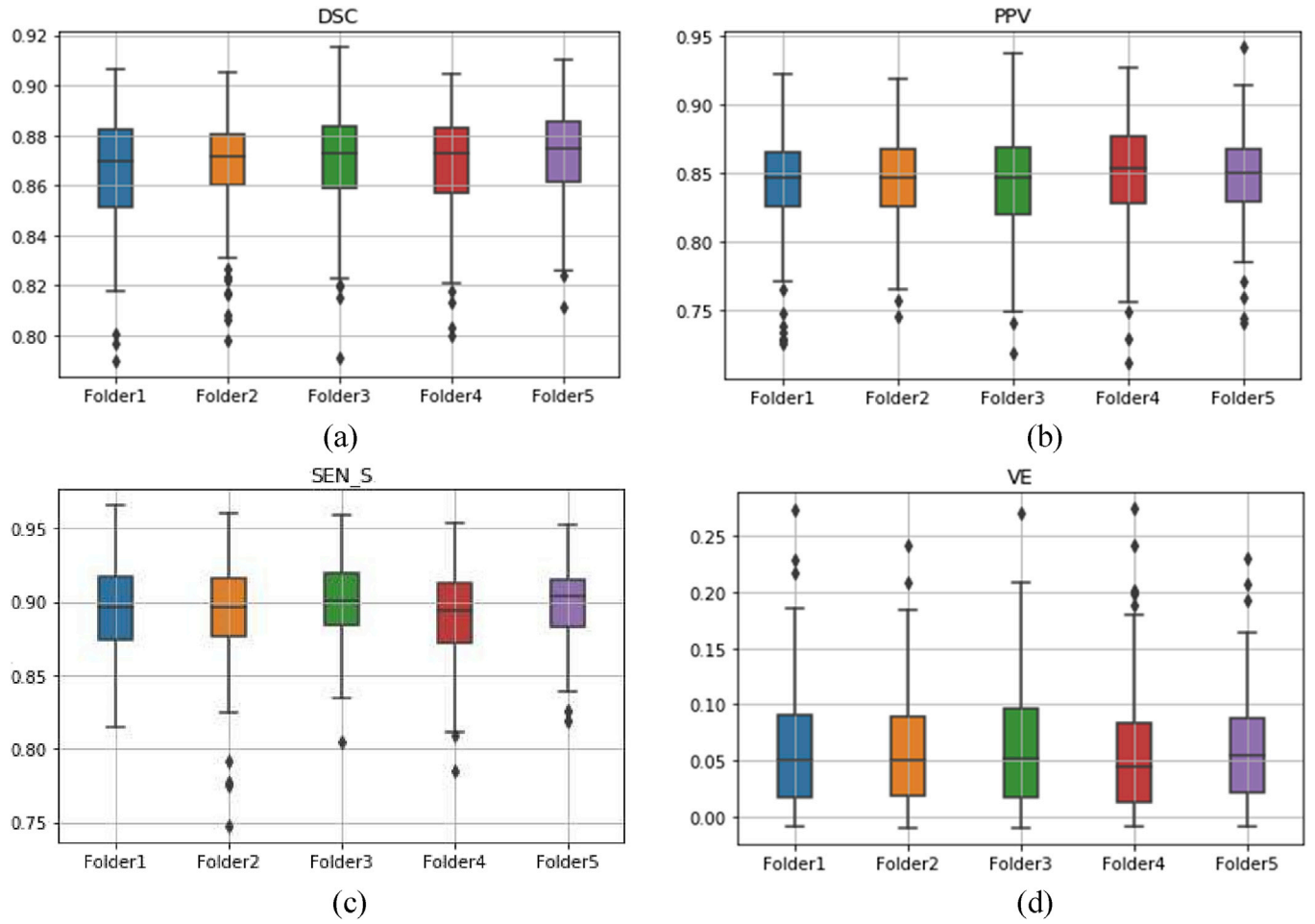

Fig. 5. The segmentation performances of (a) DSC, (b) PPV, (c) SEN_S and (d) VE in 5 folders.

Table 7
Comparison of segmentation results by different bounding cube sizes for multi-task deep CNN model (the patch unit is $1 \times 1 \times 1$ mm^3). DSC: Dice similarity coefficient; PPV: positive predicted value; SEN_S: segmentation sensitivity; VE: volume error.

Performance (%)	DSC	PPV	SEN_S	VE	Training time (hour)	Testing time (minute)
$48 \times 40 \times 48$	86.2	84.8	87.5	6.4	1.1	1.8
$64 \times 48 \times 64$	86.7	84.7	88.9	6.1	1.4	2.2
$88 \times 80 \times 88$	85.7	83.8	88.2	6.8	1.8	2.6

segmentation method on a new dataset with the EADC-ADNI Harmonized Protocol (HarP) for manual hippocampus labeling (Boccardi et al., 2015). This dataset consists of 135 subjects from the ADNI MRI cohort balanced by magnet field strength, age ranges, diagnosis, medial temporal atrophy, and scanner manufacturers. We used the MNC MR images and the nifti hippocampal labels released on their website (http://www.hippocampal-protocol.net/SOPs/index.php). First, we directly tested our trained segmentation model on this new dataset, denoted as "Our model on the new dataset". Second, we retrained and tested our segmentation model on this new dataset by 10-fold cross-validation strategy, denoted as "Retrained model on the new dataset". The comparison of results in Table 9 demonstrate that the performance is poor when our model is directly tested on the new dataset due to potentially different strategies for labeling the hippocampus. But the retrained model can work well on this new HarP dataset.

3.3. Results on disease classification

Our proposed method by integrating the multi-task deep CNN and DenseNet models was tested on disease classification. The first experiment was to test the classification performances of DenseNet when compared to other deep networks such as the popular LeNet (Lécun et al., 1998) and VGGNet (Simonyan and Zisserman, 2014). The LeNet network consists of 2 convolutional layers followed by 2 fully connected layers, while VGGNet consists of 13 convolutional layers and 3 fully connected layers. Thus, VGGNet has a deeper architecture with significantly more layers than LeNet. They were implemented with the released codes by replacing the 2D convolutions with 3D ones. Table 10 compares the classification results of AD vs. NC and MCI vs. NC by different deep networks. We could see that our DenseNet outperformed LeNet and VGGNet.

The second experiment was to test the effects of different patch sizes used to train DenseNet on the classification performance. The large patch covers more information near the hippocampus for classification. In our experiment, the patch size was gradually increased from $50 \times 40 \times 48$ to $68 \times 50 \times 64$. Table 11 shows the results of different patch sizes using DenseNet for classifying AD vs. NC and MCI vs. NC. We also compared the classification results obtained with the method combining multi-task deep CNN model and DenseNet, as shown in Table 12. The comparison demonstrates the classification results are improved by increasing the patch size from $50 \times 40 \times 48$ to $62 \times 48 \times 58$. The large patch requires more memory and computational time for both training and testing. The patch of $62 \times 48 \times 58$ was used to train DenseNets in our following experiments. It should be mentioned that this patch size is optimal based

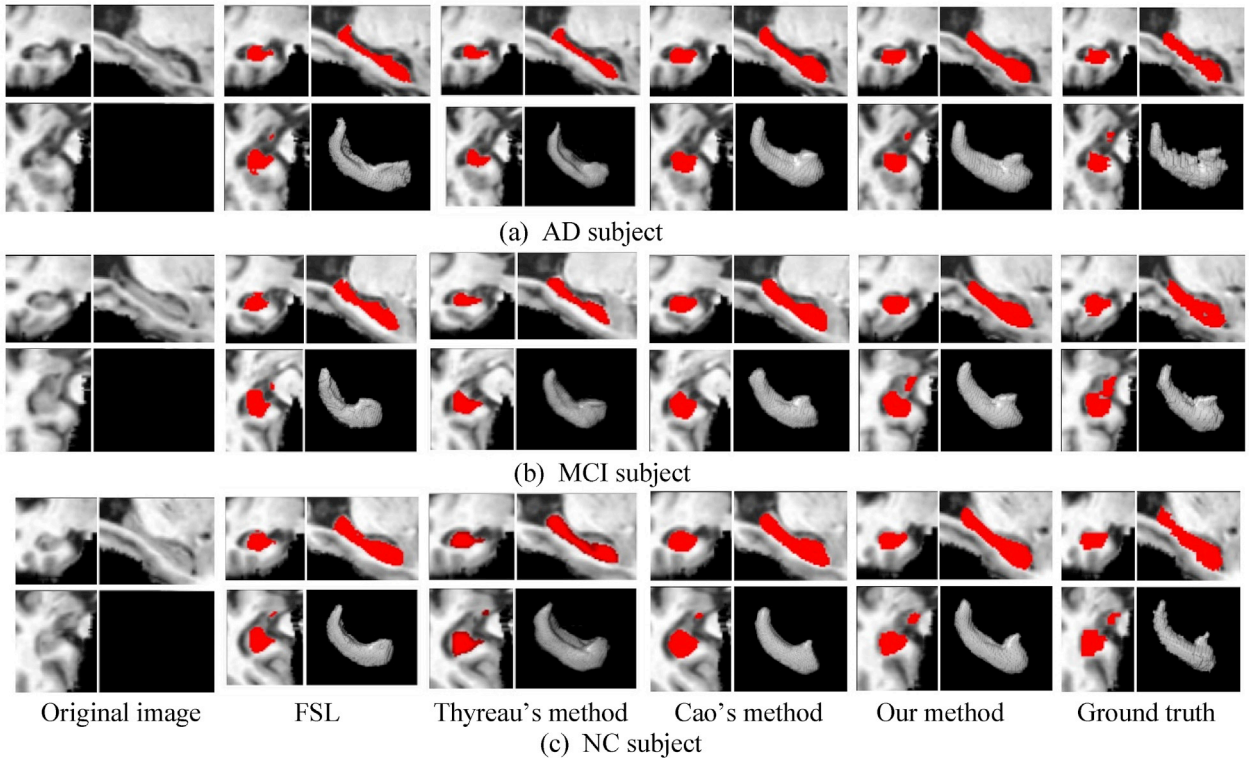

Fig. 6. Comparison of segmented hippocampal regions by different methods for three example subjects from the test data. Shown from top left, top right, bottom left to bottom right are coronal view, sagittal view, axial view and 3D rendering, respectively.

Table 8
Comparison of the segmentation results of the proposed method to other methods. DSC: Dice similarity coefficient; PPV: positive predicted value; SEN_S: segmentation sensitivity; VE: volume error.

Performance (%)	FSL	Thyreau's method	Cao's method	Proposed Method
DSC	72.2	73.8	84.6	87.0
PPV	70.4	68.8	83.7	84.6
SEN_S	74.6	80.2	85.7	89.7
VE	20.3	18.4	7.2	6.0

Table 9
Segmentation results in the deep model trained and tested on different datasets. DSC: Dice similarity coefficient; PPV: positive predicted value; SEN_S: segmentation sensitivity; VE: volume error.

Performance (%)	DSC	PPV	SEN_S	VE
Our model on our dataset	87.0	84.6	89.7	6.0
Our model on the new dataset	72.2	81.0	65.7	21.6
Retrained model on the new dataset	86.7	85.3	88.2	6.6

Table 10
Comparison of different deep networks for classification of AD vs. NC and MCI vs. NC. AD: Alzheimer's disease; MCI: mild cognitive impairment; NC: normal control; ACC: classification accuracy; SEN_C: classification sensitivity; SPE: specificity; AUC: area under receiver operating characteristic curve.

Methods	AD vs. NC (%)				MCI vs. NC (%)			
	ACC	SEN_C	SPE	AUC	ACC	SEN_C	SPE	AUC
LeNet	83.8	73.2	92.4	84.3	68.8	79.3	63.4	70.8
VGGNet	84.7	77.3	90.8	85.8	70.9	81.9	65.2	72.6
DenseNet	86.6	79.4	92.4	87.2	74.1	77.2	68.1	72.8

on the results of our dataset but it may not be optimal for other datasets. The results show that there is no great difference in the AUC with the patch size over $62 \times 48 \times 58$. For simplicity, the same patch size can be used for both hippocampus segmentation and disease classification.

The third experiment was to test the classification performances of the multi-task deep CNN model, the DenseNet model and the combination model. Table 13 shows the results of three models for classifying AD vs. NC and MCI vs. NC. Fig. 7 (a) and (b) show the ROC curves of three models for classifying AD vs. NC and MCI vs. NC, respectively. The DenseNet model greatly outperformed the multi-task deep CNN model for disease classification, suggesting that the features learned by DenseNet is more important for disease classification. The combination model outperformed both the multi-task deep CNN and DenseNet models, indicating that the features of individual models could contain complementary information for disease classification.

3.4. Comparison with other methods

We compared our proposed method to other existing methods based on structural MRI data. First, the proposed deep learning-based method was compared to other methods based on hand-crafted features such as the hippocampal volume, the GM volumes of ROIs and voxel-wise features used in some previous publications (Chupin et al., 2009; Liu et al., 2012; Zhang et al., 2011). An automatic method was proposed for hippocampal segmentation and hippocampal volumes were extracted for AD diagnosis (Chupin et al., 2009). Instead of just focusing on hippocampal volumes, 93 ROIs were parcellated in the brain and their volumetric features were calculated to train SVM classifiers for AD classification (Zhang et al., 2011). To capture fine-level features, the voxel-wise tissue densities were calculated for ensemble sparse classification of AD (Liu et al., 2012). These features are hand-crafted from one-region, multi--region to voxel levels and are widely used for AD diagnosis.

To extract the hippocampal volumes, we used the segmentation masks obtained by the multi-task deep model. The volumes from the left

Table 11
Classification results of different patch sizes for DenseNets (the patch unit is $1 \times 1 \times 1$ mm^3). AD: Alzheimer's disease; MCI: mild cognitive impairment; NC: normal control; ACC: classification accuracy; SEN_C: classification sensitivity; SPE: specificity; AUC: area under receiver operating characteristic curve.

Patch size	AD vs. NC (%)				MCI vs. NC (%)			
	ACC	SEN_C	SPE	AUC	ACC	SEN_C	SPE	AUC
50×40×48	85.2	76.2	92.4	86.3	72.3	74.5	68.1	70.8
56×42×52	85.6	82.5	88.2	86.8	72.9	77.2	64.7	71.9
62×48×58	**86.6**	**79.4**	**92.4**	**87.2**	**74.1**	**77.2**	**68.1**	**72.8**
68×50×64	86.1	84.5	87.4	87.3	73.2	74.6	70.7	73.2

Table 12
Classification results of different patch sizes for the combination method (the patch unit is $1 \times 1 \times 1$ mm^3). AD: Alzheimer's disease; MCI: mild cognitive impairment; NC: normal control; ACC: classification accuracy; SEN_C: classification sensitivity; SPE: specificity; AUC: area under receiver operating characteristic curve.

Combination	AD vs. NC (%)				MCI vs. NC (%)			
	ACC	SEN_C	SPE	AUC	ACC	SEN_C	SPE	AUC
50 × 40 × 48	87.0	85.6	88.2	90.3	74.1	72.8	76.7	76.4
56 × 42 × 52	87.9	84.5	90.8	90.9	74.4	77.6	68.1	77.1
62 × 48 × 58	**88.9**	**86.6**	**90.8**	**92.5**	**76.2**	**79.5**	**69.8**	**77.5**
68 × 50 × 64	88.0	87.6	88.2	91.4	74.7	76.8	70.9	76.9

Table 13
Comparison of the multi-task deep CNN, DenseNet and their combination methods for classifications of AD vs. NC, and MCI vs. NC. AD: Alzheimer's disease; MCI: mild cognitive impairment; NC: normal control; ACC: classification accuracy; SEN_C: classification sensitivity; SPE: specificity; AUC: area under receiver operating characteristic curve.

Methods	AD vs. NC (%)				MCI vs. NC (%)			
	ACC	SEN_C	SPE	AUC	ACC	SEN_C	SPE	AUC
Multi-task deep CNN	80.1	79.9	80.3	86.6	71.5	73.2	69.8	74.4
DenseNet	86.6	79.4	92.4	87.2	74.1	77.2	68.1	72.8
Combination	**88.9**	**86.6**	**90.8**	**92.5**	**76.2**	**79.5**	**69.8**	**77.5**

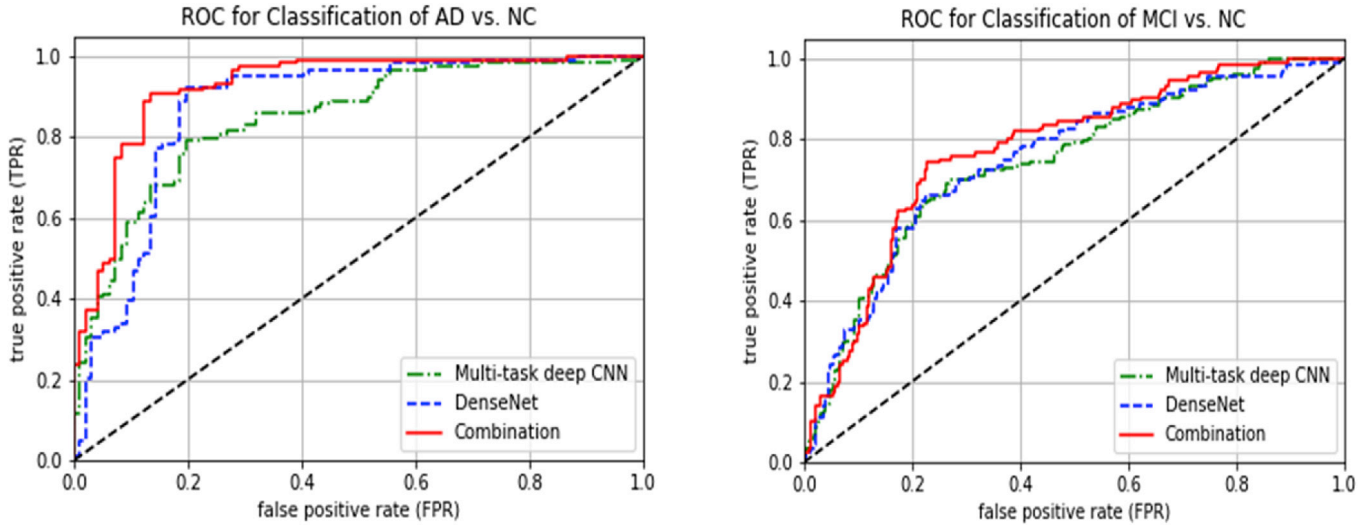

Fig. 7. Comparison of the ROC curves with the multi-task deep CNN, DenseNet and their combination models for classifying (a) AD vs. NC and (b) MCI vs. NC.

and right hippocampi were normalized by the total intracranial volume for classification. For extraction of the ROI and voxel-wise features, we followed the same procedures as those in (Liu et al., 2012; Zhang et al., 2011). We performed the tissue segmentation and nonlinear registration using the FAST model in the FSL package (Zhang et al., 2001) and the registration tool of HAMMER (Shen and Davatzikos, 2002). FAST was used to segment the MRIs into three different tissues: GM, WM, and CSF, while HAMMER was used for the nonlinear image registration and mapping the image onto 93 manually labeled ROIs (Kabani et al., 1998). For each labeled image, the normalized GM volumes from 93 ROIs were calculated as features for classification. After the image warping through HAMMER, the warped mass-preserving tissue volumes reflected the spatial distribution of tissues in an original brain and were used as the voxel-wise features for classification. Furthermore, Lasso (Kim and Kim, 2004) was used to select the most discriminative ROI and voxel-wise features for classification. To evaluate the impact of features, the same preprocessing procedures were performed for each image, and the Multi-Layer Perceptron (MLP) model with two fully connected layers followed by a softmax layer was used for classification with different features.

Table 14
Comparison of different features for classification of AD vs. NC, and MCI vs. NC. AD: Alzheimer's disease; MCI: mild cognitive impairment; NC: normal control; ACC: classification accuracy; SEN_C: classification sensitivity; SPE: specificity; AUC: area under receiver operating characteristic curve.

Method	AD vs. NC (%)				MCI vs. NC (%)			
	ACC	SEN_C	SPE	AUC	ACC	SEN_C	SPE	AUC
Hippocampal volumes	80.3	73.5	86.2	86.3	68.7	84.4	38.5	70.0
ROI features	84.7	77.4	90.5	88.1	73.5	83.9	53.4	73.6
Whole MRI DesnseNet	84.3	83.8	84.5	86.2	70.0	83.4	48.1	74.7
Voxel-wise features	86.1	84.0	87.9	90.1	74.4	87.1	50.0	75.3
Proposed method	**88.9**	**86.6**	**90.8**	**92.5**	**76.2**	**79.5**	**69.8**	**77.5**

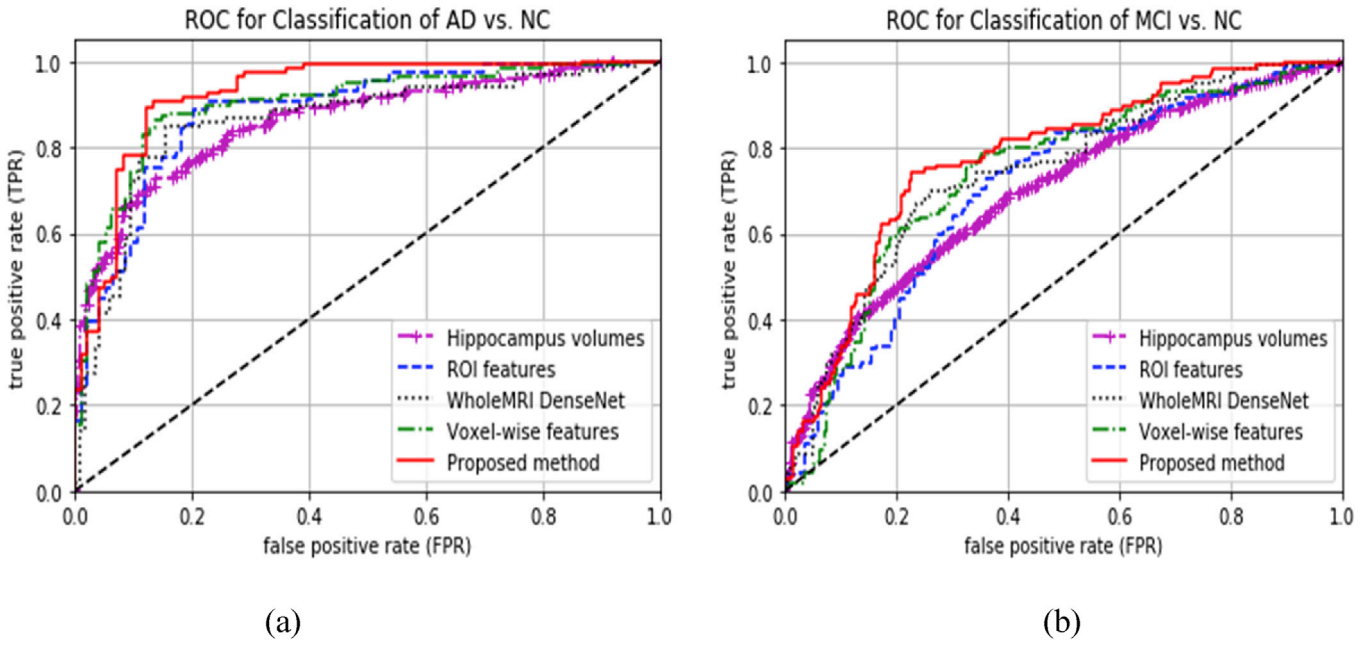

Fig. 8. Comparison of ROC curves with different features for classifying (a) AD vs. NC and (b) MCI vs. NC, where AD: Alzheimer's disease; MCI: mild cognitive impairment; NC: normal control.

In addition, one direct method is to build a deep DenseNet model on the whole brain image, which is denoted as "Whole MRI DesnseNet". The whole MR image of 256 × 256 × 256 voxels is too large to train the DenseNet model with our GPU GTX 1080 Ti because of the memory limit. Therefore, we removed the voxels of zeros values and down-sampled the images by 2, and finally obtained images of 98 × 78 × 76 voxels for training a DenseNet model. We also performed data augmentation by shifting the down-sampled image to train a robust model. Table 14 shows the comparison of classification performances with the above features as well as our proposed method. Fig. 8 (a) and (b) illustrate the ROC curves of our proposed method and other methods for classifying AD vs. NC and MCI vs. NC, respectively. It is worth noting that the results in Table 14 and Fig. 8 were obtained based on different feature extraction methods, not the design of classifier, so the results may be different from those reported in the literature.

Furthermore, we have compared our classification results with those reported in the literature using the baseline sMRI data of ADNI, as shown in Table 15. The compared methods include the conventional learning-based methods (Cao et al., 2017; Ye et al., 2016; Zhang et al., 2016) and deep learning-based methods (Korolev et al., 2017; Lian et al., 2018; Liu et al., 2014), as briefly described as follows. In the conventional learning-based methods (Ye et al., 2016; Zhang et al., 2016), the SVM classifier was used for classification with the engineered features from ROIs or landmarks. A multi-kernel learning was proposed to simultaneously conduct feature selection, manifold learning and over-sampling with the ROI features for AD diagnosis (Cao et al., 2017). The deep learning models such as Residual and plain 3D CNNs and stacked auto-encoders were investigated for AD and MCI diagnosis (Korolev

Table 15
Comparison of our results with published results for the classification of AD vs. NC and MCI vs. NC. AD: Alzheimer's disease; MCI: mild cognitive impairment; NC: normal control; ACC: classification accuracy; SEN_C: classification sensitivity; SPE: specificity; AUC: area under receiver operating characteristic curve.

Method	Subjects	AD vs. NC (%)				MCI vs. NC (%)			
		ACC	SEN_C	SPE	AUC	ACC	SEN_C	SPE	AUC
Ye et al. (2016)	51AD+99MCI+52NC	87.3	88.4	86.2	93.0	68.2	76.9	51.1	71.0
Zhang et al. (2016)	51AD+99MCI+52NC	83.1	80.5	85.1	–	73.6	75.3	69.7	–
Cao et al. (2017)	192AD+397MCI+229NC	88.6	85.7	90.4	89.8	71.9	79.0	60.7	75.3
Lian et al. (2018)	358 AD+429 NC	90.3	82.4	96.5	95.1	–	–	–	–
Liu et al. (2014)	65 AD+169 MCI+77 NC	87.8	88.6	87.2	–	76.92	74.29	78.13	–
Korolev et al. (2017)	50 AD+43 pMCI+61 NC	80.0	–	–	87.0	61.00	–	–	65.0
Proposed method	**97 AD+233 MCI+119 NC**	**88.9**	**86.6**	**90.8**	**92.5**	**76.2**	**79.5**	**69.8**	**77.5**

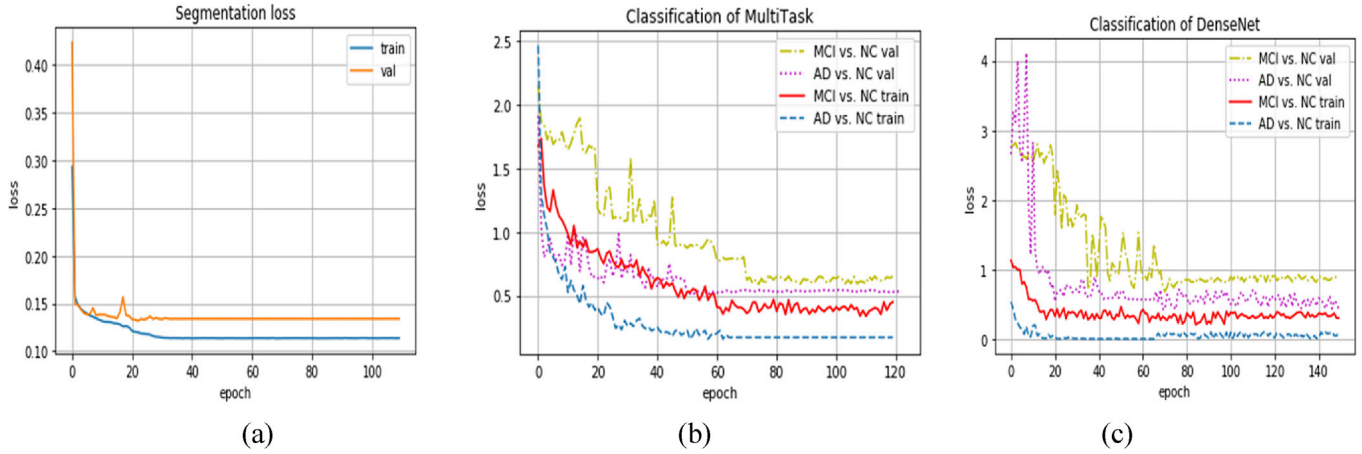

Fig. 9. The loss curves of the multi-task deep model for (a) hippocampus segmentation and (b) disease classification, and (c) the loss curves of DenseNet model for classifying AD vs. NC and MCI vs. NC on both training and validation, denoted as "AD vs. NC train", "MCI vs. NC train", "AD vs. NC val", "MCI vs. NC val", respectively.

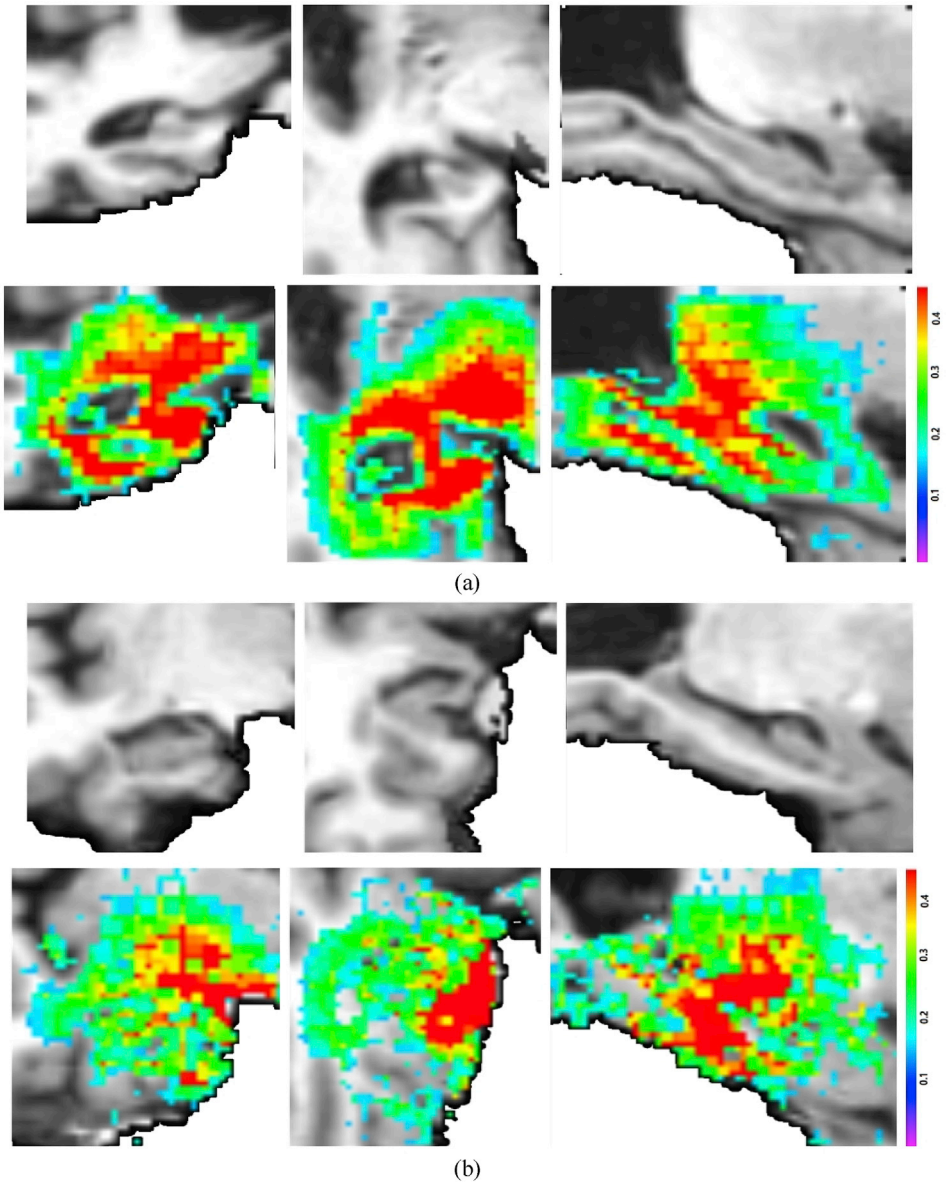

Fig. 10. The saliency maps of hippocampal regions by the proposed method for (a) AD and (b) MCI patients, where the red color highlighted regions are more relevant to AD/MCI diagnosis.

et al., 2017; Liu et al., 2014). A fully hierarchical convolutional network was proposed to identify discriminative patches and regions from the whole brain sMRI, upon which multi-scale features were jointly learned and fused to construct hierarchical classification models for AD diagnosis (Lian et al., 2018). The results show that our proposed method based on hippocampal region obtained competitive performances when compared with the methods that use whole-brain sMRI. These empirical comparison studies validated the efficacy of our proposed method. It is worth noting that the different results may be caused not only by the methods and but also by different ADNI subjects and partitions of training and testing sets.

3.5. Discussion

Based on the above results, our proposed method can not only yield accurate hippocampal segmentation results but also automatically learn multi-model features from the MRI data for improving the performance of disease classification. The proposed multi-model learning-based method outperforms the single model methods for both segmentation and classification tasks. As for the segmentation, our method achieved the DSC of 87.0% on the ADNI dataset and DSC of 86.7% on the HarP dataset. The result comparison in Table 8 shows our method outperforms a few competing methods including the FSL, Thyreau's and Cao's methods in terms of DSC, PPV, SEN_S, and VE.

As for the classification, Table 10 shows that our proposed DenseNet model performs better than the other popular deep networks such as the LeNet and VGGNet. The effects of patch sizes on the classification performance were also analyzed for DenseNet as shown in Tables 11 and 12. Furthermore, the multi-model deep network by integrating the multi-task deep model and DenseNet improved the performance of disease classification as shown in Table 13. This indicates that the two individual deep models learned complementary features for disease classification. When compared with other methods in Table 14, our method achieved a very competitive classification accuracy of 88.9% for AD vs. NC classification, which is higher than the ROI features (84.7%) and voxel-wise features (86.1%). Our method also achieved an accuracy of 76.2% for MCI vs. NC classification, which is higher than the ROI features (73.5%) and voxel-wise features (74.4%). The features from multiple ROIs were more informative than hippocampal volumes and yielded better performance. The voxel-wise features also demonstrated better discriminant ability than the ROI features because of the fine-level features. Compared to the DenseNet on the whole image, our method built the DenseNet on the hippocampus, which was easier to be trained than on the whole image with several hundreds of training subjects. We tested the proposed classification method on the additional HarP dataset consisting of 45 AD, 46 MCI, and 44 NC subjects. The AUCs are 89.8% and 72.2% for classifying AD vs. NC and MCI vs. NC, respectively, which shows the robustness when compared with the other results on the same dataset (Azar et al.,).

Training a deep model is not an easy task as the current datasets for hippocampal segmentation and disease classification are relatively small compared to the computer vision tasks. To alleviate this challenge, the data augmentation by shifting three coordinates was used to improve the robustness of the model. Fig. 9 (a) and (b) show the loss curves of the multi-task deep model for both training and validation on hippocampus segmentation and disease classification, respectively, while Fig. 9 (c) shows the loss curves of DenseNet model for both training and validation on the classifications of AD vs. NC and MCI vs. NC. The results show that the loss curves of hippocampal segmentation converges after 40 epochs, while the loss curves of all disease classifications converge when the training epoch grows to 80. The loss convergence of AD vs. NC classification is faster than that of MCI vs. NC because the classification of MCI vs. NC is more challenging than that of AD vs. NC.

As for the computational complexity, the proposed deep learning combination method includes both the offline training and online testing stages. In the offline training stage, the computational cost includes training the multi-task deep CNN model and the DenseNet model, which take 0.93 h and 1.40 h in our experiments, respectively. Thus, it takes about 2.33 h to train the whole combination model. In the online testing stage, it takes 0.29s and 0.85s on average to test the proposed algorithm for segmentation and classification of a given image, respectively, which demonstrates the usefulness of the proposed method in a real application. All the experiments were conducted on a PC with Ubuntu14.04-x64 and GPU NVIDIA GeForce GTX1080 Ti of 11 GB memory. The over-fitting problem was alleviated by using dropout techniques and data augmentation. The inverted dropout was used on the CNN layers, which performed the scaling at training time, leaving the forward pass at test time untouched.

Although the proposed method can jointly learn the feature extraction and classification model to achieve optimal diagnosis performance, it has limitations in medical interpretation and characterization of the learned features relevant to disease (i.e., AD or MCI) for clinical application. The learned features are limited in providing sufficient clinical information for understanding brain abnormalities. To facilitate the interpretability, we adopted a visualization technique proposed in (Simonyan et al., 2013) based on computing the gradient of the class score for the input image. An image-specific class saliency map was generated to highlight the discriminative areas of a given image to disease classification. We generated the saliency maps of the hippocampus patches of all test images and then calculated the average of them for illustration as shown in Fig. 10. The saliency maps indicate how the network learns the importance of corresponding areas in the prediction of disease status. The deeper the highlighted color is, the more relevant the covered region is to disease diagnosis. The highlighted areas were observed covering the regions that are important for AD diagnosis, such as hippocampus, amygdala, and parahippocampal gyrus, etc..

4. Conclusion

In this study, we proposed a new classification framework based on multi-model deep CNNs for jointly learning hippocampal segmentation and disease classification. First, a multi-task deep CNN model was constructed to jointly learn the features for hippocampal segmentation and disease classification. Based on the segmented hippocampal region, an additional 3D DenseNet was built to learn the rich and detailed image features for disease classification. Finally, the learned features from the multi-task CNN and DenseNet models are combined to classify disease status. The proposed framework can not only output the disease status, but also provide the hippocampal segmentation result. No tissue segmentation and nonlinear registration are required for MR image processing. The experimental results based on the ADNI dataset have demonstrated that our proposed approach has achieved promising performance for AD and MCI diagnosis.

Author Contributions Section

Manhua Liu: Conceptualization, Methodology, Writing- original draft, Writing, Project administration; Fan Li: Software, Writing- original draft, Writing- Original draft preparation; Hao Yan: Software, Visualization; Kundong Wang: Data curation, Supervision; Yixin Ma: Supervision, Validation; Li Shen: Writing-reviewing and editing; Mingqing Xu: Supervision, Writing-reviewing and editing, Project administration.

Declaration of competing interest

The authors report no biomedical financial interests or potential conflicts of interest.

Acknowledgments

This work was supported in part by National Natural Science Foundation of China under grants (No. 6181101049, 61981340415,

61773263) and by Shanghai Jiao Tong University Scientific and Technological Innovation Funds (No. 2019QYB02), the Shanghai Municipal Commission of Science and Technology Program (13JC1403700), "Eastern Scholar" project supported by Shanghai Municipal Education Commission (No. ZXDF089002), and Shanghai Key Laboratory of Psychotic Disorders (13dz2260500, 14-K06). Data collection and sharing were funded by the Alzheimer's Disease Neuroimaging Initiative (National Institutes of Health Grant U01 AG024904). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: Abbott, AstraZeneca AB, Bayer Schering Pharma AG, Bristol-Myers Squibb, Eisai Global Clinical Development, Elan Corporation, Genentech, GE Healthcare, GlaxoSmithKline, Innogenetics, Johnson and Johnson, Eli Lilly and Co.,Medpace, Inc.,Merck and Co., Inc., Novartis AG, Pfizer Inc., F. Hoffman-La Roche, Schering-Plough, Synarc, Inc., as well as non-profit partners the Alzheimer's Association and Alzheimer's Drug Discovery Foundation, with participation from the U.S. Food and Drug Administration. Private sector contributions to ADNI are facilitated by the Foundation for the National Institutes of Health (https://www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of California, Los Angeles.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.neuroimage.2019.116459.

References

Ahmed, O.B., Benois-Pineau, J., Allard, M., Amar, C.B., Catheline, G., 2015. Classification of Alzheimer's disease subjects from MRI using hippocampal visual features. Multimed. Tools Appl. 74, 1249–1266.

Amoroso, N., Rocca, M., Bellotti, R., Fanizzi, A., Monaco, A., Tangaro, S., 2018. Alzheimer's disease diagnosis based on the Hippocampal Unified Multi-Atlas Network (HUMAN) algorithm. Biomed. Eng. Online 17, 6.

Beg, M.F., Raamana, P.R., Barbieri, S., Wang, L., 2013. Comparison of four shape features for detecting hippocampal shape changes in early Alzheimer's. Stat. Methods Med. Res. 22, 439–462.

Boccardi, M., Bocchetta, M., Morency, F.C., Collins, D.L., Nishikawa, M., Ganzola, R., Grothe, M.J., Wolf, D., Redolfi, A., Pievani, M., 2015. Training labels for hippocampal segmentation based on the EADC-ADNI harmonized hippocampal protocol. Alzheimer's Dementia 11, 175–183.

Cao, L., Li, L., Zheng, J., Fan, X., Yin, F., Shen, H., Zhang, J., 2018. Multi-task Neural Networks for Joint hippocampus Segmentation and Clinical Score Regression. Multimedia Tools & Applications, , 1st77. Springer, pp. 1–18.

Cao, P., Liu, X., Yang, J., Zhao, D., Huang, M., Zhang, J., Zaiane, O., 2017. Nonlinearity-aware based dimensionality reduction and over-sampling for AD/MCI classification from MRI measures. Comput. Biol. Med. 91.

Chen, J., Yang, L., Zhang, Y., Alber, M., Chen, D.Z., 2016. Combining fully convolutional and recurrent neural networks for 3D biomedical image segmentation. Adv. Neural Inf. Process. Syst. 3036–3044.

Cheng, D., Wang, Y., 2017. Classification of MR brain images by combination of multi-CNNs for AD diagnosis. In: International Conference on Digital Image Processing, p. 1042042.

Chupin, M., Gérardin, E., Cuingnet, R., Boutet, C., Lemieux, L., Lehéricy, S., Benali, H., Garnero, L., Colliot, O., 2009. Fully automatic hippocampus segmentation and classification in Alzheimer's disease and mild cognitive impairment applied on data from ADNI. Hippocampus 19, 579–587.

Gerardin, E., Chételat, G., Chupin, M., Cuingnet, R., Desgranges, B., Kim, H.S., Niethammer, M., Dubois, B., Lehéricy, S., Garnero, L., 2009. Multidimensional classification of hippocampal shape features discriminates Alzheimer's disease and mild cognitive impairment from normal aging. Neuroimage 47, 1476.

Gray, K.R., Wolz, R., Keihaninejad, S., Heckemann, R.A., Aljabar, P., Hammers, A., Rueckert, D., 2011. Regional analysis of FDG-PET for use in the classification of Alzheimer'S Disease. In: IEEE International Symposium on Biomedical Imaging: from Nano to Macro, pp. 1082–1085.

Herrup, K., 2011. Commentary on "Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease." Addressing the challenge of Alzheimer's disease in the 21st century. Alzheimers Dementia J. Alzheimers Ass. 7, 335.

Ho, A.J., Raji, C.A., Saharan, P., Degiorgio, A., Madsen, S.K., Hibar, D.P., Stein, J.L., Becker, J.T., Lopez, O.L., Toga, A.W., 2011. Hippocampal volume is related to body mass index in Alzheimer's disease. Neuroreport 22, 10–14.

Hosseini-Asl, E., Keynton, R., El-Baz, A., 2016. Alzheimer's disease diagnostics by adaptation of 3d convolutional network. IEEE Int. Conf. Image Process. 126–130, 2016.

Huang, G., Liu, Z., Weinberger, K.Q., 2016. Densely Connected Convolutional Networks. CVPR.

Jack Jr., C.R., Bernstein, M.A., Fox, N.C., Thompson, P., Alexander, G., Harvey, D., Borowski, B., Britson, P.J., J, L.W., Ward, C., Dale, A.M., Felmlee, J.P., Gunter, J.L., Hill, D.L., Killiany, R., Schuff, N., Fox-Bosetti, S., Lin, C., Studholme, C., DeCarli, C.S., Krueger, G., Ward, H.A., Metzger, G.J., Scott, K.T., Mallozzi, R., Blezek, D., Levy, J., Debbins, J.P., Fleisher, A.S., Albert, M., Green, R., Bartzokis, G., Glover, G., Mugler, J., Weiner, M.W., 2008. The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods. J. Magn. Reson. Imaging : JMRI 27, 685–691.

Jr, J.C., Albert, M.S., Knopman, D.S., Mckhann, G.M., Sperling, R.A., Carrillo, M.C., Thies, B., Phelps, C.H., 2011. Introduction to the recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimers Dementia J. Alzheimers Ass. 7, 257.

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

Kim, Y., Kim, J., 2004. Gradient LASSO for feature selection. In: International Conference on Machine Learning.

Korolev, S., Safiullin, A., Belyaev, M., Dodonova, Y., 2017. Residual and plain convolutional neural networks for 3D brain MRI classification. In: IEEE International Symposium on Biomedical Imaging.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. ImageNet classification with deep convolutional neural networks. Int. Conf. Neural Inf. Process. Syst. 1097–1105.

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

Leung, K.K., Barnes, J., Ridgway, G.R., Bartlett, J.W., Clarkson, M.J., Macdonald, K., et al., 2010. Automated cross-sectional and longitudinal hippocampal volume measurement in mild cognitive impairment and Alzheimer's disease. Neuroimage 51, 1345–1359.

Lian, C., Liu, M., Zhang, J., Shen, D., 2018. Hierarchical Fully Convolutional Network for Joint Atrophy Localization and Alzheimer's Disease Diagnosis Using Structural MRI. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1-1. In press.

Lindberg, O., Walterfang, M., Looi, J.C., Malykhin, N., Ostberg, P., Zandbelt, B., Styner, M., Paniagua, B., Velakoulis, D., Orndahl, E., 2012. Hippocampal shape analysis in Alzheimer's disease and frontotemporal lobar degeneration subtypes. J. Alzheimers Dis. Jad 30, 355.

Liu, M., Zhang, J., Nie, D., Yap, P.T., Shen, D., 2018a. Anatomical landmark based deep feature representation for MR images in brain disease diagnosis. IEEE J. Biomed. Health Inform. 22 (5), 1476–1485.

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

Liu, M., Zhang, J., Adeli, E., Shen, D., 2017. Landmark-based deep multi-instance learning for brain disease diagnosis. Med. Image Anal. 43, 157–168.

Liu, M., Cheng, D., Wang, K., Wang, Y., 2018b. Multi-Modality Cascaded Convolutional Neural Networks for Alzheimer's Disease Diagnosis, 16. Neuroinformatics, pp. 295–308.

Liu, S., Liu, S., Cai, W., Pujol, S., Kikinis, R., Feng, D., 2014. Early diagnosis of Alzheimer's disease with deep learning. In: IEEE International Symposium on Biomedical Imaging, 29 April-2 May. IEEE, Beijing, China.

Milletari, F., Navab, N., Ahmadi, S.A., 2016. V-net: fully convolutional neural networks for volumetric medical image segmentation. Fourth Int. Conf. on 3d Vis. 565–571.

Morey, R.A., Petty, C.M., Xu, Y., Hayes, J.P., Wagner II, H.R, Lewis, D.V., LaBar, K.S., Styner, M., McCarthy, G., 2009. A comparison of automated segmentation and manual tracing for quantifying hippocampal and amygdala volumes. Neuroimage 45, 855–866.

Ng, Y.H., Hausknecht, M., Vijayanarasimhan, S., Vinyals, O., Monga, R., Toderici, G., 2015. Beyond short snippets: deep networks for video classification. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 7–12 June 2015, 4694–4702.

Ortiz, A., Jorge, M., Gorriz, J.M., Ramírez, J., 2016. Ensembles of deep learning architectures for the early diagnosis of Alzheimer's disease. Int. J. Neural Syst. 26, 1650025.

Platero, C., Tobar, M.C., 2016. A fast approach for hippocampal segmentation from T1-MRI for predicting progression in Alzheimer's disease from elderly controls. J. Neurosci. Methods 270, 61–75.

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

Shen, K.K., Fripp, J., Mériaudeau, F., Chételat, G., Salvado, O., Bourgeat, P., 2012. Detecting global and local hippocampal shape changes in Alzheimer's disease using statistical shape models. Neuroimage 59, 2155–2166.

Silveira M, M.J., 2015. Boosting Alzheimer disease diagnosis using PET images. In: 20th IEEE International Conference on Pattern Recognition (ICPR), pp. 2556–2559.

Simonyan, K., Vedaldi, A., Zisserman, A., 2013. Deep inside convolutional networks: visualising image classification models and saliency maps. Computer Science.

Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

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.

Suk, H.I., Lee, S.W., Shen, D., 2015. Latent feature representation with stacked auto-encoder for AD/MCI diagnosis. Brain Struct. Funct. 220, 841–859.

Thyreau, B., Sato, K., Fukuda, H., Taki, Y., 2018. Segmentation of the hippocampus by transferring algorithmic knowledge for large cohort processing. Med. Image Anal. 43, 214–228.

Wang, X., Gao, L., Song, J., Shen, H., 2017. Beyond frame-level CNN: saliency-aware 3-D CNN with LSTM for video action recognition. IEEE Signal Process. Lett. 24, 510–514.

Wang, Y., Nie, J., Yap, P.T., Shi, F., Guo, L., Shen, D., 2011. Robust deformable-surface-based skull-stripping for large-scale studies. Med. Image Comput. Comput. Assist. Interv. 14, 635–642.
Xin, L., Hong, X., Zhen, Z., Tong, L., 2012. 3D texture analysis of hippocampus based on MR images in patients with alzheimer disease and mild cognitive impairment. J. Beijing Univ. Technol. 38, 942–948.
Ye, T., Zu, C., Jie, B., Shen, D., Zhang, D., 2016. Discriminative multi-task feature selection for multi-modality classification of Alzheimer's disease. Brain Imag. Behav. 10 (3), 739–749.
Zhan, L., Zhou, J., Wang, Y., Jin, Y., Jahanshad, N., Prasad, G., Nir, T.M., Leonardo, C.D., Ye, J., Thompson, P.M., 2015. Comparison of nine tractography algorithms for detecting abnormal structural brain networks in Alzheimer's disease. Front. Aging Neurosci. 7, 48.
Zhang, D., Wang, Y., Zhou, L., Yuan, H., Shen, D., 2011. Multimodal classification of Alzheimer's disease and mild cognitive impairment. Neuroimage 55 (3), 856–867.
Zhang, J., Gao, Y., Gao, Y., Munsell, B., Shen, D., 2016. Detecting anatomical landmarks for fast Alzheimer's disease diagnosis. IEEE Trans. Med. Imaging 35, 2524–2533.
Zhang, J., Liu, M., An, L., Gao, Y., Shen, D., 2017. Alzheimer's disease diagnosis using landmark-based features from longitudinal structural MR images. IEEE J. Biomed. Health Inform. 21 (6), 1607–1616.
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.
Zhu, X., Suk, H.I., Zhu, Y., Thung, K.H., Wu, G., Shen, D., 2015. Multi-view classification for identification of Alzheimer's disease, 9352. Multi-view Classification for Identification of Alzheimer's Disease[C]// International Workshop on Machine Learning in Medical Imaging, p. 255.
Azar, Z., S., F.V., C., P.J., D., L.C., The EADC-ADNI Harmonized Protocol for Hippocampal Segmentation: A Validation Study. Neuroimage, S1053811918305846-.