IET Image Processing

*Research Article*

IET Journals
The Institution of Engineering and Technology

# Ensemble of deep convolutional neural networks based multi-modality images for Alzheimer's disease diagnosis

ISSN 1751-9659
Received on 21st May 2019
Revised 12th September 2019
Accepted on 17th October 2019
E-First on 19th December 2019
doi: 10.1049/iet-ipr.2019.0617
www.ietdl.org

*Xusheng Fang[1], Zhenbing Liu[2] ✉, Mingchang Xu[2]*

*[1]School of Electronic Engineering and Automation, Guilin University of Electronic Technology, Guilin 541004, People's Republic of China*
*[2]School of Computer Science and Information Security, Guilin University of Electronic Technology, Guilin 541004, People's Republic of China*
✉ *E-mail: zbliu@guet.edu.cn*

**Abstract:** Alzheimer's disease (AD) is one of the most common progressive neurodegenerative diseases. Structural magnetic resonance imaging (MRI) would provide abundant information on the anatomical structure of human organs. Fluorodeoxy-glucose positron emission tomography (PET) obtains the metabolic activity of the brain. Previous studies have demonstrated that multi-modality images could contribute to improve diagnosis of AD. However, these methods need to extract the handcrafted features that demand domain specific knowledge and image processing stage is time consuming. In order to tackle these problems, in this study, the authors propose a novel framework that ensembles three state-of-the-art deep convolutional neural networks (DCNNs) with multi-modality images for AD classification. In detail, they extract some slices from each subject of each modality, and every DCNN generates a probabilistic score for the input slices. Furthermore, a 'dropout' mechanism is introduced to discard low discrimination slices of the category probabilities. Then average reserved slices of each subject are acquired as a new feature. Finally, they train the Adaboost ensemble classifier based on single decision tree classifier with the MRI and PET probabilistic scores of each DCNN. Evaluations on Alzheimer's Disease Neuroimaging Initiative database show that the proposed algorithm has better performance compared to existing method, the algorithm proposed in this study significantly improved the classification accuracy.

## 1 Introduction

In common dementia worldwide, Alzheimer's disease (AD) has always been an important domain in psychology and medicine. It has been reported that AD has become the sixth leading cause of death in 2015 [1]. AD usually has concealed onset and advances gradually, which is diagnosed only after the patients have irreversible behavioural and cognitive impairments. However, there are no available drugs and methods to cure AD. Thus, early detection at the prodromal stage, also known as mild cognitive impairment (MCI), is very important for delaying the onset and therapy of AD.

With the rapid progress of computer technology, medical imaging technology has been unprecedentedly developed. An increasing number of medical images with different modalities contribute to computer-aided diagnosis. Different modalities provide different kinds of information to identify AD and MCI or healthy normal control (NC) [2, 3]. Structural magnetic resonance imaging (MRI) with high-resolution and contrast enhancement of soft tissue would provide more structural details [4–6]. Fluorodeoxy-glucose positron emission tomography (PET) could capture the cerebral metabolic activation of glucose [7–9]. The modal of biomarkers, such as pathological amyloid depositions measured through cerebrospinal fluid (CSF) [10, 11]. Functional magnetic resonance imaging measures functional brain activity and changes in the brain [12–14]. Many studies have demonstrated that the use of integrated information from multi-modality images could contribute to improve diagnosis of AD.

In the last decades, pattern recognition and machine learning methods have been widely used in brain disease diagnosis, which extracts different kinds of the features from neuroimaging modalities to learn a model and predict class labels on an unknown object. In general, these feature extraction methods can be summarised into four categories: voxel-based morphometry (VBM) approach, region of interest (ROI)-based approach, patch-based approach and landmark-based approach. The VBM approach is a simple and direct approach that analyses the changes in brain grey matter (GM) and white matter volume of each voxel in MRI [15–18]. Although this method is simple and easy to implement, it can easily cause the curse of dimensionality and ignorance of regional information. The ROI-based approach uses MRI to compute the volume of GM tissue with that ROI region as a feature. PET is aligned to MRI and the average intensity of ROI is computed as a feature [19–21]. However, these feature extraction methods are tedious and require expert knowledge in practice. The ROI-based segment may not adapt well to the diseased-related pathology. The patch-based approach dissects brain areas into small patches and extracts features from each selected patch, combining the features hierarchically in each classification level [22–25]. However, the disease-related structural changes occur in multiple brain regions or span several patches, so that whole brain information could not be captured by using only these independent patches. The landmark-based approach does not require non-linear image registration or brain tissue segmentation [26] reducing these two time-consuming steps. It mainly includes the landmark definition, landmark detection and extracting morphological features around the detected landmarks for AD diagnosis [27–29]. This method relies on a large training set, and limited training subjects could affect the accuracy of identifying landmarks so that final classification effect is not good [26].

Recently, the performance of deep learning has been dramatically improved in speech recognition, image recognition, natural language processing and many other domains [30–32]. In particular, the deep convolutional neural network (DCNN) and stack auto-encoder network (SAE), by stacking layers of the neuron, can utilise more abstract hierarchical feature representations of the image data. Some researchers have used SAE to mine the potential feature representations form of multi-modality images, i.e. MRI, PET and CSF independently, constructing a multi-kernel support vector machine (SVM) for classification [33, 34]. Several studies have proposed extracting features using a deep belief network [35, 36], then training classification model. Meanwhile, a large number of neuroimaging studies focus on CNN for diagnosis of AD. In particular, some classic DCNN such as VGG [37], GoogLeNet [38], ResNet [39,

40], DenseNet [41, 42] networks, obtained better classification results and were superior in AD diagnosis.

In this paper, to address the limitations of traditional machine learning and inspired by the idea of the aforementioned deep learning models, we propose a novel classification framework based on the ensemble DCNNs of multi-modality images by Adaboost learner, named DCMA for short. The main contributions are as follows:

(i) There is no need to registration and segment required for pre-processing multi-modality images and defined the ROIs or extracted patch images for handcrafted features. Our method directly extracts 2D slices of the original image to input DCNN network, which simplifies data pre-processing and reduce the time consumption.
(ii) By employing three different state-of-the-art DCNNs to automatically learn the hierarchical representations from the images data. We use a stacking strategy for each DCNN to improve the stability of the model and introducing a 'dropout' mechanism to discard image slices of the lower probabilistic score. We used 'dropout' in single quotes to distinguish it from the dropout method in neural networks, We acquired the average probabilistic scores of each DCNN as a new feature for subjects.
(iii) The MRI and PET probabilistic scores are fused by the Adaboost algorithm based on a single decision tree (DT) classifier for the final classification.

The remainder of this paper is organised as follows. In Section 2, we present the preliminary knowledge of the proposed method. Section 3 describes the proposed framework in detail. The experimental materials and experimental setup are presented in Section 4. Section 5 shows the experimental results and discusses the effects of the proposed method. Finally, the conclusion of this paper is drawn in Section 6.

## 2 Preliminaries

### 2.1 CNN model

CNN [41] is a special artificial neural network, its main characteristic is weight sharing and local perception. It has excellent performance in many fields, especially in image-related tasks such as image classification. Usually, a basic architecture of the CNN includes input layer, convolution layer, activation layer, batch normalisation (BN) [43] layer, dropout layer, pooling layer, fully connected layer and softmax layer.

GoogLeNet is the champion of the ImageNet ILSVRC classification challenge in 2014 [44]. It contains 22 layers of deep networks and its main contribution was to introduce the inception model [45]. The function of employing inception layers is to increase depth and width simultaneously without additional computational overhead. Another advantage is to simultaneously extract abstract features from different scales.

ResNet won first place on the ImageNet ILSVRC 2015 classification and detection task [46]. ResNet introduced the 'shortcut connections', where the outputs are added to the outputs of the stacked layers. Through residual learning, it effectively solves the notorious problem of vanishing/exploding gradients. It is easy to train the deeper neural network. Additionally, it neither adds an extra parameter nor increases computational complexity. Experimental results show that the accuracy gains from increasing depth are better than that of the previous network. It mainly includes 18/34-layers of the style ResNets and 50/101/152-layers of the style ResNets.

DenseNet [47] different from ResNet [46], adds outputs to each layer that has direct connections to all subsequent layers. Each layer obtains feature-maps from all previous layers. It can achieve feature reusing throughout the architecture, relieve the vanishing-gradient problem and substantially reduce the number of parameters. With the different number of the dense blocks, it mainly includes DenseNet-121/169/201/264 architectures.

### 2.2 Decision tree

DT learning is used to produce a tree with a strong ability to generalise unseen instances. The basic process that follows the tree-structured decision is simple and intuitive in a divide-and-conquer way. DT learning algorithms are generally the recursive processes of splitting the dataset. The key of the DT algorithm is how to select the splits. The classification and regression tree (CART) algorithm [48] is well known. It was employed for splitting selection in the information Gini index. For training set $D$, the proportion of samples of class $k$ is $p_k(k = 1, 2, \ldots, |Y|)$, suppose there is a feature $a$ with $V$ possible values $\{a^1, a^2, \ldots, a^V\}$. Equation (1) used Gini index to define the purity of $D$. It reflects the inconsistent probability of the label's category. Thus, the smaller the value of Gini($D$) is, the higher the purity of the data $D$

$$\text{Gini}(D) = \sum_{k=1}^{|Y|} \sum_{k' \neq k} p_k p_{k'} = 1 - \sum_{k=1}^{|Y|} p_k^2 \tag{1}$$

where the defined feature Gini index of $a$ is as follows:

$$\text{Gini}(D, a) = \sum_{v=1}^{V} \frac{|D^v|}{|D|} \text{Gini}(D^v) \tag{2}$$

For the candidate feature set $A$, which selected the minimum Gini index of feature-value pair for the split

$$a_* = \arg\min_{a \in A} \text{Gini\_index}(D, a) \tag{3}$$

The CART use a binary tree model that improves the computational efficiency, and could solve classification and regression tasks. It can handle both discrete and continuous values.

### 2.3 Adaboost

Boosting is a famous ensemble algorithm that is able to convert weak learners to strong learners. A weak learner is slightly better than a random guess, and a strong learner is very close to perfect performance. Adaboost (adaptive boosting) [49] is the specific implementation of the boosting algorithm. For the data distribution $D$, let $f$ denote the ground-truth function and $h$ denote weak classifier. Each iteration $t$ raises the weight of misclassified subjects so that they add up to 1/2 and lowers those of the correctly classified ones, so that they too add up to 1/2. When it is unable to do so, the algorithm takes a break; or else, it continues until a predetermined number of $T$ classifiers are generated. The different weights $\alpha_t$ for every classifier are learned, using an additive weighted combination of weak learners as follows:

$$H(x) = \sum_{t=1}^{T} \alpha_t h_t(x) \tag{4}$$

That achieved by minimising the exponential loss function as follows:

$$(h|D) = \mathbb{E}_{x \sim D}\left[e^{-f(x)h(x)}\right] \tag{5}$$

## 3 Propose framework

Fig. 1 shows the flow chart of the proposed DCMA method. Our method is an ensemble of three DCNNs models – GoogLeNet, ResNet-50 and DenseNet-121 based multi-modality (i.e. MRI and PET) by Adaboost algorithm. Algorithm 1 (see Fig. 2) summarises the detailed procedure of the DCMA algorithm.

### 3.1 DCNN fine-tuning

We first acquired the same size slice of MRI and PET individually by pre-processing and fine-tuning the DCNN model that had been pertained on the ImageNet [50] natural image dataset. In order to accommodate the DCNN for AD diagnosis, we replaced the last

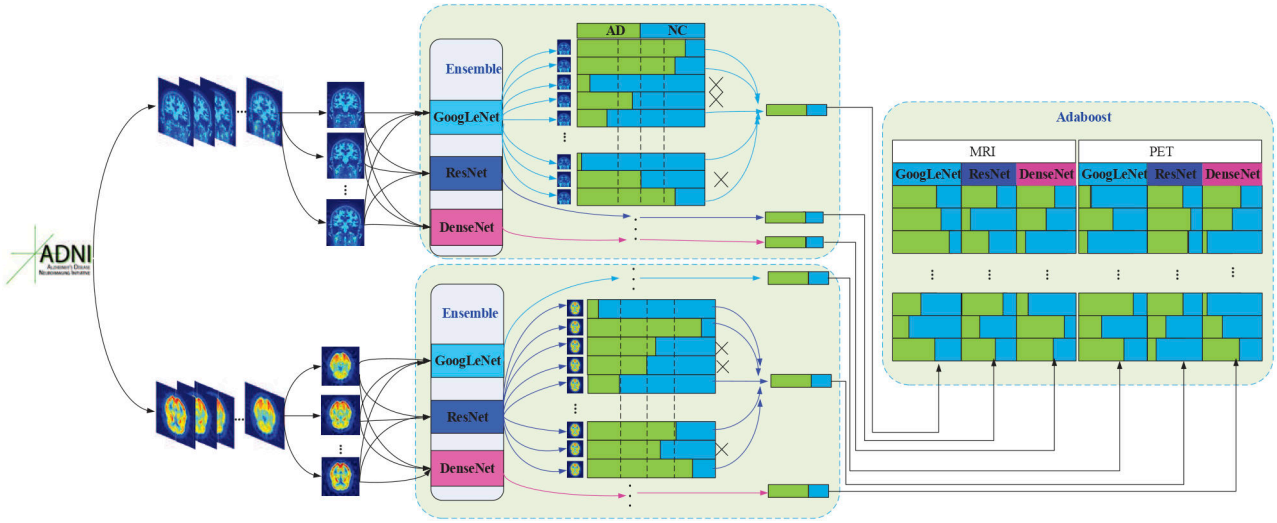


**Fig. 1** *Flow chart of the proposed DCMA method*

**Input:**
Data set $D = \{(x_1, y_1), (x_2, y_2), ..., (x_i, y_i),\ i = 1, 2, ..., N,$ $y_i \in c_k = \{-1, +1\}$, $y_i$ is a label and $x_{ij}$ is the $j$ slice of the subject $x_i$, $j = 1, 2, ..., n$. $l = 5$
First-stage acquired probabilistic score.
1: Compute probabilistic score of the input slices $p_{ij}^{Gs}(m)$, $p_{ij}^{Rs}(m)$, $p_{ij}^{Ds}(m)$
2: Average the slice probabilistic score $p_{ij}^{G}(m)$ via Eq. (6)
3: Retain the probabilistic score of the discrimination slices $\tilde{p}_{ij}^{G}(m)$ via Eq. (7)
4: Obtain the subject probabilistic score $p_i^G(m)$ via Eq. (8)
Second-stage MRI and PET fusion.
$T$ is the number of iterations and $DT$ denotes the base classifier
5: $D_1(x) = 1/N$
6: for $t = 1, 2, ..., T$
7: $h_t = DT(D, D_t)$
8: Evaluate of the error of the $h_t$ via Eq. (9)
9: If $e_t > 0.5$ then break
10: Determine the weight of the $h_t$ via Eq. (10)
11: Update the distribution $D_{t+1}(x)$ via Eq. (11)
12: end
**Output:** Finally classifier $H(x)$ via Eq. (13)

**Fig. 2** *Algorithm 1: The algorithm of DCMA*

fully connected layer for two classes. The initial DCNN filer weights that come from the ImageNet pertained model were then fine-tuned by back-propagation algorithm. We can better apply the model in AD dataset. In order to prevent over-fitting, we add BN [43] and dropout in the network.

For the train stage, especially, we construct stacking strategy for every DCNN model, schematic diagram as shown Fig. 3. We random split the training data into $l$ equal parts, one is used as the validate data and the others are used as training data in turn. Owing to the differences between the training data and testing data in every training model, which improve the stability of the model and effectively avoid the phenomenon of overfitting. The specific parameters setting are presented in Section 4.3.

### 3.2 Probabilistic score

We used the fully trained network by fine-tuning to predict the brain disease of different stages. Our proposed DCMA framework includes two stages: the first stage is acquiring a probabilistic score for the subjects by DCNN, and the second stage is the Adaboost ensemble algorithm that fuses MRI and PET. The procedural details of the DCMA algorithm are shown in Algorithm 1 (Fig. 2).

With the every DCNN, we acquired $l$ trained network models. Take the MRI modality for example, after the softmax layer, a probabilistic score for each input slices $p_{ij}^{G}(m)$ is generated, where

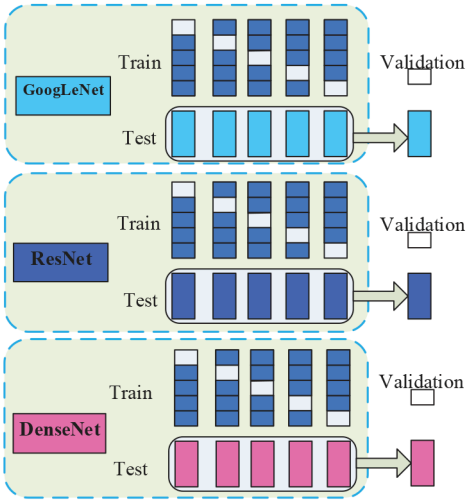


**Fig. 3** *Schematic diagram stacking strategy*

superscript G denote the GoogLeNet, R denote the ResNet, D denote the DenseNet, m denote the MRI. Averaging the $l$ times probabilistic score of each gets as the slice probabilistic score $p_{ij}^{G}(m)$ as follows:

$$p_{ij}^{G}(m) = \frac{1}{l}\sum_{s=1}^{l} p_{ij}^{Gs}(m), \quad s = 1, ..., 5. \tag{6}$$

For example, as shown in Fig. 1, in the classification of AD and NC, for the image slices, the proportion of green colour denotes a classification as an AD probabilistic score and the proportion of blue colour denotes a classification as an NC probabilistic score. The more proportion there is, the higher the probabilistic score. The dotted line in the middle denotes a probabilistic score of 0.5. The left dotted line of probabilistic score used $\alpha$ denotes and the right dotted line of probabilistic score used $\beta$ denotes. The hypothesis that the probabilistic score between $\alpha$ and $\beta$ is a low discrimination or noise. Suppose the number of slices probabilistic score between $\alpha$ and $\beta$ is $d$. In order to decrease the effects of the noise and improve the accuracy of classification, we remove the low discrimination slices of the category probabilities score. In the middle part of Fig. 1, we can see that the slices proportion of probabilistic score between $\alpha$ and $\beta$ were removed, and do not participate in the next calculation. That is, we introduce a 'dropout' mechanism and its schematic diagram as shown in Fig. 4. The discrimination image slices probabilistic score is retained, the function of retained slice probabilistic score $\tilde{p}_{ij}^{G}(m)$ is defined as follows:

$$\tilde{p}_{ij}^{\mathrm{G}}(m)=\begin{cases}p_{ij}^{\mathrm{G}}(m) & \text{if } \left(0<p_{ij}^{\mathrm{G}}(m)<\alpha\right)\text{or}\left(\beta<p_{ij}^{\mathrm{G}}(m)<1\right)\\ 0 & \text{otherwise}\end{cases} \tag{7}$$

where $\alpha+\beta=1$. All probabilistic scores of the retained slices are averaged as the subject probabilistic score $p_i^{\mathrm{G}}(m)$ is defined as

$$p_i^{\mathrm{G}}(m)=\frac{1}{n-d}\sum_{j=1}^{n-d}\tilde{p}_{ij}^{\mathrm{G}}(m) \tag{8}$$

Meanwhile, we can obtain the $p_i^{\mathrm{R}}(m)$ and $p_i^{\mathrm{D}}(m)$. The method for the PET modality is similar to that of the MRI modality, and we can obtain the $p_i^{\mathrm{G}}(p)$, $p_i^{\mathrm{R}}(p)$ and $p_i^{\mathrm{D}}(p)$.

### 3.3 Adaboost classification

Different modal images can in future learn the complementary features information to enhance classification accuracy [16, 17]. In the second stage, we fused the MRI and PET modalities by the Adaboost algorithm. In detail, from the first stage, we acquired the subject probability score of each DCNN model for each modality individually. For each subject, we combined six scores as the feature. We selected the single CART as the base learner (i.e. weak learner). The single CART is simple in construction and does not easily lead to overfitting.

First for the train set $D$ with $N$ subjects, as shown in Algorithm 1 (Fig. 2), the sample distribution, $D_t(x)$ essentially assigns a weight to each training subject $x_i$, $i=1,2,\ldots,N$, from all training data $D$ are drawn for each consecutive classifier (hypothesis) $h_t$. The distribution is initialised to be uniform; hence, all subjects have equal probability to be drawn into the first training dataset. The training error $e_t$ of classifier $h_t$ is then evaluated as the sum of these distributing weights of the subjects misclassified by $h_t$

$$e_t=P_{x\sim D_t}(h_t(x)\neq f(x)) \tag{9}$$

The algorithm requires that this error be <0.5. If $e_t>0.5$, the iteration is stopped, if $e_t<0.5$, it continues until predetermined $T$ classifiers are generated. Then, determine the weights of $h_t$ as $a_t$ are learned as follows:

$$a_t=\frac{1}{2}\ln\left(\frac{1-e_t}{e_t}\right) \tag{10}$$

Then sampling distribution $D_{t+1}(x)$ is updated as follows:

$$D_{t+1}(x)=\frac{D_t(x)}{Z_t}\times\begin{cases}\exp(-a_t) & \text{if } h_t(x)=f(x)\\ \exp(a_t) & \text{if } h_t(x)\neq f(x)\end{cases} \tag{11}$$

where the $Z_t$ is a normalisation factor which enables $D_{t+1}$ to be a proper distribution, the weights of the misclassified subjects are effectively increased. The $Z_t$ formulation is as follows:

$$Z_t=\sum_{i=1}^{N}D_t\exp(-a_t f(x)h_t(x)) \tag{12}$$

Finally, with each new classifier added to the ensemble, an additive weighted combination of weak learners as strong learner $H(x)$ is defined as

$$H(x)=\text{sign}\left(\sum_{t=1}^{T}a_t h_t(x)\right) \tag{13}$$

### 3.4 Discussion

In this part, compared with some existing DCNN-based approaches, we have introduced the advantages of the DCMA method. The DCMA method does not need to segment images to GM or extracted GM patches and directly uses slices of the original image as the input of the DCNN network compared with the existing DCNN-based methods [39, 41], which simplifies the stage of data pre-processing and reduce the time consumption. We employed the stacking trick, similar to cross-validation, which effectively avoids over-fitting and improve the stability of the model. Ensemble methods [51] that train multiple learners and then combine them for use are usually significantly more accurate than a single learner. In our method, we ensembled three different state-of-the-art DCNN models. Compared with the only ensemble single network of DenseNet style [41], the DCMA method acquired a better result. We exploited the 'dropout' mechanism for the probabilistic score of a slice to discard the slices of low discrimination to that increase the robustness of the model. In the final stage, different from the single modal methods [4, 7], the DCMA model combines the multi-modality (i.e. MRI and PET) probabilistic score of subjects with the Adaboost method rather than using a simple equal-weighted method [23, 52]. Due to the above advantages, the experiment results (Section 5) show that the DCMA method is able to significantly improve the accuracy of classification.

## 4 Materials and experimental setup

### 4.1 Dataset

We used the dataset acquired from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (www.loni.usc.edn). In total, 398 subjects were used. Although there are more than 800 subjects in the ADNI database, only 398 subjects had baseline data containing the modalities of both MRI and PET [23]. In particular, it includes 93 AD subjects, 204 MCI subjects and 101 NC subjects. The demographics of the subjects and the general exclusion criteria are detailed in Table 1.

We used 1.5T T1-weighted MRI, in NIFTI format downloaded from the ADNI. Details include: 256 × 256 × 166 voxel and 192 × 192 × 160 voxel, and 1.2 mm slice thickness. The PET image with fluorine 18 (18F) fluorodeoxyglucose were acquired 30–60 min post-injection, smoothed, averaged, spatially aligned, AC-PC orient baseline corrected, interpolated to a standard voxel size and intensity normalised.

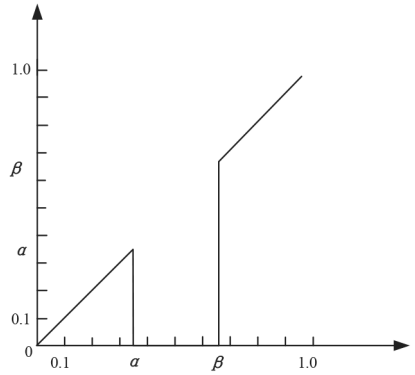


**Fig. 4** *Schematic diagram of the 'dropout' mechanism*

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

| | AD | | | MCI | | | NC | | |
|---|---|---|---|---|---|---|---|---|---|
| | Mean | SD | Range | Mean | SD | Range | Mean | SD | Range |
| age | 75.49 | 7.4 | 55–58 | 74.94 | 7.2 | 55–89 | 75.39 | 4.8 | 62–78 |
| education | 14.66 | 3.2 | 4–20 | 15.75 | 2.9 | 7–20 | 15.83 | 3.2 | 7–20 |
| MMSE | 23.45 | 2.1 | 18–27 | 27.18 | 1.7 | 24–30 | 28.93 | 1.1 | 25–30 |
| CDR | 0.8 | 0.25 | 0.5–1 | 0.5 | 0.03 | 0–0.5 | 0 | 0 | 0 |

MMSE: mini-mental state examination, CDR: clinical dementia rating.

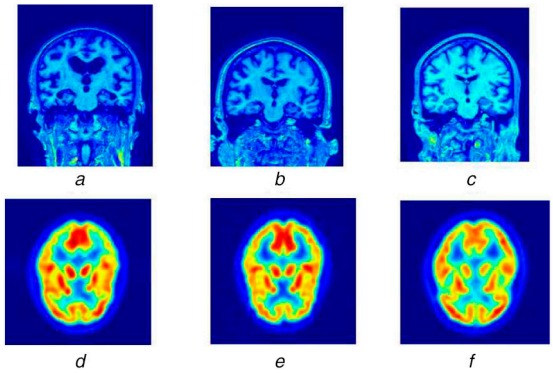


**Fig. 5** *Examples of different modal of different stages. Category of the MRI (a–c) and category of the PET (d–f)*
***(a)*** Alzheimer's disease, ***(b)*** Mild cognitive impairment, ***(c)*** Normal control, ***(d)*** Alzheimer's disease, ***(e)*** Mild cognitive impairment, ***(f)*** Normal control

**Table 2** Performance comparison of the different methods for AD versus NC

| Modal | Model | Acc | Sen | Spe |
|---|---|---|---|---|
| MRI | GoogLeNet | 95.05 | 94.61 | 96.28 |
| | ResNet | 96.64 | 94.18 | 96.63 |
| | DenseNet | 96.79 | 94.67 | 96.25 |
| | Ensemble | 98.16 | 95.74 | 97.77 |
| | D-Ensemble | 98.58 | 98.26 | 98.30 |
| PET | GoogLeNet | 90.23 | 84.81 | 92.08 |
| | ResNet | 91.58 | 92.65 | 93.03 |
| | DenseNet | 91.97 | 93.33 | 94.69 |
| | Ensemble | 93.51 | 94.59 | 95.17 |
| | D-Ensemble | 94.56 | 95.58 | 95.21 |
| MRI + PET | DCMA | 99.27 | 95.89 | 98.72 |

### *4.2 Experimental platform and pretreatment*

Our hardware of experimental environment is a desktop PC equipped with Inter core i7, 8 GB memory and GPU with 16G NVIDIA P100 × 8. The software for the environment is MATLAB 2014b with statistical parametric mapping (SPM12) for image pre-processing and an Ubuntu 16.04 system with a Tensorflow and Keras framework for training and testing. The Python programming language was used.

Based on MATLAB 2014b with SPM software, the steps for pre-processing were as follows: (i) motion correction and conformation, (ii) non-uniform intensity normalisation, (iii) Talairach transform computation, (iv) intensity normalisation and (v) resliced to 192 × 192 × 160. After the pre-processing, from the coronal view, the size of the image was 192 × 160 and there were 192 slices. For these slices, only the slices with indices 92–107 were used in the study, on account that these slices included the important regions of whole brain information. The slices were resized to 224 × 224 and conversion to a 3-channel pseudo-colour in jpg format. Similarly, slices of the PET modality with indices of 35 to 60 were used in the study. The slices were pre-processed as shown in Fig. 5.

### *4.3 Experimental setup*

The three DCNNs adopt the same preferences. Detailed parameters are as follows: (i) batch size is set to 64; (ii) the initial learning rate is set to $1 \times 10^{-3}$, and the decay rate is set to $1 \times 10^{-6}$; (iii) the number of iterations is set to 100; (iv) training the network with SGD plus momentum and using a weight decay of $1 \times 10^{-4}$ and momentum of 0.9; (v) loss function adopts the 'categorical-crossentropy'; (vi) dropout rate is set to 0.5; (vii) the $l$ is set to 5. In the experimental data, 65% of the data were randomly selected for training, 10% of the data were randomly selected for validation data and the remaining 25% of subjects were used as the test data. We repeated experiments for each classification problem ten times. The final result is an average of ten times. For the test stage, $\alpha$ is set to 0.35 and $\beta$ is set to 0.65.

For the Adaboost we use the 'scikit-learn' machine learning package, and based on classification, we selected single CART; the number of weak learners is set to 200, and the learning rate is set to 1. The algorithm of boosting selected 'SAMME.R'.

## 5 Experimental results

We validated the effectiveness of DCMA on 398 ADNI participants using the corresponding MRI and PET. Therefore, we considered three binary classification experiments: AD versus NC, AD versus MCI and MCI versus NC. The proposed method was compared with three individual DCNN models, GoogLeNet, ResNet-50 and DenseNet-121, for a single modality. We also compared the effective mechanism of the 'dropout' and ensemble three DCNN model using the Adaboost ensemble method combined with the MRI and PET probabilistic scores.

### *5.1 Assessment criteria*

We evaluated the performance of different methods by classification accuracy (Acc), sensitivity (Sen), the specificity (Spe) and area under the receiver operating characteristic curve (ROC), where TP, FN, TN and FP represent true positives, false negatives, true negatives and false positives, respectively

$$\text{Acc} = \frac{\text{TP}}{\text{TP} + \text{FN} + \text{TN} + \text{FP}} \tag{14}$$

$$\text{Sen} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{15}$$

$$\text{Spe} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{16}$$

### *5.2 Results of AD versus NC classification*

In the classification result of AD versus NC, as presented in Table 2, for the MRI modal, the ensemble method showed a better accuracy of classification than any independent DCNN model (i.e. GoogLeNet, ResNet and DenseNet), which was improved by 3.11, 1.52 and 1.37%, respectively, compared to the best performances (DenseNet) among the three independent DCNN. It is clear that the ensemble method outperforms the independent DCNN. By employing the 'dropout' mechanism with ensemble (D-Ensemble), the accuracy of classifying was improved by 0.42% compared with the ensemble method, which achieved 98.58% accuracy. It shows the effectiveness of the 'dropout' mechanism to enhance the classification accuracy. In the same way, for the PET modality, the

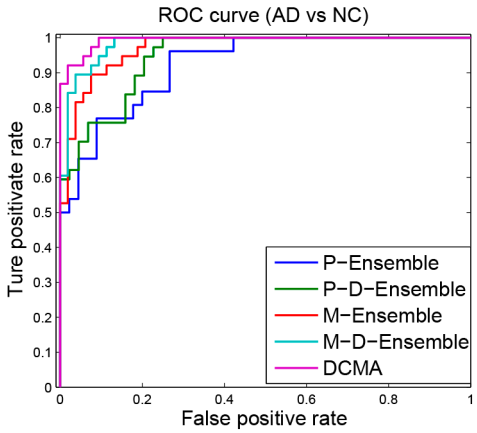


**Fig. 6** *ROC curves of different methods for classification of AD versus NC*

**Table 3** Performance comparison of the different methods for AD/NC versus MCI

| Modal | Model | AD versus MCI | | | MCI versus NC | | |
|---|---|---|---|---|---|---|---|
| | | Acc | Sen | Spe | Acc | Sen | Spe |
| MRI | GoogLeNet | 86.32 | 85.26 | 87.91 | 86.62 | 80.24 | 89.39 |
| | ResNet | 86.57 | 85.61 | 89.25 | 87.89 | 84.06 | 90.31 |
| | DenseNet | 87.64 | 85.64 | 89.71 | 88.24 | 83.27 | 90.16 |
| | Ensemble | 88.89 | 89.16 | 90.10 | 88.73 | 86.17 | 91.58 |
| | D-Ensemble | **89.98** | **89.22** | **90.67** | **88.93** | **86.33** | **91.88** |
| PET | GoogLeNet | 80.26 | 78.35 | 80.14 | 80.31 | 75.36 | 82.21 |
| | ResNet | 81.61 | 79.98 | 81.56 | 81.59 | 77.28 | 83.10 |
| | DenseNet | 82.25 | 80.29 | 83.27 | 81.34 | 78.64 | 83.19 |
| | Ensemble | 83.69 | 80.31 | 85.28 | 82.49 | 80.51 | 84.08 |
| | D-Ensemble | **85.00** | **81.75** | **87.55** | **84.21** | **82.46** | **85.82** |
| MRI + PET | DCMA | **92.57** | **89.71** | **93.59** | **90.35** | **88.36** | **92.56** |

ensemble method also showed a better classification accuracy than any other independent DCNN model and improved by 3.28, 1.93 and 1.54%, respectively. The classification accuracy of the D-Ensemble model was improved by 1.05% compared with the ensemble method. Finally, combining the MRI and PET modalities, the accuracy of classification under the DCMA method achieved 99.27%, sensitivity achieved 95.89% and specificity achieved 98.72%, which showed the best classification performance. The accuracy was improved by 0.69 and 4.71% compared to the best performance among the same method with single modality of MRI and PET, respectively. Fig. 6 further shows the ROC curves of the different methods for AD classification (prefixes with 'P' or 'M' for their modal acronym). The ROC curves of the D-Ensemble method are higher than the ensemble method and show the effectiveness of the 'dropout' mechanism. The ROC curves of DCMA contain all other curves indicating excellent diagnostic power. Based on these results, we believe that the proposed ensemble and 'dropout' method achieved the best results, outperforming all the other methods.

### *5.3 Results of AD/NC versus MCI classification*

The classification results of AD versus MCI and MCI versus NC produced by different methods are listed in Table 3. From Table 3, the DCMA method consistently achieved the best performance than other methods for the classification between AD/NC and MCI. Specifically, for AD versus MCI, the ensemble method outperformed for three independent models of DCNN, which were improved by 2.57, 2.32 and 1.25%, respectively. For the single-modality method, the D-Ensemble method showed the best accuracy of 89.98% on MRI and 85.00% on PET, which was improved by 1.09 and 1.31%, respectively. The sensitivity and specificity also showed the best results of 89.22 and 90.67% on MRI. That explained the effectiveness of the 'dropout' mechanism to enhance the classification result. The improvement of accuracy through the proposed DCMA was 2.59 and 7.57%, respectively, compared to the best performances among the competing method with individual modality. The DCMA method achieved a sensitivity up to 89.71% and a specificity up to 93.59%, also outperformed than the single modality. We also compare the ROC curves of the other methods on classification AD versus MCI and MCI versus NC problems in Fig. 7. Fig. 7*a* shows the ROC curve of the different methods for AD versus MCI. For the classification of NC from MCI, the DCMA method achieved a classification accuracy of 90.35%, a sensitivity of 88.36% and a specificity of 92.56%. The right of Fig. 7*b* shows the ROC curve of the different methods for NC versus MCI. It is clear that the ROC curve of the DCMA method, closer the left-hand border and top border, demonstrates the best model performance. Based on these results, we believe that proposed ensemble and 'dropout' method achieved the excellent diagnostic power, outperforming all the other methods.

### *5.4 Comparison with existing methods*

Meanwhile, we compared the classification results of the DCMA method with some existing methods, as shown in Table 4, including the data of single modality and multi-modality of the ADNI. Suk *et al.* (2014) [23] used 93 AD subjects, 204 MCI subjects and 101 NC subjects, and derive a algorithm for unique feature representation based on the paired patches of MRI and PET with a multimodal deep Boltzmann machine. Finally, they obtained an accuracy of 95.35% with of AD versus NC. Liu *et al.* [32] proposed cascaded CNNs to learn the multi-level and multimodal features for AD classification. However, this method needs to extract image patch that damaged the integrity of ROI. Shi *et al.* (2015) [53] used 51 AD, 99 MCI and 52 NC subjects with three modalities based on stacked auto-encoder network (i.e. MRI, PET and CSF) and obtained an excellent accuracy of 98.8% for AD versus NC, 83.7% for AD versus MCI and 90.7% for MCI versus NC. It demonstrated that the multi-modality images (i.e. MRI, PET and CSF) could contribute to improving diagnosis of AD. Zhu *et al.* [20] used the same dataset with two modalities (MRI and PET) and proposed a relational regularisation feature selection method

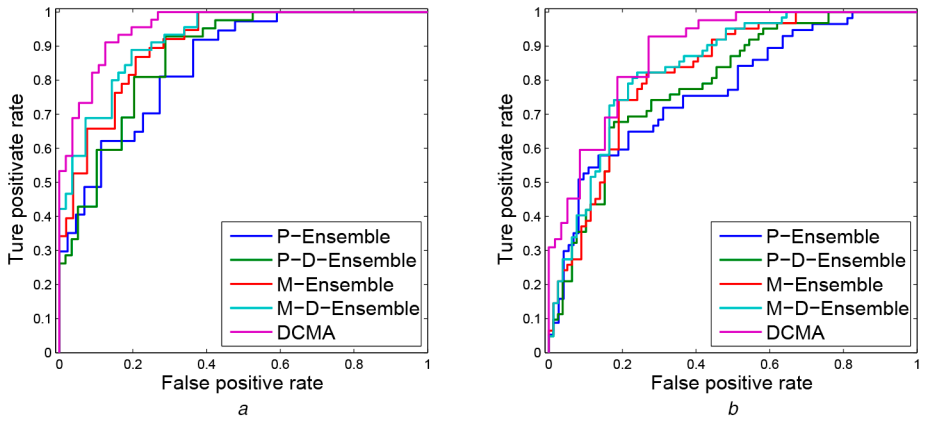


**Fig. 7** *ROC curves of different methods for classification of AD/NC versus MCI*
***(a)*** AD vs MCI, ***(b)*** NC vs MCI

**Table 4** Performance comparison of the different existing methods

| Articles | Subject | Modalities | AD versus NC | AD versus MCI | MCI versus NC |
|---|---|---|---|---|---|
| Suk *et al.* [23] | 93AD + 204MCI + 101NC | MRI + PET | 95.35 | — | 85.67 |
| Liu *et al.* [32] | 93AD + 204MCI + 100NC | MRI + PET | 93.26 | — | 74.34 |
| Heung-Il *et al.* [34] | 51AD + 99MCI + 52NC | MRI + PET + CSF | 98.8 | 83.7 | 90.7 |
| Zhu *et al.* [20] | 51AD + 99MCI + 52NC | MRI + PET | 95.7 | — | 75.9 |
| Shi *et al.* [53] | 51AD + 99MCI + 52NC | MRI + PET | 97.13 | — | 87.24 |
| Li *et al.* [24] | 199AD + 403MCI + 229NC | MRI | 89.5 | — | 73.8 |
| Altaf *et al.* [4] | 92AD + 105MCI + 90NC | MRI | 97.8 | 85.3 | **91.8** |
| Lu *et al.* [7] | 304AD + 226NC | PET | 93.58 | — | — |
| proposed | 93AD + 204MCI + 101NC | MRI + PET | **99.27** | **92.57** | **90.35** |

**Table 5** Performance comparison of the different classifier for AD versus NC

| Classifier | Acc | Sen | Spe |
|---|---|---|---|
| SVM | 98.10 | 94.54 | 96.19 |
| LR | 97.18 | 93.61 | 96.34 |
| Adaboost | **99.27** | **95.89** | **98.72** |

and reported an accuracy of 95.7% for AD versus NC and an accuracy of 75.9% for MCI versus NC. Shi *et al.* [53] exploited stacked deep polynomial networks to fusion of multimodal neuroimaging data. Li and Liu [24] proposed multiple cluster DenseNets to obtained the various local features based on MRI modal. However, the classification accuracy was less than satisfactory. Altaf *et al.* (2018) [4] used 287 subjects (92 AD + 105 MCI + 90 NC) extracted hybrid feature based MRI modality and obtain accuracies of 97.8, 85.3 and 91.8% for AD, NC and MCI, respectively. Lu *et al.* [7] exploited the deep neural network method of the PET modality and obtained an accuracy of 93.58% for AD from NC.

Table 4 indicates that our proposed method consistently outperformed the existing methods, based on these results, we firmly believe that our DCMA method achieved best performances among the competing methods, which further validates the efficacy of the DCMA method for AD diagnosis.

### *5.5 Effect of Adaboost learner*

We compared the proposed method with different classifiers such as SVM and logistic regression (LR). The LR model which use maximum likelihood to determine the parameters for two-class classification. The softmax classifier is a multi-class LR, which is generalised linear model, the extension of LR in multi-category. In this paper, we considered the three binary classifications. Table 5 summarises the classification result of competing methods for AD versus NC. The proposed method outperformed than SVM and LR methods in experiments. Our method improved the classification accuracy by 1.17% compared to SVM and 2.09% compared to LR, respectively. Meanwhile, the Adaboost method improved the classification sensitivity by 1.35% (SVM) and 1.27% (LR), respectively. The classification specificity is better than the other method. We argue that the proposed method helped enhance classification performances.

We also investigated the effect of the number of base learners on the classification performance of our proposed method. For example, AD versus NC. As seen from Fig. 8, ordinate represents error rate and the abscissa represents the number of base learner. With the increasing number of base learners, the training error and testing error gradually declines. The accuracy of classification is gradually increasing, and can achieve a relatively stable value. When the number of base learners is set to 150, the training error and the testing error tend to be stable, achieving a small value, in which the training error tends to be 0, and the testing error tends to be 0.08. This demonstrates the effectiveness of the Adaboost algorithm with the MRI and PET modalities for AD classification.

## 6 Conclusion

This paper proposed a new multi-modality data fusion and classification method based on ensemble DCNN by the Adaboost algorithm. Compared with conventional machine learning with the feature fusion method, we provided a new way to use deep learning, which employed three state-of-the-art DCNNs for the prediction of each modality, especially employing the 'dropout' mechanism to discard the low discrimination slices, and fuse their probabilistic scores by the Adaboost ensemble method for the modality data. In our lot of experiments, we validated the effectiveness of the DCMA algorithm by comparing both the single modality and multi-modality. In future work, we will continue to improve our proposed framework to predict the MCI stage, which aims to classify the MCI converters and MCI non-converters.

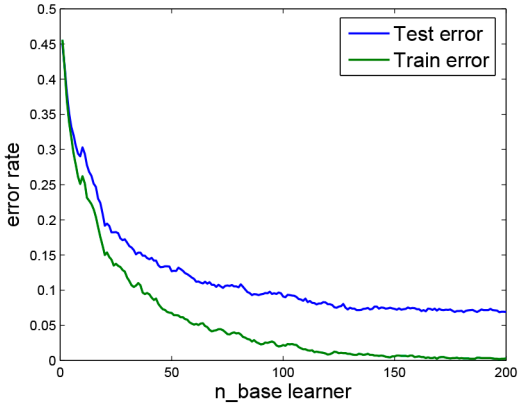


**Fig. 8** *Classification error rate with different of base classifiers*

## 7 Acknowledgments

Data acquisition and sharing were downloaded from ADNI database. This work was supported in part by the National Natural Science Foundation of China under grants (61562013 and 61866009), and the Guangxi Natural Science Foundation under grant (2017GXNFDA198025).

## 8 References

[1] Alzheimer's Association: '2018 Alzheimer's disease facts and figures', *Alzheimer's Dementia*, 2018, **14**, (3), pp. 367–429

[2] Zhang, D., Wang, Y., Zhou, L.: 'Multimodal classification of Alzheimer's disease and mild cognitive impairment', *Neuroimage*, 2011, **55**, (3), pp. 856–867

[3] Liu, S., Liu, S., Cai, W., *et al.*: 'Multimodal neuroimaging feature learning for multiclass diagnosis of Alzheimer's disease', *IEEE Trans. Biomed. Eng.*, 2015, **62**, (4), pp. 1132–1140

[4] Altaf, T., Anwar, S.M., Gul, N., *et al.*: 'Multi-class Alzheimer's disease classification using image and clinical features', *Biomed. Signal Proc. Control*, 2018, **43**, pp. 64–74

[5] Liu, M., Zhang, J., Adeli, E., *et al.*: 'Landmark-based deep multi-instance learning for brain disease diagnosis', *Med. Image Anal.*, 2018, **43**, pp. 157–168

[6] Salim, L., Mounir, B.: 'New approach for automatic classification of Alzheimer's disease, mild cognitive impairment and healthy brain magnetic resonance images', *Inst. Eng. Technol.*, 2014, **1**, (1), pp. 32–36

[7] Lu, D., Popuri, K., Ding, G.W., *et al.*: 'Multiscale deep neural networks based analysis of FDG-PET images for the early diagnosis of Alzheimer's disease', *Med. Image Anal.*, 2018, **46**, pp. 26–34

[8] Nordberg, A., Rinne, J.O., Kadir, A., *et al.*: 'The use of pet in Alzheimer disease', *Nat. Rev. Neurol.*, 2010, **6**, (2), pp. 78–87

[9] Ding, Y., Sohn, J.H., Kawczynski, M.G., *et al.*: 'A deep learning model to predict a diagnosis of Alzheimer disease by using 18F-FDG PET of the brain', *Radiology*, 2018, **290**, (2), pp. 456–464

[10] Tong, T., Gray, K., Gao, Q., *et al.*: 'Multi-modality classification of Alzheimer's disease using nonlinear graph fusion', *Pattern Recognit.*, 2017, **63**, pp. 171–181

[11] Corinna, B., Howard, C., Ronald, K.: 'Multimodal discrimination between normal aging, mild cognitive impairment and Alzheimer's disease and prediction of cognitive decline', *Diagnostics*, 2018, **8**, (1), pp. 14–34

[12] Cao, B., Zhan, L., Kong, X., *et al.*: 'Identification of discriminative subgraph patterns in fMRI brain networks in bipolar affective disorder', *Lect. Notes Comput. Sci.*, 2015, **9250**, pp. 105–114

[13] Lu, S., Xia, Y., Cai, W., *et al.*: 'Early identification of mild cognitive impairment using incomplete random forest-robust support vector machine and FDG-PET imaging', *Comput. Med. Imaging Graph.*, 2017, **60**, pp. 35–41

[14] Cai, L., Wei, X., Wang, J., *et al.*: 'Reconstruction of functional brain network in Alzheimer's disease via cross-frequency phase synchronization', *Neurocomputing*, 2018, **314**, pp. 490–500

[15] Lehmann, M., Crutch, S.J., Ridgway, G.R., *et al.*: 'Cortical thickness and voxel-based morphometry in posterior cortical atrophy and typical Alzheimer's disease', *Neurobiol. Aging*, 2011, **32**, (8), pp. 1466–1476

[16] Imabayashi, E., Matsuda, H., Tabira, T., *et al.*: 'Comparison between brain CT and MRI for voxel-based morphometry of Alzheimer's disease', *Brain Behav.*, 2013, **3**, (4), pp. 487–493

[17] Tsao, S., Gajawelli, N., Zhou, J., *et al.*: 'Feature selective temporal prediction of Alzheimer's disease progression using hippocampus surface morphometry', *Brain Behav.*, 2017, **7**, (7), p. e00733

[18] Lan, R., Zhou, Y.: 'Medical image retrieval via histogram of compressed scattering coefficients', *IEEE. J. Biomed. Health. Inform.*, 2018, **21**, (5), pp. 1338–1346

[19] Zu, C., Jie, B., Liu, M., *et al.*: 'Label-aligned multi-task feature learning for multimodal classification of Alzheimer's disease and mild cognitive impairment', *Brain Imaging Behav.*, 2016, **10**, (4), pp. 1148–1159

[20] Zhu, X., Suk, H.-I., Wang, L., *et al.*: 'A novel relational regularization feature selection method for joint regression and classification in ad diagnosis', *Med. Image Anal.*, 2017, **38**, (6), pp. 205–214

[21] Zhang, D., Shen, D.: 'Multi-modality multi-task learning for joint prediction of multiple regression and classification variables in Alzheimer's disease', *Neuroimage*, 2012, **59**, (2), pp. 895–907

[22] Liu, M., Zhang, D., Shen, D.: 'Hierarchical fusion of features and classifier decisions for Alzheimer's disease diagnoses', *Hum. Brain Mapp.*, 2014, **35**, (4), pp. 1305–1319

[23] Suk, H.I., Lee, S.W., Shen, D.: 'Hierarchical feature representation and multimodal fusion with deep learning for ad/mci diagnosis', *NeuroImage*, 2014, **101**, pp. 569–582

[24] Li, F., Liu, M.: 'Alzheimer's disease diagnosis based on multiple cluster dense convolutional networks', *Comput. Med. Imaging Graph.*, 2018, **70**, pp. 101–110

[25] Liu, M., Zhang, D., Shen, D.: 'Ensemble sparse classification of Alzheimer's disease diagnosis', *Neuroimage*, 2012, **10**, (4), pp. 1106–1116

[26] Zhang, J., Gao, Y., Gao, Y., *et al.*: 'Detecting anatomical landmarks for fast Alzheimer's disease diagnosis', *IEEE Trans. Med. Imaging*, 2016, **35**, (12), pp. 2524–2533

[27] Liu, M., Zhang, J., Nie, D., *et al.*: 'Anatomical landmark based deep feature representation for MR images in brain disease diagnosis', *IEEE. J. Biomed. Health. Inform.*, 2018, **22**, (5), pp. 1476–1485

[28] Lan, R., He, J., Wang, S., *et al.*: 'Integrated chaotic systems for image encryption', *Signal Process.*, 2018, **147**, (5), pp. 133–145

[29] Zhang, J., Liu, M., Shen, D.: 'Detecting anatomical landmarks from limited medical imaging data using two-stage task-oriented deep neural networks', *IEEE Trans. Image Process.*, 2017, **26**, (10), pp. 4753–4764

[30] LeCun, Y., Bengio, Y., Hinton, G.: 'Deep learning', *Nature*, 2015, **521**, (7553), pp. 436–444

[31] Ker, J., Wang, L., Rao, J., *et al.*: 'Deep learning applications in medical image analysis', *IEEE Access*, 2018, **6**, pp. 9375–9389

[32] Liu, M., Cheng, D., Wang, K., *et al.*: 'Multi-modality cascaded convolutional neural networks for Alzheimer's disease diagnosis', *Neuroinformatics*, 2018, **16**, pp. 295–308

[33] Suk, H.I., Shen, D.: 'Deep learning-based feature representation for AD/MCI classification', *Med. Image Comput. Comput. Assist. Interv.*, 2013, **16**, (2), pp. 583–590

[34] Suk, H.I., Lee, S.W., Shen, D.: 'Latent feature representation with stacked auto-encoder for AD/MCI diagnosis', *Brain Struct. Funct.*, 2015, **220**, (2), pp. 841–859

[35] Ortiz, A., Munilla, J., Górriz, J.M., *et al.*: 'Ensembles of deep learning architectures for the early diagnosis of the Alzheimer's disease', *Int. J. Neural Syst.*, 2016, **26**, (7), p. 1650025

[36] Suk, H.I., Wee, Y.C., Lee, S.W., *et al.*: 'State-space model with deep learning for functional dynamics estimation in resting-state fMRI', *NeuroImage*, 2016, **129**, pp. 292–307

[37] Billones, C.D., Demetria, O.J.L.D., Hostallero, D.E.D., *et al.*: 'DemNet: a convolutional neural network for the detection of Alzheimer's disease and mild cognitive impairment'. 2016 IEEE Region 10 Conf. (TENCON), November 2016, pp. 3724–3727

[38] Sarrfa, S., Tofighi, G.: 'Classification of Alzheimer's disease using fMRI data and deep learning convolutional neural networks', March 2016. Available at https://doi.org/10.1101/070441

[39] Faroop, A., Anwar, S.M., Awais, M., *et al.*: 'A deep CNN based multi-class classification of Alzheimer's disease using MRI'. 2017 IEEE Int. Conf. on Imaging Systems and Techniques (IST), Beijing, 2017, pp. 1–6

[40] Yang, C., Rangarajan, A., Ranka, S.: 'Visual explanations from deep 3D convolutional neural networks for Alzheimer's disease classification', July 2018

[41] Islam, J., Zhang, Y.: 'Brain MRI analysis for Alzheimer's disease diagnosis using an ensemble system of deep convolutional neural networks', *Brain. Inform.*, 2018, **5**, (2), p. 2

[42] Hongfei, W., Yanyan, S., Shuqiang, W.: 'Ensemble of 3D densely connected convolutional network for diagnosis of mild cognitive impairment and Alzheimer's disease', *Neurocomputing*, 2019, **333**, pp. 145–156

[43] Ioffe, S., Szegedy, C.: 'Batch normalization: accelerating deep network training by reducing internal covariate shift', In: CoRR, December 2015, pp. 189–193. Available at http://arxiv.org/abs/1502.03167

[44] Szegedy, C., Liu, W., Jia, Y., *et al.*: 'Going deeper with convolutions'. IEEE Conf. Computer Vision and Pattern Recognition, Boston, MA, USA, October 2015, pp. 1–9

[45] Lin, M., Chen, Q., Yang, S.: 'Network in network', 2013. Available at https://arxiv.org/abs/1312.4404
[46] He, K., Zhang, X., Ren, S., *et al.*: 'Deep residual learning for image recognition'. Proc. IEEE Conf. Computer Vision Pattern Recognition, Las Vegas, Nevada, USA, 27–30 June 2016, pp. 770–778
[47] Huang, G., Liu, Z., Weinberger, K.Q., *et al.*: 'Densely connected convolutional networks', August 2016. Available at https://arxiv.org/abs/1608.06993
[48] Utgoff, P.E.: 'Incremental intuction of decision tree', *Mach. Learn.*, 1989, **4**, (2), pp. 161–186
[49] Freund, Y., Schapire, R.E.: 'A decision-theoretic generalization of on-line learning and an application to boosting', *J. Comput. Syst. Sci.*, 1997, **55**, (1), pp. 119–139
[50] Russakovsky, O., Deng, J., Su, H., *et al.*: 'Imagenet large scale visual recognition challenge', *Int. J. Comput. Vis.*, 2015, **115**, (3), pp. 211–252, doi:10.1007/s11263-015-0816-y
[51] Rokach, L.: 'Ensemble-based classifiers', *Artif. Intell. Rev.*, 2010, **33**, (1), pp. 1–39
[52] Lan, R., Lu, H., Zhou, Y., *et al.*: 'An LBP encoding scheme jointly using quaternionic representation and angular information', *Neural Comput. Appl.*, 2019, **45**, p. 1984
[53] Shi, J., Zheng, X., Li, Y., *et al.*: 'Multimodal neuroimaging feature learning with multimodal stacked deep polynomial networks for diagnosis of Alzheimer's disease', *IEEE. J. Biomed. Health. Inform.*, 2018, **22**, pp. 173–183