



Ensemble of 3D densely connected convolutional network for diagnosis of mild cognitive impairment and Alzheimer's disease[☆]

Hongfei Wang^{a,b}, Yanyan Shen^a, Shuqiang Wang^{a,*}, Tengfei Xiao^d, Liming Deng^c,
Xiangyu Wang^a, Xinyan Zhao^{a,b}

^aShenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, Guangdong Province 518000, China

^bSchool of Software Engineering, University of Science and Technology of China, Hefei 230026, China

^cDepartment of Systems Engineering and Engineering Management, City University of Hong Kong, Hong Kong, China

^dSchool of Data and Computer Science, Sun Yat-Sen University, Guangzhou 510006, China

ARTICLE INFO

Article history:

Received 20 August 2018
Revised 27 November 2018
Accepted 7 December 2018
Available online 29 December 2018

Communicated by Jun Yu

Keywords:

3D dense connection
Ensemble learning
Convolutional neural network
Alzheimer's disease

ABSTRACT

Automatic diagnosis of Alzheimer's disease (AD) and mild cognition impairment (MCI) from 3D brain magnetic resonance (MR) images plays an important role in early treatment of dementia disease. Deep learning architectures can extract potential features of dementia disease and capture brain anatomical changes from MRI scans. This paper proposes an ensemble of 3D densely connected convolutional networks (3D-DenseNets) for AD and MCI diagnosis. First, dense connections were introduced to maximize the information flow, where each layer connects with all subsequent layers directly. Then probability-based fusion method was employed to combine 3D-DenseNets with different architectures. Extensive experiments were conducted to analyze the performance of 3D-DenseNet with different hyper-parameters and architectures. Superior performance of the proposed model was demonstrated on ADNI dataset.

© 2018 Elsevier B.V. All rights reserved.

1. Introduction

Alzheimer's disease (AD) has been known as one of progressive neurodegenerative diseases and currently been ranked as the fourth most common cause of death in developed countries. AD is caused by the damage and destruction of nerve cells in brain regions which are related to memory, and its most common symptoms are memory loss and cognitive decline [1]. Mild cognitive impairment (MCI) is an intermediate transition between Normal and AD [2]. In a recent research, 32% of patients with MCI worsened Alzheimer's disease within 5 years [3]. Early diagnosis and intervention of MCI play an important role in reducing the incidence of AD. Common practice of dementia diagnosis is the use of magnetic resonance imaging (MRI) [4,5] which can create a 3D representation and capture changes in the structure of the brain through magnetic fields and radio waves [6]. Numerous of ma-

chine learning methods have been proposed for automatic recognition of dementia diseases [7–9], in particular, deep learning shows prominence in computer-aided diagnosis of AD and MCI [10–12]. In the latest studies, convolutional neural networks (CNNs) have been proved to be excellent in automatic diagnosis of cognitive disease from brain MR images [13]. Compared with 2D convolutions on slices, 3D convolutions on a whole MRI can capture potential 3D structural information which may be essential for discrimination [14]. Comparatively, 3D-CNN has presented outstanding performance on AD and MCI classification [15]. Due to the complex structure of 3D MRI and its high-dimensional features, the 3D CNNs would be designed deeper to model high-level abstractions of the data. However, performance of 3D CNNs are very limited when the gradient information passes through many layers, because the gradient information may vanish during the forward and backward propagation. In this paper, we proposed an ensemble of 3D densely connected convolutional networks for AD and MCI diagnosis. Dense connections were introduced to improve the feature utilization, then the network could be deeper due to less feature increment in each layer and fewer parameters. What is more, the ensemble approach can decrease the misrecognition risk of selecting a single classifier and is becoming popular for medical image analysis [16,17]. A probability-based ensemble method was employed to further increase the model performance in this paper. The main contributions are listed as follows:

[☆] This work was supported in part by Shenzhen Overseas High-level Talents Innovation Funds under Grant No. KQJSCX20170331162115349, in part by National Natural Science Foundation of China under Grant Nos. 61502473 and 61872351, and in part by the Natural Science Foundation of Guangdong Province under Grant 2016A030313176, and Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database.

* Corresponding author.

E-mail address: sq.wang@siat.ac.cn (S. Wang).

(1) We integrated dense connections to maximum information and gradients flow between layers within the deep 3D convolutional networks and developed the 3D-DenseNet to train the automated AD and MCI detection model. The network is easier to train and avoids over-fitting due to its less parameters.

(2) We presented a novel ensemble of 3D-DenseNet to boost the performance of dementia detection model. It was constructed by varying hyper-parameters and architecture around the optimal values for base 3D-DenseNets.

(3) We set sufficient experiments and analyze primary factors affecting the performance of the 3D-DenseNet in detail. Effect of primary hyper-parameters on model performance were demonstrated. We revealed the important role of dense connections in improving the performance of 3D networks.

The rest of this paper is organized as follows. In Section II, we review the previous machine learning and deep learning methods for medical image analysis and dementia disease diagnosis. In Section III, the pre-processing techniques, proposed learning framework and the training strategies are introduced. The experiments and parameter analysis of this study are presented in Section IV. We discuss the proposed methods and conclusions of the paper in Section V.

2. Related works

Precomputed medical descriptors together with statistical and conventional machine learning methods have been widely used to aid the classification of AD. Risacher et al. [18] calculated the hippocampal volumes grey matter (GM) density, and cortical thickness values from segmented regions of interest (ROI). Then voxel-based morphometry (VBM) method was used for MRI analysis and AD classification. Cai et al. [19] designed a 3D pathology-centric masks to extract cerebral metabolic rate of glucose consumption (CMRGlc), and a content-based retrieval method was proposed for 3D neurological images. Liu et al. [20] extracted 83 ROIs from 3D brain MRI and PET scans, and proposed a Multifold Bayesian Kernelization (MBK) based on a Bayesian framework to diagnose AD. Zhang et al. [21] extracted 93 ROIs from the MRI and PET scans using standardized templates, and the multimodal features were combined through a multi-kernel support vector machine (SVM). F Zhang et al. [22] used k-means clustering to build a low-level features dictionary involving lesion metabolism index, mean index and Fisher index. Then, Probabilistic Latent Semantic Analysis (PLSA) and Canonical Correlation Analysis (CCA) were used to combine the features and capture the latent associations.

As a popular machine-learning tool, deep learning methods have been widely used in computer-aided diagnosis [23]. Liu et al. [24] trained a deep neural network contained auto-encoders to combine multimodal features which were extracted from 83 ROIs of PET and MRI scans. Li et al. [25] achieved multimodal fusion of PET and MRI features through restricted Boltzmann machine (RBM) and improved the classification accuracies by designing a multi-task deep learning network with dropout. ROI-based methods can significantly extract representative features and partly reduce the feature dimension, but the ROIs are too empirical to capture the larvaceous features entirely which are associated with AD diagnose. Convolutional neural networks (CNNs) have been widely used in pattern recognition and present an outstanding performance on AD classification through medical images. Ciprian D. Billones et al. [26] selected 20 successive slices from MRI, under the hypothesis that the slices cover the significant areas for the dementia detection. And each serial number of the 2D slices was used to train a 2D-CNN respectively modified from the VGGNet. 3D-CNN can capture more complete spatial features through its space association ability. Hosseini-Asl et al. [27] proposed a 3D convolutional neural network which combines a 3D

convolutional autoencoder pre-trained with registered images. Payan et al. [28] built a learning algorithm by combining 3D convolutions and sparse autoencoders, and used it on a whole MRI. Cheng et al. [29] extracted a number of 3D patches from the whole MRI and a patch transformed into features by 3D-CNN. Finally, multiple 3D-CNNs were used to combine the features and demonstrated the effectiveness for AD classification.

A successful group of approaches are proposed to relief the challenges which caused by enlargement of data dimension. Junior et al. [30] proposed a morphological analysis method based on contour feature extraction and randomized neural network modeling. The main advantage of this method is that the classification accuracy is improved by aggregating the weights of feature vectors. Hong et al. [31] fused multimodal features and developed a 3D pose recovery multi-layer network based on non-linear mapping, in which the recovered poses are closer to the original 2D frames than through the linear regression. Yu et al. [32] joint the deep CNNs and the tree classifier and develop a multi-task learning algorithm to detect privacy-sensitive object classes from images. This method achieved competitive results with respect to accuracy and computational efficiency. Osipov et al. [33] bound the controlled elements with signal and proposed a dynamic space Ctime structures in RNNs, which can be used to create multi-level structure artificial neurons. 3D graph contains more abundant structure and space information and become important in this days computer vision system, but its complexity of structure makes it difficult to be represented effectively. Wu et al. [34] proposed a convolutional deep belief network to represent a geometric 3D shape and estimate the distribution of 2D image on a 3D voxel grid. Qi et al. [35] analyzed the performance of volumetric CNNs and multi-view CNNs on classification tasks, and introduced multi-resolution filtering to achieve state-of-the-art 3D classification results. This research may help relief the bottleneck of 3D resolution and further improve the efficiency of 3D Convolution network. For reducing data dimension and optimizing feature expression, Zhang et al. [36] proposed an unsupervised deep-learning data dimensionality reduction method named LDFA, which can learn both local and global features of the sample. The proposed local stacked contractive auto-encoder (SCAE) are demonstrated to improve the performance of the unsupervised learning. Due to the limitation of depth, traditional neural network may not achieve well performance when dealing with irregular data. To solve this problem, Liu et al. [37] used a gating strategy to control information flow and increase network stability, and proposed a gate-based deep network which can be deeper and extract high-level features.

3. Methods

3.1. Data acquisition and pre-processing

In this work, we obtained neuroimaging data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database [38]. The study involves more than 1,000 participants including people with MCI, patients with diagnosed AD and normal contrasts. Most of the participants were collected repeatedly for two to six times, and the interval between neighbor scans was more than a year. Time sequence scans provide researchers with a novel discovery of AD progression.

As shown in Fig. 1, a total of 833 T1-weighted MRIs were employed, which were collected from 624 participants, including both male and female, and their ages range from 70 to 90. Since a given participant's brain structure makes a difference after a period of time, we selected two scans with the longest interval of one participant as different subjects, as long as the interval is more than three years. And when 10-fold cross-validation was employed to evaluate our models, the subjects selected from the same

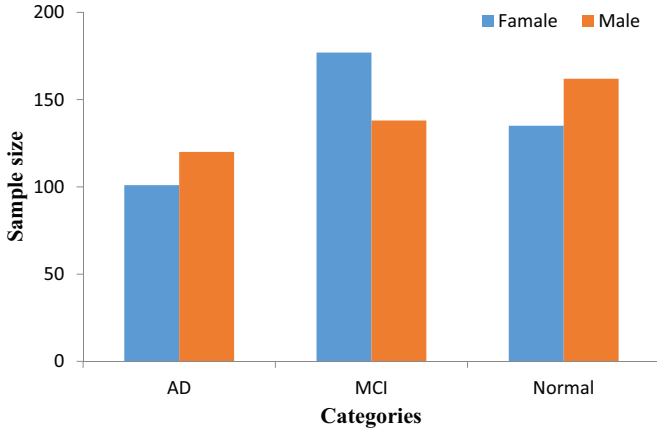


Fig. 1. Distribution of the subjects.

participant were bound up and placed in a same subgroup, so that they were forbidden to appear both in training and testing dataset.

The selected MRIs were considered the best in the quality ratings and have undergone grad-warping, intensity correction, and have been scaled for gradient drift using the phantom data. Brain Extraction Tool (FSL-BET) [39] was used to strip non-brain tissue from an image of the whole head, and images were aligned to the standardized template using FSL FLIRT [40]. The dimension of each image is $91 \times 109 \times 91$ in Neuroimaging Informatics Technology Initiative (Nifti) file format. Preprocessed steps were shown in Fig. 2.

3.2. Proposed method

Consider a traditional network comprises l layers, we denote x_l as the output of the l th layer, and each layer implements a non-linear transformation $H_l(\cdot)$, where l indexes the layer. To boost the training against the vanishing gradients and improve the information flow within the network, the DenseNet [41] implements the connections from a layer to all its subsequent layers. We extended the idea of dense connectivity to 3D volumetric image processing tasks. In particular, x_l is defined as:

$$x_l = H_l([x_0, x_1, \dots, x_{l-1}]) \quad (1)$$

where x_0, x_1, \dots, x_{l-1} are 3D feature volumes produced in preceding layers, [...] refers to the concatenation operation. Fig. 3(a) illustrates a dense unit. Composite function $H_l(\cdot)$ consists of three operations: a batch normalization (BN) to reduce internal covariate transform [42], a rectified linear unit (ReLU) to accelerate training process, and spatial convolution with $k3 \times 3 \times 3$ convolution kernels to generate 3D feature volumes. The basic framework of a 3D dense block is shown in Fig. 3(b). A dense unit is regarded as one layer in a dense block and each layer is connected with all subsequent layers directly. With this dense connection mechanism, feature utilization become more effective and fewer feature increments are added to each layer than traditional CNNs. Therefore, the network are very narrow and has fewer parameters.

Bottleneck layers and transition layers are also used to reduce parameters. $1 \times 1 \times 1$ convolution is employed as bottleneck layer to reduce the input feature-volumes, before convolution layer [43,44]. After the mechanism of bottleneck layer, multichannel feature-volumes are fused and only a small set of feature-volumes were added to the next layer while the preceding features are remained. Transition layers were also introduced to further improve model compactness with the hyper-parameter theta controlling the degree of compression. For a dense block contains m feature-volumes, the output feature-volumes of the following transition layer decrease to $\lfloor \theta m \rfloor$, where $0 < \theta \leq 1$. Therefore, the 3D-

DenseNet layers become very narrow and require fewer parameters than traditional network, but it can perform well by making the best use of features through dense connections. To further eliminate redundancy and improve the feature expression performance of the model, we introduce dropout between pooling layer and linear layer. The 3D-DenseNet with two dense blocks is illustrated schematically in Fig. 3(c).

As mentioned above, the 3D-DenseNets with different hyper-parameter sets appeared various architectures. We demonstrated that the performance of 3D-DenseNet was sensitive to its hyper-parameters through extensive experiments in Section IV. So training with different hyper-parameters can adjust the instability of the base 3D-DenseNets and enhance their diversity. Based on extensive experimental results with varying hyper-parameter sets, we generated base networks with different structures by changing hyper-parameters randomly around the optimal value. All base 3D-DenseNets work independently and output the class probabilistic score by a softmax layer. We fused their outputs by probability-based fusion method. The proposed ensemble model is shown in Fig. 4.

In traditional majority voting method, the prediction results of most classifiers are used as the final prediction labels. Each classifier is independent and the error rates between different classifiers are irrelevant, so that the performance of the ensemble model is better than a single classifier. But for multi-classification tasks, this method may not be very effective. Single classifiers perform well on most subjects, but for some subjects which are difficult to classify, the error rates will increased due to the uncertainty among multiple categories. For example, three classifiers are considered, the output probabilities of softmax lay for $\{AD, MCI, Normal\}$ are I: $\{0.8, 0.1, 0.1\}$, II: $\{0.4, 0.5, 0.1\}$, III: $\{0.3, 0.4, 0.3\}$, respectively. Based on the majority voting method, the prediction result is MCI. But it is not completely correct, since the prediction result of classifier I is more credible while the II and III has more uncertainty.

In our approach, a simple probability-based ensemble method was employed [45], in which the output probabilities of softmax layer from base classifiers will be reintegrated. Meanwhile, the predictions of each classifier will not be ignored. In ternary classification, i base classifiers were selected, the probabilities of 3D-DenseNet _{i} assigned to categories on testing set were:

$$P^i = (\alpha_1^i, \alpha_2^i, \alpha_3^i) \quad (2)$$

where α_j^i indicates the probabilities of the class j . We normalized the P^i by:

$$P^i = \frac{P^i}{\max[\alpha_1^i, \alpha_2^i, \alpha_3^i]} \quad (3)$$

When outputs of m base 3D-DenseNets have been computed, the final class label was determined by the proposed fusion model as follows:

$$y = \arg \max \left(\prod_{i=1}^m \alpha_1^i, \prod_{i=1}^m \alpha_2^i, \prod_{i=1}^m \alpha_3^i \right) \quad (4)$$

4. Experiments

4.1. Data and Implementation

833 MR subjects were used for the experiments, including 221 AD subjects, 297 MCI subjects and 315 Normal control subjects. We utilized the 10-fold cross-validation method to test the model performance. The original dataset was randomly partitioned into 10 equal sized subsamples, a single subsample was retained as the validation data for testing the model, and the remaining 9 subsamples are used as training data. The cross-validation process was

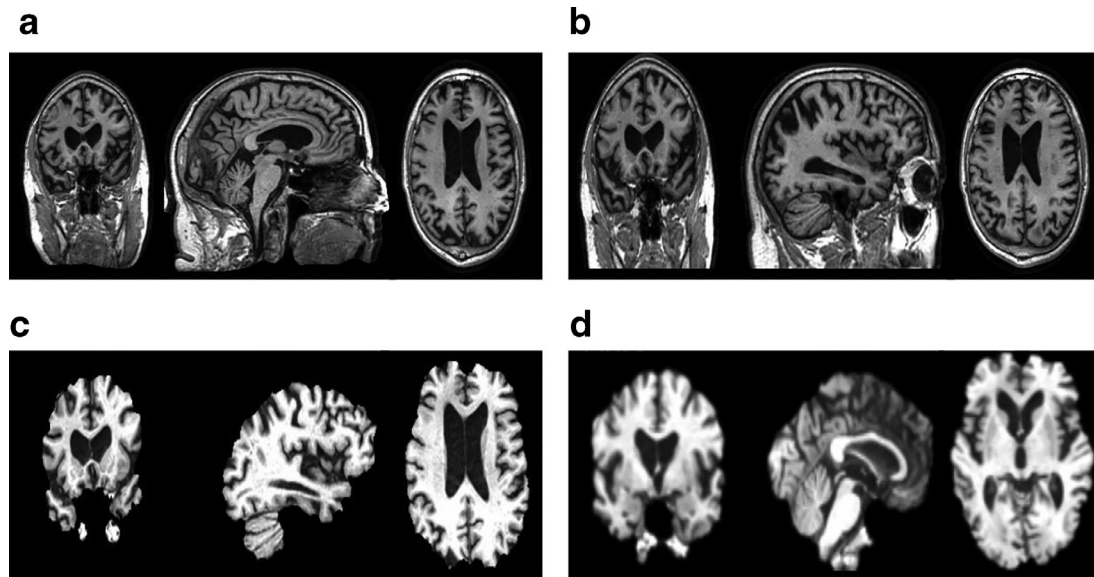


Fig. 2. Images after each pre-processing steps: (a) Original image; (b) image after removal of redundant tissues; (c) image after brain extraction; (d) image aligned to the MIN152 template.

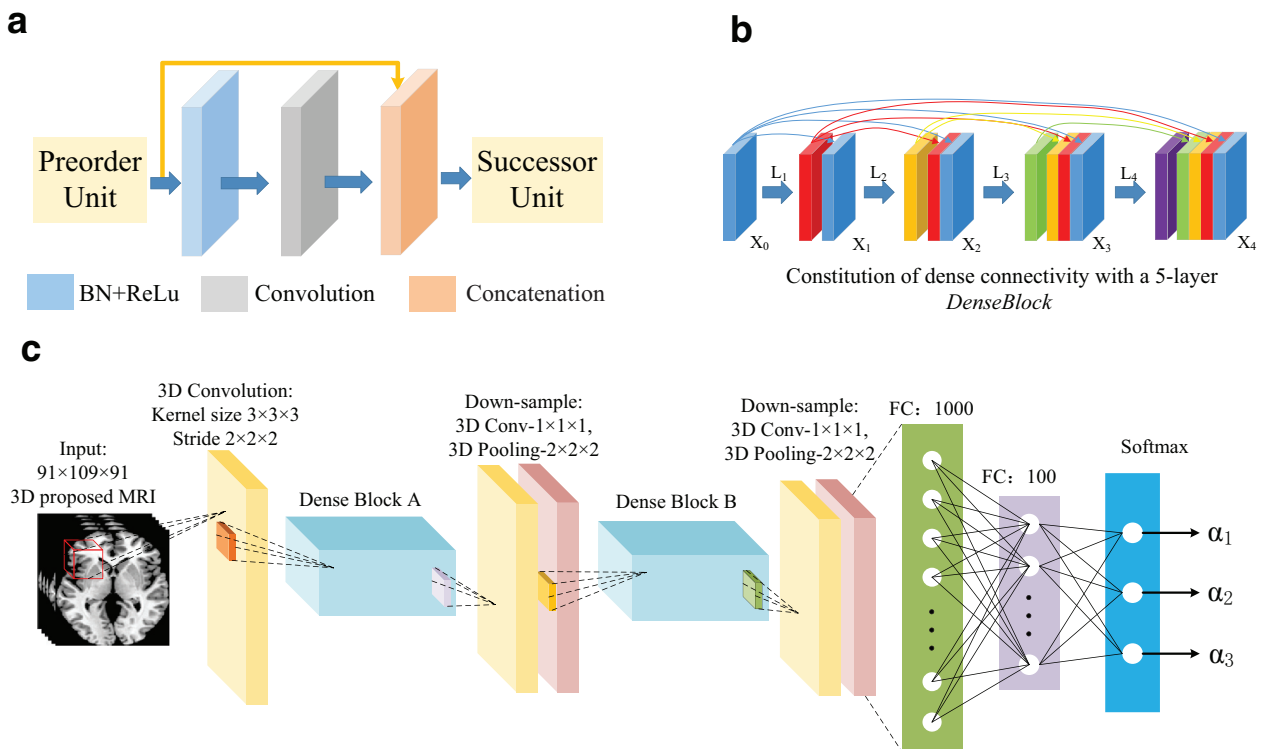


Fig. 3. (a): A composite dense unit; (b): Constitution of dense connectivity with a 5-layer dense block; (c): A 3D-DenseNet with 2 blocks, the number of feature-volumes were changed between blocks through transition layers.

then repeated 10 times, with each of the 10 subsamples used exactly once as the validation data. A ternary classifier (AD vs. MCI vs. Normal) and three respective binary classifiers (AD vs. Normal, AD vs. MCI, MCI vs. Normal) are used to report the classification results. The subjects selected from the same participant were forbade to appear in both the training set and the testing set. All the experiments were performed on a system with NVIDIA Tesla P100 GPU.

4.2. Experimental steps and evaluation

A 3D-DenseNet was selected as base classifier for comparison with the ensemble method. The base 3D-DenseNet was trained and a series of experiments were conducted to choose optimal hyper-parameters. Subsequently, some 3D-DenseNets were generated by varying primary hyper-parameters around the selected optimal values randomly. Then comparison of the ensemble method

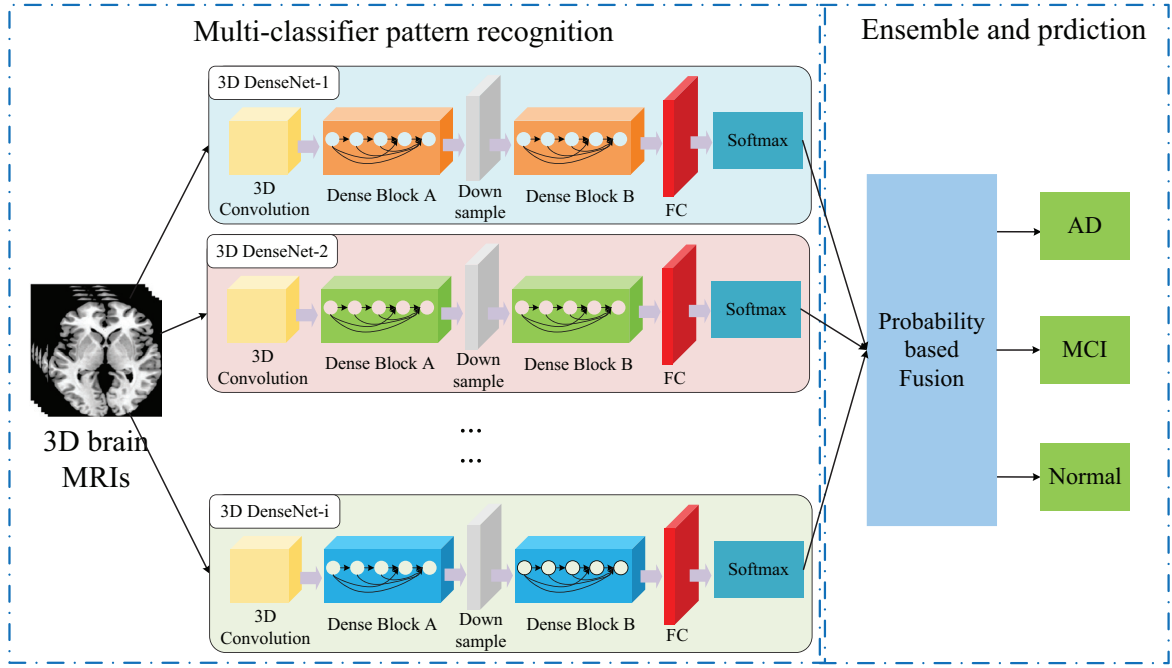


Fig. 4. Architectures of proposed ensemble 3D-DenseNet framework for AD and MCI diagnosis.

Table 1
Confusion matrix of binary classification.

	Predicted positive (Class A)	Predicted negative (Class B)
Actual positive (Class A)	True positive (TP)	False negative (FN)
Actual negative (Class B)	False positive (FP)	True negative (TN)

Table 2
Confusion matrix of binary classification.

	Predicted Class A	Predicted Class B	Predicted Class C
Actual Class A	True A (T_A)	False AB (F_{AB})	False AC (F_{AC})
Actual Class B	False BA (F_{BA})	True B (T_B)	False BC (F_{BC})
Actual Class C	False CA (F_{CA})	False CB (F_{CB})	True C (T_C)

and base classifiers was conducted to prove superiority of the ensemble method.

The performance of the classifiers can be interpreted from the confusion matrices, which record model performance across categories. The confusion matrices of binary and ternary classification problems are shown in Table 1 and Table 2, respectively. The average of 10-fold cross-validation was regarded as the final results.

The performance of each model is defined as follows:

(a) Accuracy that indicates the proportion of correctly classified subjects among the whole subset,

$$Accuracy_{bin} = \frac{TP + TN}{TP + TN + FP + FN}, \quad (5)$$

$$Accuracy_{ter} = \frac{T_A + T_B + T_C}{T + F}. \quad (6)$$

(b) Precision that quantifies the proportion of samples correctly classified among the classification,

$$Precision_{bin} = \frac{TP}{TP + FP}, \quad (7)$$

$$Precision_{ter-classA} = \frac{T_A}{T_A + F_{BA} + F_{CA}}. \quad (8)$$

(c) Recall is the fraction of relevant instances that have been retrieved over the total amount of relevant instances,

$$Recall_{bin} = \frac{TP}{TP + FN}, \quad (9)$$

$$Recall_{ter-classA} = \frac{T_A}{T_A + F_{AB} + F_{AC}}. \quad (10)$$

(d) F1-score considers both the precision and recall and evaluate the model performance synthetically,

$$F_1 - score = \frac{2 \times Precision \times Recall}{Precision + Recall}. \quad (11)$$

4.3. Parametric analyses

A series of experiments were conducted to analyze performance of the 3D-DenseNet with different hyper-parameters sets including depth, growth rate and compression factor. We display the testing accuracy and errors among the 10-fold cross-validation via box-plots which represent the interquartile ranges. The boxes indicate the quartiles of the dataset while the whiskers extend to show the points that are determined to be outliers using a method which is a function of the inter-quartile range.

Analysis of growth rate. We refer to the hyper-parameter k as the growth rate of the network which indicates the number of new feature-volumes increased at each layer. As shown in Fig. 5, the accuracy of classifiers vary significantly according to different k . The AD/MCI classifier obtains state-of-the-art result with $k=15$, while MCI/Normal with $k=12$. The model with $k=9$ is sufficient to classify AD and Normal. As for the ternary classification task, state-of-the-art accuracy is achieved when $k=24$. The model obtains poor accuracy with small growth rate, because essential features for classification are not fully extracted. Relatively large growth rate can

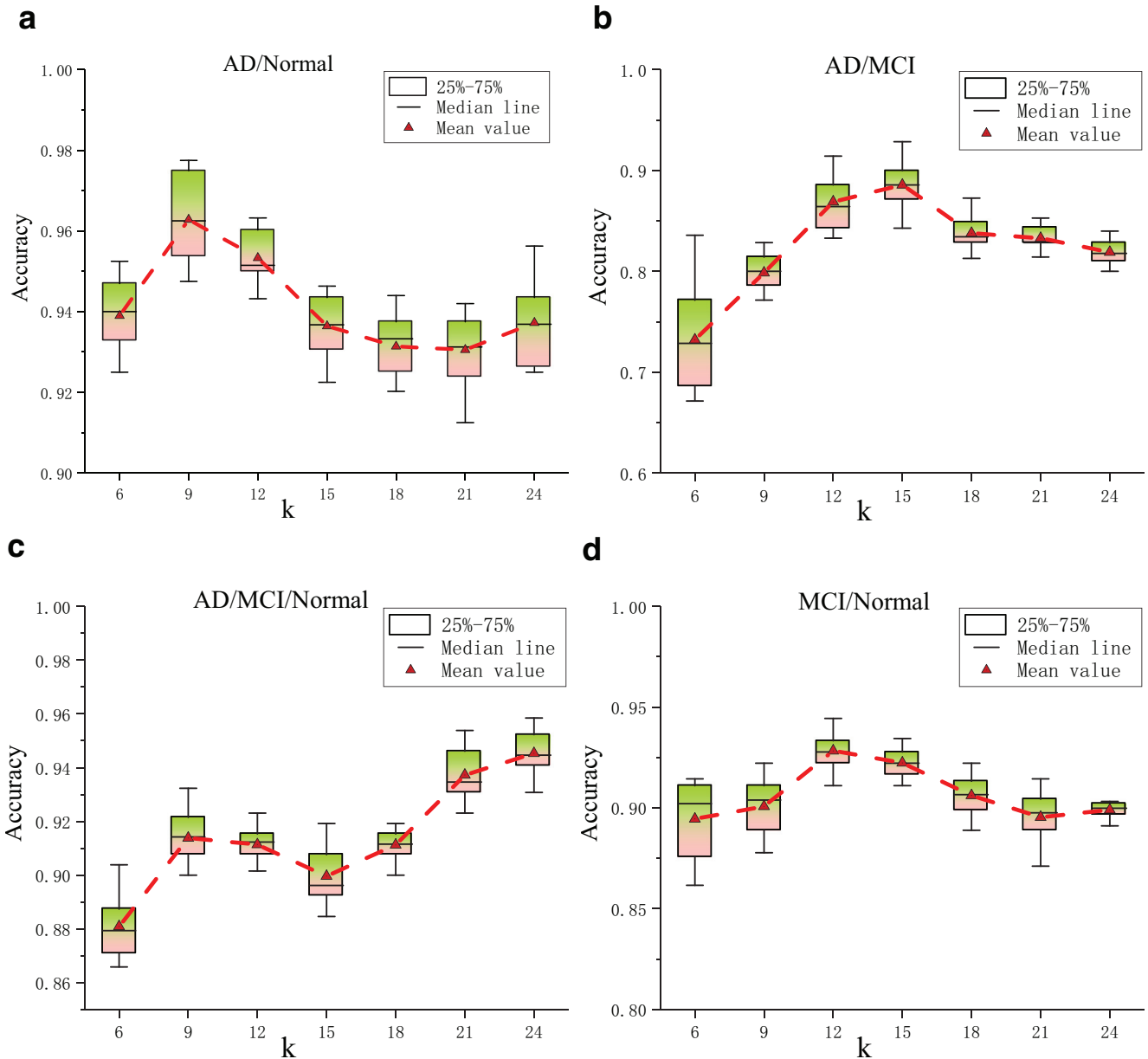


Fig. 5. Comparison of different growth rates.

boost the performance by introducing more feature-volumes. However overlarge growth rate may decrease the accuracy, because the complicated model is deficiently trained with limited training data. What is more, the model with larger growth rate has smaller range of errors, the explanation for this is that complex structure with more parameters can improve the performance of the model.

Analysis of depth. The depth is referred to as the total number of layers of all blocks in the 3D-DenseNet. To investigate the impact of depth on accuracy, the model with different depth was trained for experiments. The comparison of model performance with different depth is shown in Fig. 6. State-of-the-art accuracy can be obtained to categorize AD/Normal with depth=15, while other classifiers obtain optimal accuracy with depth=20. The mean accuracy of AD/MCI is lower than others, because AD is usually developed from MCI and the anatomical shape variations captured from MRIs are not obvious enough to identify categories. Similar to the growth rate, the network with few layers cannot express features adequately, so increasing depth properly can improve the classification accuracy. But the network with overlarge depth may

obtain poor accuracy, because parameters may not be fully trained due to limited dataset.

Analysis of compression factor. The θ is referred to as compression factor, which indicate degree of feature reduction in transition layers. Fig. 7 shows that the variation of compression factor has an prominent impact on classification accuracy. The MCI/Normal and ternary classifiers obtain optimal accuracy with a medium compression factor. The accuracy of AD/MCI declines as θ increase. And the AD/Normal obtains optimal accuracy when the model is compressed with larger θ . Simplified network structures can ignore irrelevant features for diagnosis of dementia disease and reduce overfitting in some extent. But excessive compression of the network can lead to inadequate expression of features and thus reduce the accuracy of the model.

Analysis of the optimization methods. To further accelerate the model converging and improve its performance, BP arithmetic is mended by appending optimization methods. As we can see in Fig. 8, the model optimized by momentum method achieved better accuracy than models with other optimizers in all classification

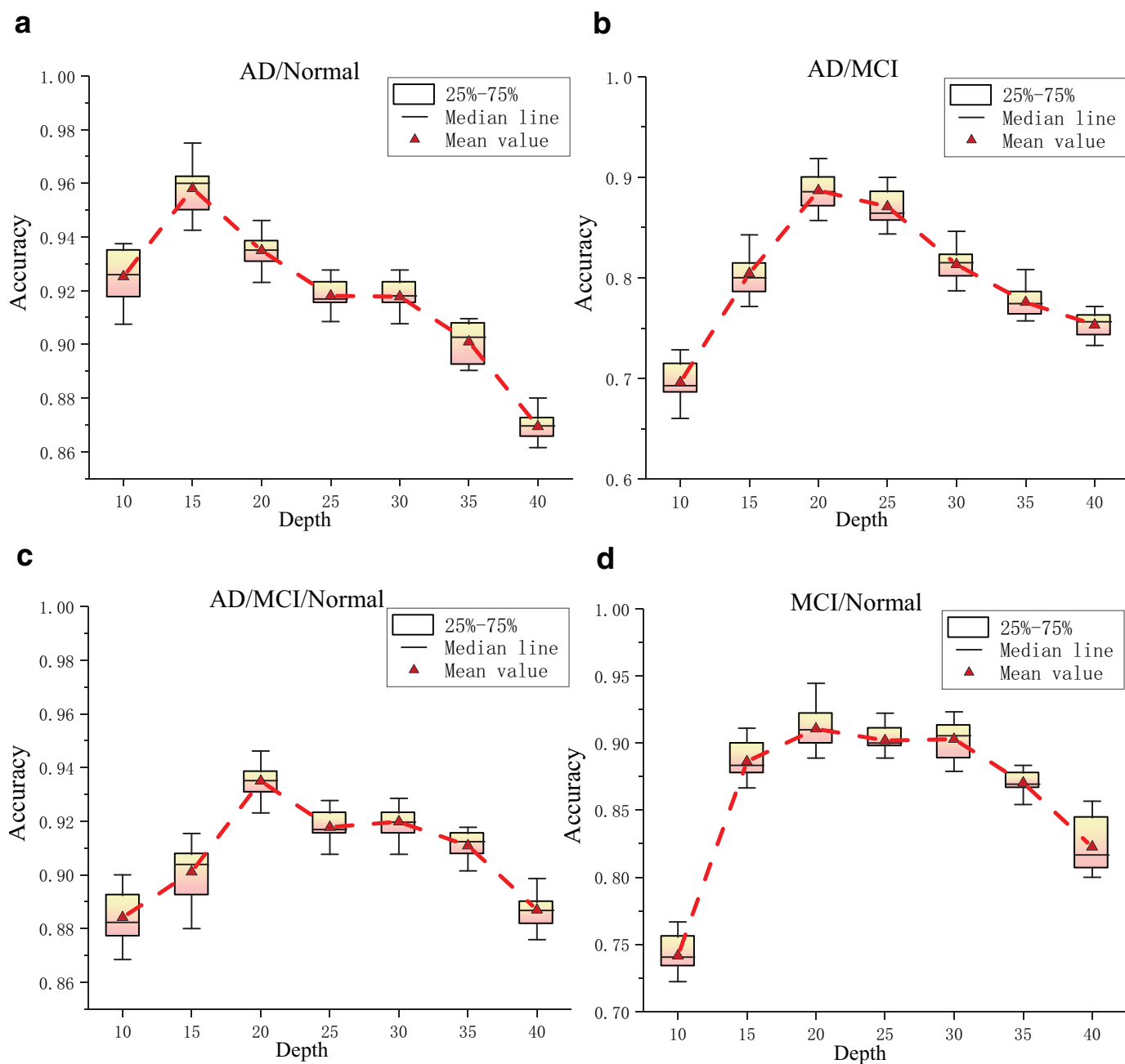


Fig. 6. Comparison of different depths.

tasks. One explanation for this is that current gradient is the accumulation of previous momentum. Larger values of momentum factor accelerate parameters update and help the model get rid of the local minimum when gradient decreases to zero.

Analysis of the model performance with different amount of training data. In order to analyze the effect of different parameters on the performance of the model, different number of samples were used for ternary classifier training. Fig. 9 shows the mean accuracy with different amount of training data, where the models with different fixed hyper-parameters were compared. Accuracy of the model with fixed growth rate reduced rapidly with reduction of training samples, while the models with fixed depth and compression factor change gently. So the depth is more insensitive to the amount of training data. In 3D-DenseNet, the growth rate focuses on the number of features which are sufficient to distinguish categories, while the depth emphasizes expressing features with

appropriate layers and mining the differences between features. And compression factors improve model compactness by reducing the amount of feature-volumes. Dense connections reuse existing features and make the information and gradients transfer effectively throughout the network, so proper depth has an important influence on the performance of the model.

Analysis of model performance with dropout. Fig. 10 shows the distribution of 10-fold cross-validation accuracy of the same architectures trained with and without dropout. As discussed previously, there may exist many redundant features in the output of the dense block, we decreased the impact of redundant features by dropout. Dropout reduced complex co-adaptations of neurons and forced classifier to give up noise by dropping some units from the network with a certain probability temporarily. Therefore, the proposed model learned more pivotal features and prevents overfitting by this way.

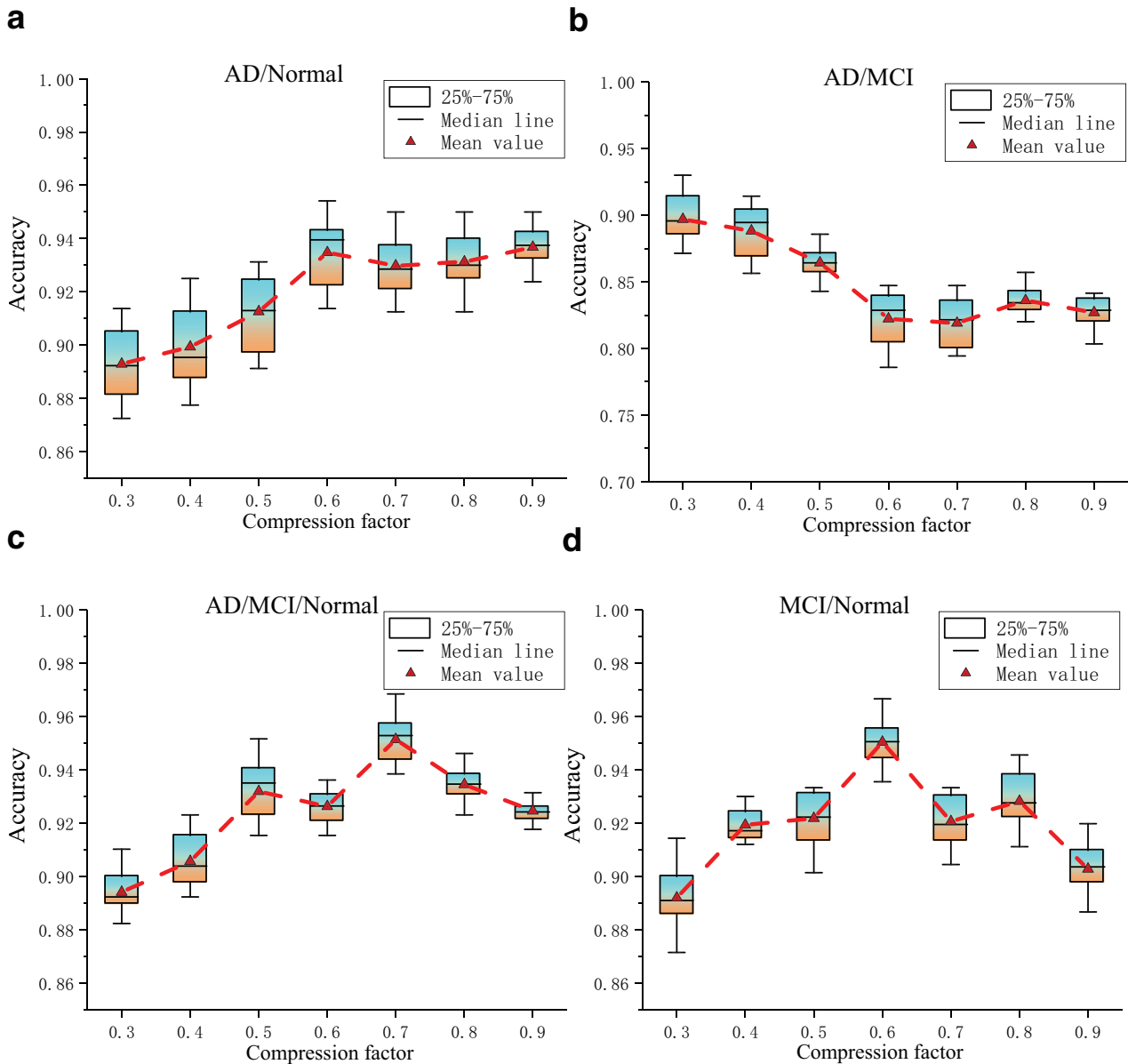


Fig. 7. Comparison of different compression factors.

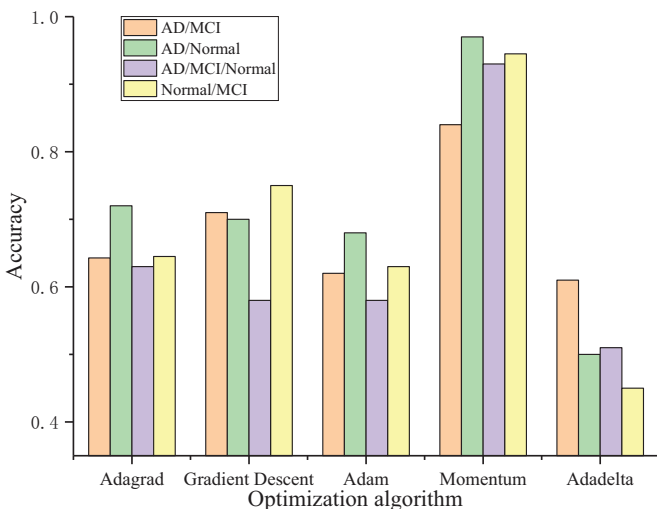


Fig. 8. Comparison of model performance with different optimization methods.

Analysis of parameter quantity and computation time. In order to investigate the computational efficiency of the 3D-DenseNet, we compared parameter quantities and training times of different structures. These were computed on a NVIDIA Tesla P100 GPU and iterated for 150 epochs with a 0.01 initial learning rate. As shown in Table 3, the network with a larger growth rate have more parameters and consume more time, because there are more feature-volumes increment at each layer. The narrow but deeper network also has higher time complexity due to its containing more layers in each dense block and extracting more abstract features. What is more, the compression of transition layer can reduce parameters significantly and consume less memory, but cannot save computing time. This may because the channels of former collective knowledge were reintegrated, but features computing in each dense block cannot be simplified.

4.4. Results

For each 3D-DenseNet, we initialized weights of the model randomly with a Gaussian distribution ($\mu = 0, \sigma = 0.01$). The ini-

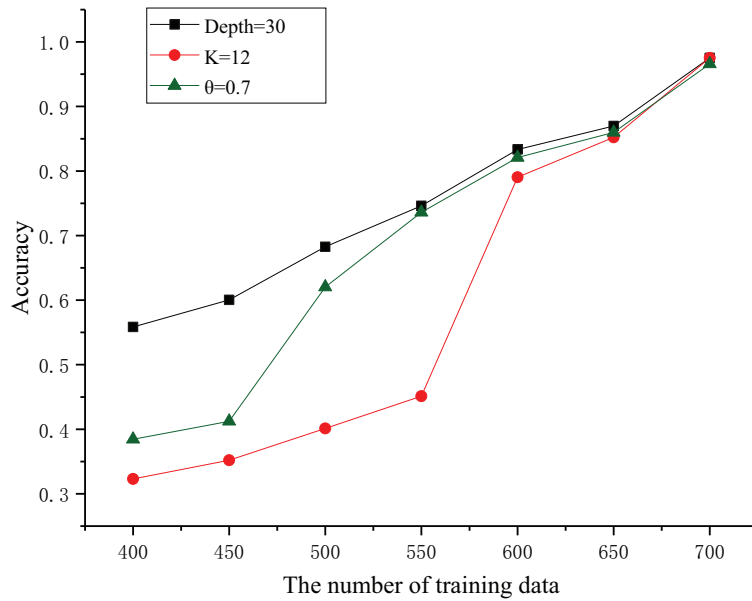


Fig. 9. Parameters sensitivity to the number of training data.

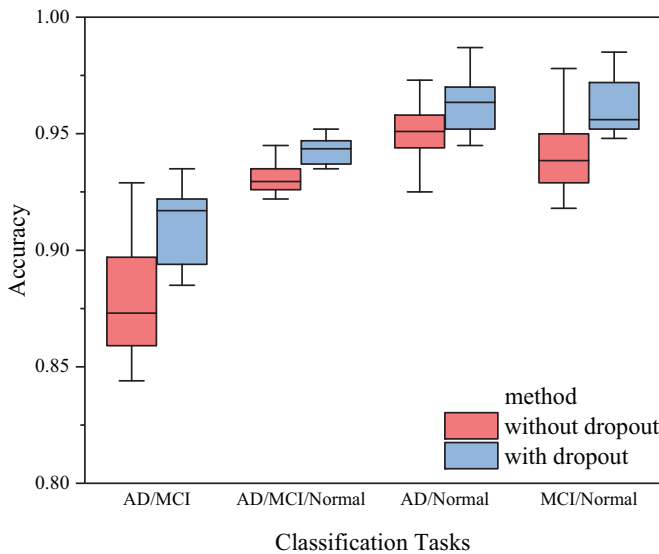


Fig. 10. Analysis of model performance with and without dropout.

tial learning rate was set to 0.01, and the poly learning rate policy was employed to update learning rate by multiplying $(1 - \frac{iter}{max_iter})^{power}$ along the training iteration. Momentum method

Table 3

Comparison of parameter quantity and computation time with different network structure.

Methods	Depth	Growth rate	θ	parameters	time
DenseNet-I	30	12	1	1.3M	2h:22min
DenseNet-II	30	24	1	5.3M	4h:15min
DenseNet-III	30	12	0.8	0.3M	2h:26min
DenseNet-IV	30	12	0.5	0.2M	2h:28min
DenseNet-V	50	12	1	4.4M	6h:23min

(with batch size=10, weight decay=0.0005 and momentum=0.9) was used to optimize training iteration. We carried out 1000 iterations for training and testing in each cross-validation. Finally, 5 different 3D-DenseNets were selected as base classifiers, and the variation ranges of accuracy among them are within 2%. The average of 10-fold cross-validation was regarded as the final results.

The experimental results are shown in Table 4. The best performance with accuracy of 97.52%, average precision of 97.13%, average recall of 97.0% and F1-score of 97.1% is given by the proposed probability-based ensemble model, while the proposed 3D-DenseNet produce the accuracy of 94.77% and the majority voting method achieve the accuracy of 95.96%. From Table 4, the following findings can be given:(1) The majority voting approach and the probability-based ensemble model can improve the model performance significantly. The fusion of multiple independent

Table 4

The performance of 3D-DenseNets and the ensemble model on testing set for AD/MCI/Normal.

Model	Class	Accuracy	Precision	Recall	F1-score
Optimal 3D-DenseNet	AD	0.9477	0.9253	0.9696	0.9469
	MCI		0.9431	0.9325	0.9405
	Normal		0.9680	0.9578	0.9628
Average of the base classifiers	AD	0.9398	0.9104	0.9242	0.9172
	MCI		0.9425	0.9213	0.9317
	Normal		0.9578	0.9680	0.9628
Majority voting	AD	0.9596	0.9365	0.9402	0.9383
	MCI		0.9435	0.9526	0.9480
	Normal		0.9684	0.9703	0.9693
Probability-based ensemble model	AD	0.9752	0.9692	0.9545	0.9617
	MCI		0.9555	0.9662	0.9598
	Normal		0.9893	0.9893	0.9893

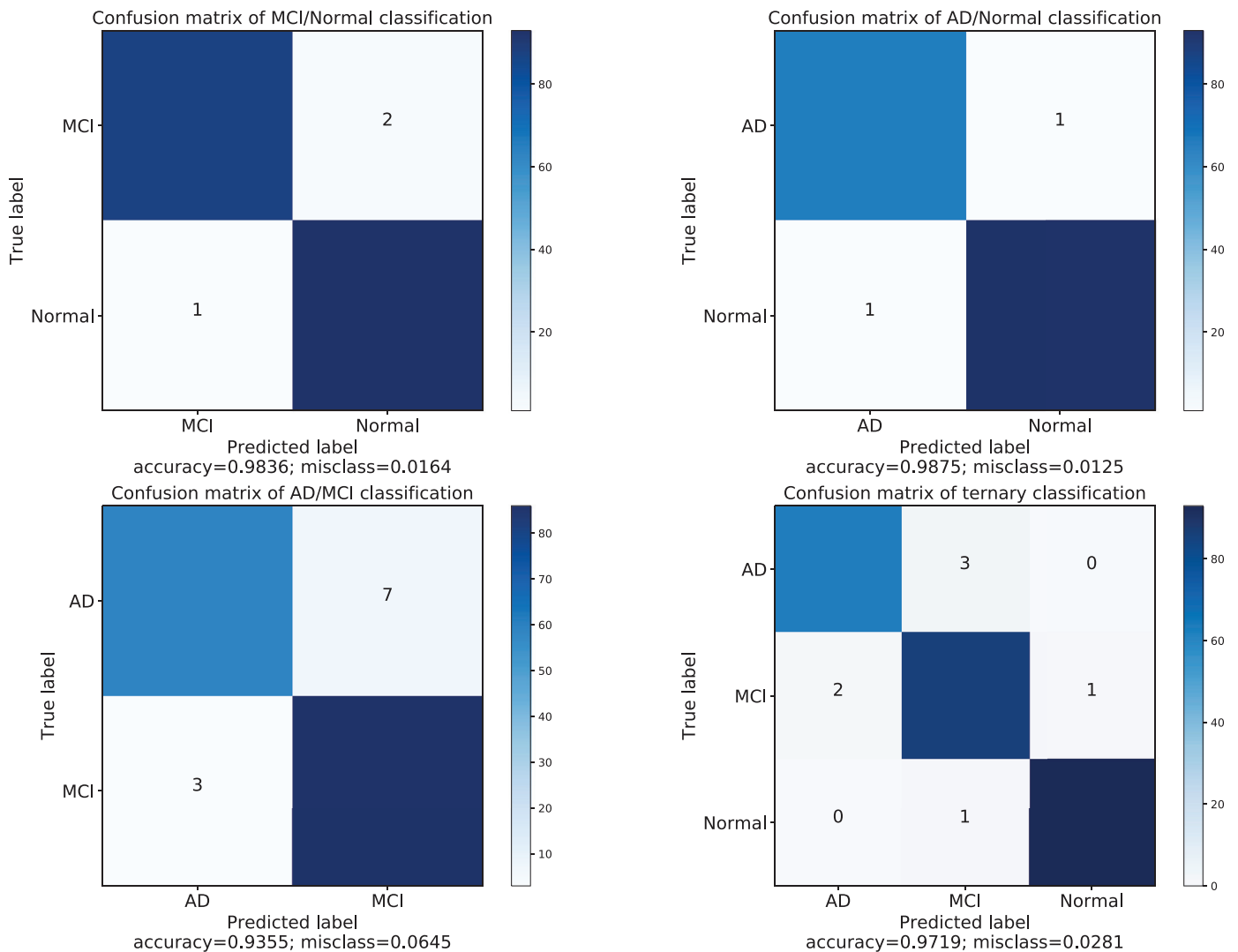


Fig. 11. Confusion matrix of four classifiers.

Table 5

Accuracy, recall and precision for different classification methods. These results correspond to stable MCI/AD classification.

Methods	Accuracy (%)	Recall (%)	Precision (%)
Payan et al. [28]	86.84	<i>n/a</i>	<i>n/a</i>
Wee et al. [46]	79.24	78.03	80.46
Wang et al. [47]	90.6	83.7	80.6
Ferreira et al. [48]	81.56	85	79
3D-DenseNet	92.20	91.12	91.68
Majority voting	92.83	91.26	92.85
Proposed ensemble method	93.61	92.45	94.59

classifiers can reduce the error rate. (2) The proposed probability-based ensemble model outperform the majority voting method. As mentioned above, this is mainly because the probability-based method can cumulate category probabilities of multi base classifiers and make prediction based on integrated information, rather than select the majority result directly.

In order to estimate generalization capability of our proposed method, experiments were also conducted on three binary classification tasks (AD vs. NC, AD vs. MCI, and MCI vs. Normal). Through the ensemble 3D-DenseNet method, encouraging accuracy results were obtained: 98.83% for AD/Normal, 93.61% for AD/MCI, 98.42% for MCI/Normal and 97.52% for AD/MCI/Normal. In order to display

the performance of the model intuitively, the confusion matrices of the ensemble model on one of the random cross-validation are shown in Fig. 11.

Tables 5–8 show the comparisons of the proposed model with previous methods on four classification tasks. N/A means that the paper did not give this evaluation index. Only accuracy are showed in Table 7, because the positive sample can not be assigned, so that the recall, precision and F1-score may not appear in multi classification. It can be seen that our proposed probability-based ensemble model perform best in four classification tasks with accuracy of 93.61% for AD/MCI, 98.42% for MCI/Normal, 98.83% for AD/Normal and 97.52% for ternary classification. Precision and recall also ahead of the state-of-the-arts, despite the recall of AD/Normal slightly lower than the DemNet [26]. This may because the DemNet are more likely to recognize the uncertain samples as positives (AD), so more true positives are recognized. But our ensemble model achieved better performance than DemNet on other challenging classification tasks. In addition, our proposed 3D-DenseNet also outperform the precious works in four classification tasks, with accuracy of 92.20% for AD/MCI, 94.12% for MCI/Normal, 95.12% for AD/Normal and 94.77% for ternary classification. Therefore, it seems that the effort of dense connections takes more effective information and gradients flow. Extracting features from a whole brain MRI and fusing probabilities of multiple classifiers

Table 6

Accuracy, recall and precision for different classification methods. These results correspond to stable MCI/Normal classification.

Methods	Accuracy (%)	Recall (%)	Precision (%)
Billones et al. [26]	91.67	92.22	91.11
Payan et al. [28]	92.11	<i>n/a</i>	<i>n/a</i>
Wee et al. [46]	83.75	83.55	83.95
3D-DenseNet	94.12	94.33	94.56
Majority voting	95.63	93.54	94.92
Proposed ensemble method	98.42	98.34	98.37

Table 7

Accuracy, recall and precision for different classification methods. These results correspond to stable AD/Normal classification.

Methods	Accuracy (%)	Recall (%)	Precision (%)
D. Cheng et al. [15]	87.15	<i>n/a</i>	86.36
Billones et al. [26]	98.33	98.89	97.78
Payan et al. [28]	95.39	<i>n/a</i>	<i>n/a</i>
Wee et al. [46]	92.35	90.35	94.31
3D-DenseNet	95.12	94.89	95.62
Majority voting	97.69	96.74	97.02
Proposed ensemble method	98.83	98.70	98.70

Table 8

Accuracy of AD/Normal/MCI classification for different classification methods.

Methods	Accuracy (%)
Billones et al. [26]	91.85
Hosseini et al. [27]	89.1
Payan et al. [28]	89.47
Cheng et al. [29]	87.15
3D-DenseNet	94.77
Majority voting	95.96
Proposed ensemble method	97.52

through the proposed ensemble 3D-DenseNet is an effective approach for diagnosis of dementia disease.

5. Conclusion

In this paper, we develop an ensemble of 3D convolutional networks involving dense connections to recognize AD and MCI using brain MR images. The distinction between AD and MCI can help to identify different categories of dementia disease and take appropriate treatments. To address the problem of limited training data, we introduce dense connections to 3D-CNN. Dense connections improve the information and gradients propagation throughout the network, then make the network easier to train due to less parameters. Extensive comparative experiments were conducted to analyze the effect of parameters on performance of 3D-DenseNet. Furthermore, each base 3D-DenseNet was constructed by varying the hyper-parameter initializations and architecture. A probability-based fusion method was used to combine the base classifiers. The ensemble model achieved obvious boosting of accuracy than doing just the simple average of the networks predictions. The experimental results demonstrate that our proposed model outperforms the previous methods in all four classification tasks.

References

- [1] A. Association, et al., 2017 alzheimer's disease facts and figures, *Alzheimer's Dementia* 13 (4) (2017) 325–373.
- [2] S. Li, O. Okonkwo, M. Albert, M.-C. Wang, Variation in variables that predict progression from MCI to AD dementia over duration of follow-up, *Am. J. Alzheimer's Dis. (Columbia, Mo.)* 2 (1) (2013) 12–28.
- [3] R. Roberts, D.S. Knopman, Classification and epidemiology of MCI, *Clin. Geriatric Med.* 29 (4) (2013) 753–772.
- [4] N. Fox, R. Black, S. Gilman, M. Rossor, S. Griffith, L. Jenkins, M. Koller, et al., Effects of A β immunization (AN1792) on MRI measures of cerebral volume in alzheimer disease, *Neurology* 64 (9) (2005) 1563–1572.
- [5] G.B. Frisoni, N.C. Fox, C.R. Jack Jr, P. Scheltens, P.M. Thompson, The clinical use of structural MRI in alzheimer disease, *Nat. Rev. Neurol.* 6 (2) (2010) 67–77.
- [6] C.R. Jack, R.C. Petersen, Y.C. Xu, P.C. O'Brien, G.E. Smith, R.J. Ivnik, B.F. Boeve, S.C. Waring, E.G. Tangalos, E. Kokmen, Prediction of AD with MRI-based hippocampal volume in mild cognitive impairment, *Neurology* 52 (7) (1999) 1397.
- [7] R. Cuingnet, E. Gerardin, J. Tessieras, G. Auzias, S. Lehéricy, M.-O. Habert, M. Chupin, H. Benali, O. Colliot, A.D.N. Initiative, et al., Automatic classification of patients with Alzheimer's disease from structural MRI: a comparison of ten methods using the adni database, *Neuroimage* 56 (2) (2011) 766–781.
- [8] F. Falahati, E. Westman, A. Simmons, Multivariate data analysis and machine learning in alzheimer's disease with a focus on structural magnetic resonance imaging, *J. Alzheimer's Dis.* 41 (3) (2014) 685–708.
- [9] E. Moradi, A. Pepe, C. Gaser, H. Huttunen, J. Tohka, A.D.N. Initiative, et al., Machine learning framework for early MRI-based Alzheimer's conversion prediction in MCI subjects, *Neuroimage* 104 (2015) 398–412.
- [10] S. Liu, S. Liu, W. Cai, S. Pujol, R. Kikinis, D. Feng, Early diagnosis of alzheimer's disease with deep learning, in: Proceedings of the IEEE 11th International Symposium on Biomedical Imaging (ISBI), IEEE, 2014, pp. 1015–1018.
- [11] H.-I. Suk, D. Shen, Deep learning-based feature representation for AD/MCI classification, in: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2013, pp. 583–590.
- [12] H.-I. Suk, S.-W. Lee, D. Shen, A.D.N. Initiative, et al., Hierarchical feature representation and multimodal fusion with deep learning for AD/MCI diagnosis, *Neuroimage* 101 (2014) 569–582.
- [13] S. Sarraf, G. Tofghi, Classification of alzheimer's disease using fMRI data and deep learning convolutional neural networks, arXiv:1603.08631(2016).
- [14] S. Ji, W. Xu, M. Yang, K. Yu, 3D convolutional neural networks for human action recognition, *IEEE Trans. Pattern Anal. Mach. Intel.* 35 (1) (2013) 221–231.
- [15] X.W. Gao, R. Hui, Z. Tian, Classification of CT brain images based on deep learning networks, *Comput. Methods Progr. Biomed.* 138 (2017) 49–56.
- [16] C. Senaras, A.C. Moberly, T. Teknos, G. Essig, C. Elmaraghy, N. Taj-Schaal, L. Yua, M.N. Gurcan, Detection of eardrum abnormalities using ensemble deep learning approaches, in: *Medical Imaging 2018: Computer-Aided Diagnosis*, 10575, International Society for Optics and Photonics, 2018, p. 105751A.
- [17] R. Rasti, M. Teshnehlab, S.L. Phung, Breast cancer diagnosis in DCE-MRI using mixture ensemble of convolutional neural networks, *Pattern Recognit.* 72 (2017) 381–390.
- [18] S.L. Risacher, A.J. Saykin, J.D. Wes, L. Shen, H.A. Firpi, B.C. McDonald, Baseline MRI predictors of conversion from MCI to probable AD in the ADNI cohort, *Current Alzheimer Res.* 6 (4) (2009) 347–361.
- [19] W. Cai, S. Liu, L. Wen, S. Eberl, M.J. Fulham, D. Feng, 3D neurological image retrieval with localized pathology-centric CMRGlc patterns, in: Proceedings of the 17th IEEE International Conference on Image Processing (ICIP), IEEE, 2010, pp. 3201–3204.
- [20] S. Liu, Y. Song, W. Cai, S. Pujol, R. Kikinis, X. Wang, D. Feng, Multifold Bayesian kernelization in Alzheimer's diagnosis, in: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2013, pp. 303–310.
- [21] D. Zhang, Y. Wang, L. Zhou, H. Yuan, D. Shen, A.D.N. Initiative, et al., Multimodal classification of Alzheimer's disease and Mild Cognitive Impairment, *Neuroimage* 55 (3) (2011) 856–867.
- [22] F. Zhang, Y. Song, S. Liu, S. Pujol, R. Kikinis, M. Fulham, D. Feng, W. Cai, Semantic association for neuroimaging classification of PET images, *J. Nucl. Med.* 55 (supplement 1) (2014) 2029.
- [23] H. Greenspan, B. van Ginneken, R.M. Summers, Guest editorial deep learning in medical imaging: Overview and future promise of an exciting new technique, *IEEE Trans. Med. Imaging* 35 (5) (2016) 1153–1159.
- [24] S. Liu, S. Liu, W. Cai, H. Che, S. Pujol, R. Kikinis, D. Feng, M.J. Fulham, et al., Multimodal neuroimaging feature learning for multiclass diagnosis of Alzheimer's disease, *IEEE Trans. Biomed. Eng.* 62 (4) (2015) 1132–1140.
- [25] F. Li, L. Tran, K.-H. Thung, S. Ji, D. Shen, J. Li, A robust deep model for improved classification of AD/MCI patients, *IEEE J. Biomed. Health Inform.* 19 (5) (2015) 1610–1616.
- [26] C.D. Billones, O.J.L.D. Demetria, D.E.D. Hostallero, P.C. Naval, Demnet: A convolutional neural network for the detection of Alzheimer's Disease and Mild Cognitive Impairment, in: Proceedings of the IEEE Region 10 Conference (TENCON), IEEE, 2016, pp. 3724–3727.
- [27] E. Hosseini-Asl, R. Keynton, A. El-Baz, Alzheimer's disease diagnostics by adaptation of 3D convolutional network, in: Proceedings of the IEEE International Conference on Image Processing (ICIP), IEEE, 2016, pp. 126–130.
- [28] A. Payan, G. Montana, Predicting Alzheimer's disease: a neuroimaging study with 3D convolutional neural networks, arXiv:1502.02506(2015).
- [29] D. Cheng, M. Liu, J. Fu, Y. Wang, Classification of MR brain images by combination of multi-CNNs for ad diagnosis, in: Proceedings of the Ninth International Conference on Digital Image Processing (ICDIP 2017), 10420, International Society for Optics and Photonics, 2017, p. 1042042.
- [30] J.J.d.M.S. Junior, A.R. Backes, O.M. Bruno, Randomized neural network based descriptors for shape classification, *Neurocomputing* (2018).
- [31] C. Hong, J. Yu, J. Wan, D. Tao, M. Wang, Multimodal deep autoencoder for human pose recovery, *IEEE Trans. Image Process.* 24 (12) (2015) 5659–5670.
- [32] J. Yu, B. Zhang, Z. Kuang, D. Lin, J. Fan, Iprivacy: image privacy protection by identifying sensitive objects via deep multi-task learning, *IEEE Trans. Inform. Forensics Secur.* 12 (5) (2017) 1005–1016.

- [33] V. Osipov, M. Osipova, Space-time signal binding in recurrent neural networks with controlled elements, *Neurocomputing* (2018).
- [34] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, J. Xiao, 3d shapenets: A deep representation for volumetric shape modeling, in: *Proceedings of the CVPR*, 1, 2015, p. 3.
- [35] C.R. Qi, H. Su, M. Nießner, A. Dai, M. Yan, L.J. Guibas, Volumetric and multi-view cnns for object classification on 3d data, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5648–5656.
- [36] J. Zhang, J. Yu, D. Tao, Local deep-feature alignment for unsupervised dimension reduction, *IEEE Trans. Image Process.* 27 (5) (2018) 2420–2432.
- [37] A. Liu, Y. Laili, Balance gate controlled deep neural network, *Neurocomputing* 320 (2018) 183–194.
- [38] C.R. Jack, M.A. Bernstein, N.C. Fox, P. Thompson, G. Alexander, D. Harvey, B. Borowski, P.J. Britson, J. L. Whitwell, C. Ward, et al., The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods, *J. Magn. Reson. Imaging* 27 (4) (2008) 685–691.
- [39] M.W. Woolrich, S. Jbabdi, B. Patenaude, M. Chappell, S. Makni, T. Behrens, C. Beckmann, M. Jenkinson, S.M. Smith, Bayesian analysis of neuroimaging data in FSL, *Neuroimage* 45 (1) (2009) S173–S186.
- [40] M. Jenkinson, P. Bannister, M. Brady, S. Smith, Improved optimization for the robust and accurate linear registration and motion correction of brain images, *Neuroimage* 17 (2) (2002) 825–841.
- [41] G. Huang, Z. Liu, K.Q. Weinberger, L. van der Maaten, Densely connected convolutional networks, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1, 2017, p. 3.
- [42] S. Ioffe, C. Szegedy, Batch normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift, in: *Proceedings of the International Conference on Machine Learning*, 2015, pp. 448–456.
- [43] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2818–2826.
- [44] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.
- [45] G. Wen, Z. Hou, H. Li, D. Li, J. Jiang, E. Xun, Ensemble of deep neural networks with probability-based fusion for facial expression recognition, *Cognit. Comput.* 9 (5) (2017) 597–610.
- [46] C.-Y. Wee, P.-T. Yap, D. Shen, A.D.N. Initiative, Prediction of Alzheimer's disease and Mild Cognitive Impairment using cortical morphological patterns, *Human Brain Map.* 34 (12) (2013) 3411–3425.
- [47] S. Wang, Y. Shen, W. Chen, T. Xiao, J. Hu, Automatic recognition of mild cognitive impairment from MRI images using expedited convolutional neural networks, in: *Proceedings of the International Conference on Artificial Neural Networks*, Springer, 2017, pp. 373–380.
- [48] F.L. Ferreira, S. Cardoso, D. Silva, M. Guerreiro, A. de Mendonça, S.C. Madeira, Improving Prognostic Prediction from Mild Cognitive Impairment to Alzheimers Disease Using Genetic Algorithms, in: *Proceedings of the International Conference on Practical Applications of Computational Biology & Bioinformatics*, Springer, 2017, pp. 180–188.



Hongfei Wang received the B.E. in Automation from Shandong Agricultural University, Tai'an, China, in 2016. He is currently pursuing the Master's degree in software engineering with the School of Software Engineering, University of Science and Technology of China, Hefei, China. He also works as a visiting student with the Institute of Advanced Computing and Digital Engineering, Shenzhen Institutes of Advanced Technology, Chinese Academy of Science, Shenzhen, China. His current research interests include machine learning, medical image computing.



Yanyan Shen received the B.S. and M. Eng. degrees in Electrical Engineering from Yanshan University, Qinhuangdao, China, in 2006 and 2009, respectively, and the Ph.D. degree in the Department of Mechanical and Biomedical Engineering from City University of Hong Kong, Hong Kong SAR, China in 2012. From 2013 to 2014, she was a postdoctoral research fellow at the School of Information and Communication Engineering, Inha University, South Korea. She is currently an associated professor in the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China. She has been the principle/co-investigator in several research projects funded by NSFC and Shenzhen Basic Research

Foundation. Her current research interests include optimization methods and machine learning in wireless networks.



Shuqiang Wang received his Ph.D. degree in System Engineering and Engineering Management from City University of Hong Kong, in 2012. He was a research scientist of Huawei Technologies Noahs Ark Lab (2012–2013). He held a postdoctoral fellowship at the University of Hong Kong from 2013 until 2014. He is currently an associate professor with the Institute of Advanced Computing and Digital Engineering, Shenzhen Institutes of Advanced Technology (SIAT), Chinese Academy of Science. His current research interests include machine learning, medical image computing, and optimization theory.



Tengfei Xiao received the M.S. degree in pattern recognition and intelligent systems from the Sun Yat-Sen University, China, in 2011 and the Ph.D. degree in systems engineering and engineering management from the City University of Hong Kong, Hong Kong, China, in 2016. Since 2016, he has been a Distinguished Associate Research Fellow with the School of Data and Computer Science, Sun Yat-Sen University, China. His research interests are focused on intelligent control and distributed parameter system.



Liming Deng received B. Sc and M. Sc degrees in Mechanic engineering in Nanchang University and Shanghai Jiao Tong University, China, in 2011 and 2014, respectively. He is a Ph.D. student in the Department of Systems Engineering and Engineering Management from City University of Hong Kong. His research interests include prognostics and health management, bioinformatics and natural language processing.



Xiangyu Wang received the B.Sc. degree from the Donghua University, Shanghai, China, in 2015. He is currently pursuing the Master's degree in software engineering with the School of Software Engineering, University of Science and Technology of China, Hefei, China. His current research interests include machine learning, medical image computing and bioinformatics.



Xinyan Zhao received the B.Sc. degree from the Hohai University, Nanjing, China, in 2016. She is currently pursuing the Ph.D. Degree with the School of Data Science, University of Science and Technology of China, Hefei, China. Her current research interests include deep learning, dialogue system and automatic question answering system.