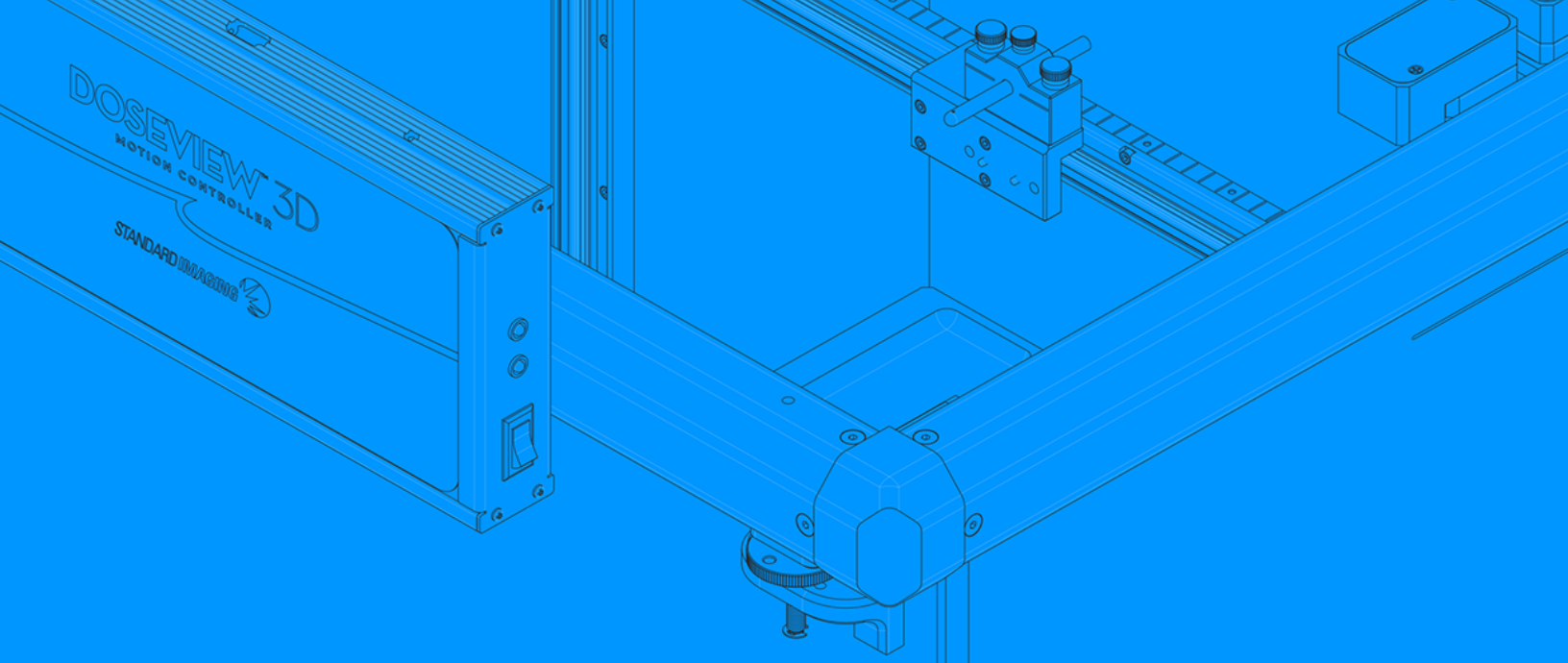


IT'S YOUR TIME
BE PRECISE

STANDARD IMAGING

Every day we spend **our time** optimizing ways to make QA easy and reliable.

Ask us how our solutions can benefit you.

WWW.STANDARDIMAGING.COM

Ensemble of Convolutional Neural Networks and Multilayer Perceptron for the Diagnosis of Mild Cognitive Impairment and Alzheimer's Disease

Minglei Li[1], Yuchen Jiang[1], Xiang Li[1], Shen Yin[2,*], Hao Luo[1,*]
[1]Department of Control Science and Engineering, Harbin Institute of Technology, Harbin 150001, Heilongjiang, China
[2]Department of Mechanical and Industrial Engineering, Norwegian University of Science and Technology, Trondheim 7033, Norway

Shen Yin. Email: shen.yin@ntnu.no
Hao Luo. Email: hao.luo@hit.edu.cn

Abstract

Background: Structural magnetic resonance imaging (sMRI) can provide morphological information about the structure and function of the brain in the same scanning process. It has been widely used in the diagnosis of Alzheimer's disease (AD) and mild cognitive impairment (MCI).

Purpose: To capture the anatomical changes in the brain caused by AD/MCI, deep learning-based MRI image analysis methods have been proposed in recent years. However, it is observed that the performance of most existing methods is limited as they only construct a single type of deep network and ignore the significance of other clinical information.

Methods: To make up for these defects, an ensemble framework that incorporates three types of dedicatedly-designed convolutional neural networks (CNNs) and a multilayer perceptron (MLP) network is proposed, where three CNNs with entropy-based multi-instance learning pooling layers have more reliable feature selection abilities. The dedicatedly-designed base classifiers can make use of the heterogeneous data, and empower the framework with enhanced diversity and robustness. In particular, to consider the interactions among the base classifiers, a novel multi-head self-attention voting scheme is designed. Moreover, considering the chance that MCI can be transformed to AD, the proposed framework is designed to diagnose AD and predict MCI conversion simultaneously, with the aid of the transfer learning technique.

Results: For performance evaluation and comparison, extensive experiments are conducted on the public dataset of the Alzheimer's Disease Neuroimaging Initiative (ADNI). The results show that the proposed ensemble framework provides superior performance under most of the evaluation metrics. Especially, the proposed framework achieves state-of-the-art diagnostic accuracy (98.61% for the AD diagnosis task, and 84.49% for the MCI conversion prediction task).

Conclusions: These promising results demonstrate the proposed ensemble framework can accurately diagnose AD patients and predict the conversion of MCI patients, which has the potential of clinical practice for diagnosing AD and MCI.

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/mp.15985

Keywords: Alzheimer's disease, magnetic resonance imaging, computer-aided diagnosis, ensemble learning, multiple instance learning

Contents

I. INTRODUCTION

Alzheimer's disease (AD) is a chronic neurodegenerative disease and contributes to 60-80% of dementias, over 30 million people around the world are diagnosed with AD[1,2]. As the most common form of dementia, AD can cause irreversible damage or destruction of neurons in brain regions over time, and gradually has a serious impact on the life of patients. Mild cognitive impairment (MCI) is often seen as a preclinical stage of AD, the predominant symptom of MCI is mild memory loss which has less impact on a person than AD[3]. Around 10% of the MCI patients worldwide develop to AD per year, while a majority of them stay stable or even revert to the normal state[4]. Those MCI patients who develop to AD are medically known as progressive MCI (pMCI), in contrast, patients who stay stable are stable MCI (sMCI). Therefore, distinguishing sMCI from pMCI has been typically considered as an early prediction of AD dementia. In particular, because there is no effective treatment to cure AD, reliable early diagnosis is crucial for the control of AD. And early diagnosis will help for the better targeted selection of individuals with MCI, thus allowing early implementation of treatment strategies and altering the course of this disease[5].

Various biomarkers (e.g., positron emission tomography (PET)[6] and MRI[7]) and biospecimens (e.g., cerebrospinal fluid (CSF)[8]) measured in vivo constitute dominant features in the diagnosis of AD. These biomarkers and biospecimens are typically employed for evaluating the development of AD, which have been well validated in many clinical settings[9]. For example, structural MRI can noninvasively capture cerebral atrophy caused by loss of neurons and dendritic pruning[10], which provides a powerful auxiliary pattern for brain research and clinical diagnosis. In addition, the clinical information of individuals can be used to partially indicate disease status, which typically includes demographic information and cognitive and neuropsychological measures. Many cognitive and neuropsychological measures, such as the Mini Mental State Examination (MMSE)[11], Clinical Dementia Rating Scale (CDRSB)[12], Alzheimer's Disease Assessment Scale (ADAS)[13], and Ray Auditory Verbal Learning Test (RAVLT)[14], etc., can reflect the cognitive level of an individual and reveal the disease progression.

Computer-aided methods have been a growing interest in the assessment and treatment of serious brain diseases, such as brain tumors[15], autism[16], and Parkinson's disease[17]. AD as one of the serious brain diseases also receives much attention. To achieve the reliable diagnosis of AD and MCI, machine learning- (ML) or deep learning- (DL) based methods have been developed in many studies based on sMRI. These existing methods include at least two main components: 1) extraction of imaging features and 2) construction of classification

Last edited *Date* :

models. According to the scale of feature extraction, these methods are usually categorized into 1) subject-level, 2) region-level, 3) patch-level and 4) slice-level[18]. The subject-level methods[19–22] extract features from voxel intensities directly, while the extracted features are high dimensional and these methods are susceptible to overfitting due to the small number of samples. The region-level methods[23–26] focus on pre-determined brain regions of structure or function, and extract representative features from these regions. Although region-level features have lower dimensions than subject-level features, they may not cover all possible pathological parts of the whole brain and miss some subtle changes in pathology. The patch-level methods[27–29] combine the above two methods, attempting to capture the disease-related pathologies in the local brain. The key step of patch-level methods is to select patches and combine them to obtain information about the brain. The slice-level methods[30,31] are closer to the diagnosis modes of physicians, which utilize 2D slice images from sMRI to extract features and then count each slice-level result to obtain a subject-level diagnosis. ML-based methods usually need to extract features manually and then construct a conventional classifier to complete diagnosis, such as support vector machine (SVM). While DL-based methods perform feature extraction and classification only by convolutional neural networks (CNNs), which have been demonstrated more powerful than ML-based methods.

In the above methods, the requirements of slice-level methods for computing resources are much lower than the use of regions, patches, or subjects. And the architectures of classifiers in slice-level methods are also simpler than other methods. In addition, the superior performance of DL-based methods often depends on numerous learnable parameters of networks. Many existing DL-based studies have been limited to using a single CNN for AD diagnosis or MCI conversion prediction. However, due to the scarcity of medical data, it is challenging for an individual CNN to achieve reliable classification with the small number of available training data.

To overcome this limitation, ensemble learning methods have been applied to the disease diagnosis, and effectively combined with the CNN[32]. There are very few works used CNN-based ensemble classifiers for AD diagnosis in recent years[33–36]. Ensemble learning is the algorithm that constructs a set of classifiers and then performs classification by aggregating their predictions[37]. And the ensemble learning methods have been proved that can enhance the reliability of diagnosis, while the main drawback of these works is that each classifier is assigned the same weight when the final results are obtained by the majority- and average-voting. These fusion methods do not perform adaptive fusion based on each classifier and may be affected by the weaker classifier in the ensemble.

I. INTRODUCTION

In this work, the target is to propose an ensemble framework that can conduct the reliable diagnosis of AD and MCI simultaneously. For clarity, the following two research tasks are defined:

1) *Task 1 (AD vs. CN)*: Distinguish between whether a subject (a patient) is cognitively normal (CN) or with AD.

2) *Task 2 (pMCI vs. sMCI)*: Distinguish between whether an MCI patient belongs to pMCI or sMCI.

The contributions of this work can be summarized as follows:

- A robust ensemble learning framework is proposed to make use of the multi-modal information/heterogeneous data. Three types of dedicatedly-designed CNNs are incorporated to exploit information from sMRI, and a shallow network (i.e., MLP) is employed to exploit the clinical information.
- A multi-head self-attention voting scheme is proposed as an ensemble approach for base classifiers. The interactions among the classifiers are considered, and the defect that common voting approaches ignore the relationships among classifiers is overcome.
- Multi-instance learning (MIL) is incorporated into base CNN classifiers. The entropy-based MIL pooling layer can reasonably consider the expressive abilities of different slices and integrate slice-level features.

II. MATERIAL AND METHODS

II.A. Data acquisition and image pre-processing

We consider a dataset obtained from ADNI-1 and ADNI-2 in the Alzheimer's Disease Neuroimaging Initiative (http://www.loni.ucla.edu/ADNI)[38]. The ADNI database is the largest publicly available Alzheimer's disease dataset and has been used in quite a few studies. Specifically, the baseline dataset contains T1-weighted MRI obtained from 771 subjects, which consists of 244 CN, 299 MCI, and 228 AD subjects. Depending on whether the MCI subjects progressed to the AD stage within 36 months after baseline assessment, they can be further divided into 170 sMCI and 129 pMCI subjects. The demographic information (age, gender, and education years), cognitive and neuropsychological measures (CDRSB, ADAS, MMSE, RAVLT) as well as the ApoE4 genotyping of the subjects are shown in Table 1.

Table 1: Information summary of the studied dataset extracted from ANDI

	Gender (M/F)	Age	Education (years)	APoE4 level			CDRSB	ADAS			MMSE	RAVLT			
				0	1	2		ADAS11	ADAS13	ADASQ4		immediate	learning	forgetting	%forgetting
CN	118/126	74.2±6.0	16.5±2.6	178	61	5	0.2±0.1	5.6±2.7	8.6±4.0	2.7±1.7	29.1±1.1	45.4±10.0	5.9±2.2	3.7±2.6	34.9±26.7
sMCI	99/71	71.8±7.4	16.2±2.9	104	53	13	1.2±0.7	8.7±3.8	13.9±5.6	4.7±2.2	28.1±1.7	37.9±11.1	4.9±2.6	4.3±2.4	50.9±30.7
pMCI	73/56	73.8±7.1	15.9±2.8	42	57	30	2.0±1.0	13.0±4.0	21.4±5.2	7.4±1.9	26.6±1.7	28.0±6.9	3.1±2.0	5.2±2.3	77.4±27.8
AD	124/104	74.9±7.8	15.2±2.9	71	115	42	4.5±1.6	19.9±6.6	30.1±7.8	8.6±1.5	23.1±2.0	22.9±7.1	2.0±1.6	4.5±1.7	88.8±21.4

*The data are presented as mean ± standard deviation (std).

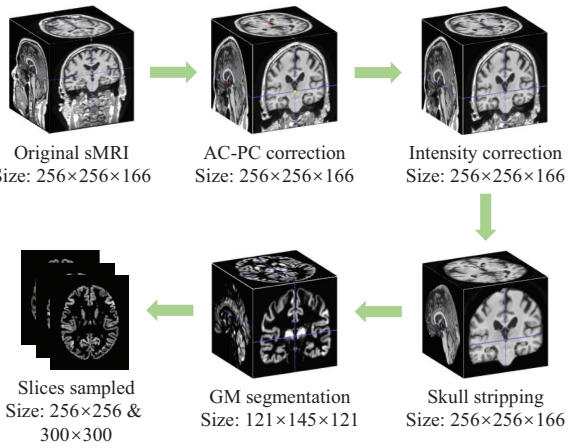


Figure 1: The preprocessing pipeline of sMRI. The pipeline includes AC-PC correction, intensity correction, skull stripping, tissue segmentation, and slice selection. Taking an sMRI with the size of 256 × 256 × 166 voxels as an example, the image size after each processing step is shown.

As shown in Fig. 1, the sMRI data go through a standard pipeline preprocessing procedure, including anterior commissure (AC)-posterior commissure (PC) correction, intensity correction, skull stripping, tissue segmentation, and slice selection. Specifically, we use the MIPAV software (https://mipav.cit.nih.gov/clickwrap.php) for AC-PC correction and adopt N3 algorithm[39] for intensity correction. Skull stripping and tissue segmentation are performed by using the CAT12 toolbox (http://dbm.neuro.uni-jena.de/cat/) via SPM12 software (http://www.fil.ion.ucl.ac.uk/spm/software/spm). Following skull stripping, the quality of the preprocessed images is checked manually. And the qualified images are then segmented to obtain the gray matter (GM) tissues, which are aligned to Montreal Neurological Institute T1 Template[40]. The GM images are smoothed with a 3.0 mm full width at half maximum (FWHM) isotropic Gaussian kernel. As a result, the sizes of obtained GM tissues are $121 \times 145 \times 121$ voxels, and the spatial resolutions are $1.5 \times 1.5 \times 1.5 mm^3$. Considering that GM is the most notably affected tissue by AD, it is used for feature extraction. Then, the 3D volumetric data is sectioned along the axial direction, and the slices are sampled from the central slice to the edges of the 3D volumetric data. The edge slices largely cover cross-sections of the brain stem, cerebellum, and cerebral cortex, which are the anatomic

areas less relevant to AD pathology. Therefore, the middle two-thirds of the slices (80 slices) are selected and resized to 256×256 and 300×300 pixels. The selected slices cover areas including ventricle, inferior temporal, and middle temporal cortices. And these areas have been reported as the regions correlated with AD pathology, which can provide rich tissue information[41].

For the clinical information, numerical normalization (i.e., Min-Max normalization) is employed to normalize the values of each separate clinical factor to the range of [0, 1].

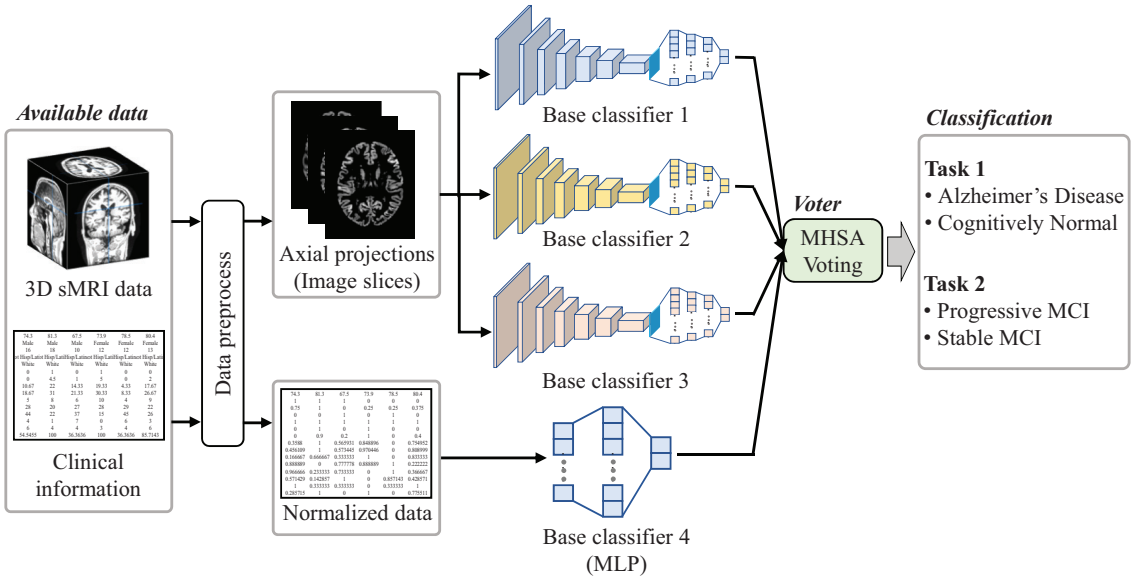


Figure 2: Illustration of the proposed ensemble framework for AD diagnosis and MCI conversion prediction. Raw 3D sMRI and corresponding clinical information of each individual are first preprocessed, multiple 2D slices are sampled from each sMRI, and the clinical information is normalized. The processed data are then fed into four different base classifiers, and an MHSA voting scheme aggregates the outputs of each base classifier for the final prediction.

II.B. Overall ensemble learning architecture

The proposed ensemble framework is illustrated in Fig. 2, where the inputs are the 3D sMRI data and clinical information, and the output is the AD diagnosis (i.e., AD or CN) or MCI conversion prediction (i.e., pMCI or sMCI). Specifically, 3D sMRI and clinical information of each individual are processed via several preprocessing steps. After that, the multiple slices sampled from 3D sMRI and normalized clinical information are as the inputs of different base classifiers. The base classifiers are designed to have different architectures, each base classifier

can play an important part in this ensemble framework. Base classifier 1, base classifier 2, and base classifier 3 are used to extract the features of images and give the initial predictions based on sMRI data, where the entropy-based multi-instance learning (MIL) pooling layer is designed to consider different information densities of slices and further improve their expression abilities. Base classifier 4 is designed as an MLP to make use of the clinical information, which can introduce different patient information than the sMRI modal. Then, four base classifiers are fused via MHSA voting to obtain the classification results for two classification tasks (i.e., AD *vs*. CN and pMCI *vs*. sMCI).

II.C. Base classifiers in the ensemble framework

In this section, the detailed architecture of each base classifier and their mentalities of designing are introduced, including three CNNs (base classifier 1, 2, and 3) and an MLP model (base classifier 4).

The architectures of base classifiers are shown in Fig. 3, all base CNN classifiers (Base classifier1, 2, and 3) have feature extraction, entropy-based MIL pooling, and classification layer three parts. Scaling up the dimension of network width, depth, and resolution has been widely used to improve the performance of networks. However, scaling up a CNN in all three dimensions of width, depth, and resolution will greatly increase the number of parameters. Considering the consumption of computing resources and the efficiency of ensemble learning, it is not necessary to design an overly complex model as one of the base classifiers. Thus, the three base CNN classifiers are scaled up in width, depth and resolution, respectively. The number of layers and the number of parameters in these classifiers are controlled. As a result, the average number of three CNNs parameters is less than that of ResNet34[42], and the layers of them are less than 19 layers. Specifically, three base CNN classifiers have different scales of network width, depth and resolution, respectively. Base classifier 1 has higher resolutions than the other two, which means that it can potentially capture more fine-grained patterns. Base classifier 2 only scales up in terms of network depth. Deeper networks can fit more complex deep features. Base classifier 3 has a wider architecture and can focus on richer features. More details of these base classifiers will be introduced as follows.

II.C.1. Entropy-based MIL pooling

AD-related pathological areas usually exist in some partial areas of the brain, and these areas in sMRI images are unlabeled, namely, only the entire sMRI image is labeled as a certain

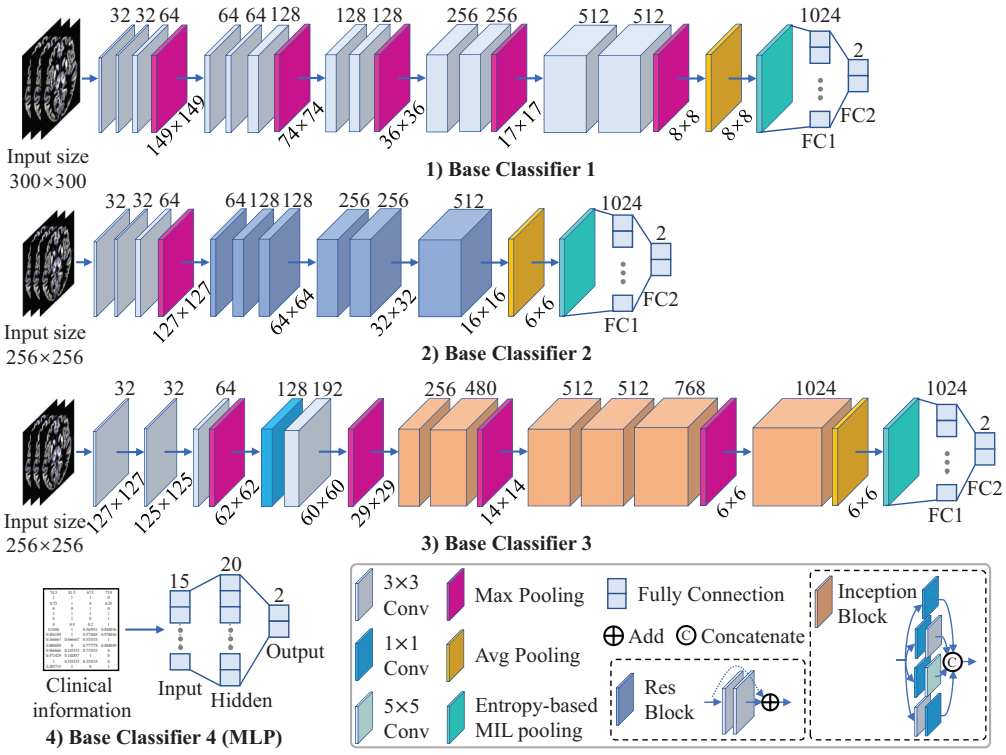


Figure 3: The architectures of base classifiers. Base classifier 1, base classifier 2, and base classifier 3 are CNN based classifiers with MRI images as inputs, which mainly consist of convolutional layers, designed special blocks (i.e., Res block, Inception block), and pooling layers. Base classifier 4 is an MLP with clinical information as inputs, and it consists of fully connected layers. The number of channels for each convolutional layer or special block is displayed above them. When the sizes of the feature maps change after passing through some layers, the sizes are shown below the convolutional layers or special blocks in the form of $H \times W$.

category. As described in Section II.A., the slices are sampled from 3D volumetric data along the axial direction and used as inputs of base CNN classifiers. These processes can be seen as the construction of bags in MIL. Considering the properties and preprocessing processes of sMRI images, both tasks in this work can be solved with the MIL strategy.

Let $X_i = \{x_{i1}, x_{i2}, ..., x_{in_i}\}$ denotes the bag of the i-th sMRI, where $x_{kl} \in \mathbb{R}^d (k = 1, 2, ..., n_k)$ represents the l-th slice of the k-th bag. Then, these slices are input into the feature extraction part of base CNN classifiers to obtain slice-level features $E_i = \{e_{i1}, e_{i2}, ..., e_{in_i}\}$, followed by a proposed entropy-based MIL pooling layer to generate embedding-level features $\mathcal{B}_i$ from slice-level features. The proposed entropy-based MIL pooling layer combines information entropy with MIL. The information entropy of an image is a statistical form of the features, which evaluates the information density of an image. In

general, the images with high entropy values have more information about target areas (e.g., brain, lung, etc.). In the clinical environment, for medical images with explicit sequences, such as MRI and CT, physicians also focus on the slices with more abundant tissue information when diagnosing diseases. Entropy as a form of reflecting image information density, combining it with MIL can not only be closer to actual clinical diagnosis, but also further improve the performance of diagnosis. This is the motivation for us to design entropy-based MIL pooling. The entropy-based MIL pooling layer can be described by the following equations:

$$\mathcal{B}_i = Concat_{l=1}^{n_i}(h_{il} \cdot e_{il}) \tag{1}$$

$$h_{il} = norm(\frac{H_{il}}{\sum_{l=1}^{n_i} H_{il}}) \tag{2}$$

where h_{il} is normalized weight that can be calculated by Eq. (2), and H_{il} in Eq. (2) is the information entropy of the l-th slice of i-th sMRI. e_{il} corresponds to the l-th slice-level features of E_i. $Concat$ is channel concatenation. In addition, mean MIL pooling and maximum MIL pooling are commonly used operators in MIL. Mean MIL pooling considers that all slices have the same ability to express the information of features, it generates embedding-level features by averaging slice-level features. Maximum MIL pooling depends on only one slice to determine the prediction of the individual. Different from these two pooling operators, entropy-based MIL pooling comprehensively considers the information entropy of different slices, which can utilize the information expression ability of these slices to achieve a more accurate diagnosis.

After obtaining embedding-level features $\mathcal{B}_i$, the classification layer is used to predict the category (i.e., AD, CN, pMCI, or sMCI) of each input sMRI.

$$P(Y|X) = f_{cls}(\mathcal{B}_i) \tag{3}$$

where $P(Y|X)$ is the probability that the subject belongs to a specific class, Y denotes the true category, $f_{cls}(\cdot)$ denotes the mapping function of the classification layer.

II.C.2. Base classifier 1

The base classifier 1 is designed to have higher resolution, and it is constructed by stacking convolutional layers without adopting more complex modules. Specifically, base classifier 1 contains twelve convolutional (Conv) layers, an entropy-based MIL pooling layer, and two fully connected (FC) layers. The number of channels for Conv layers is mainly 32, 64, 128,

256, and 512. Each Conv layer consists of one convolutional layer, batch normalization (BN), and rectified linear unit (ReLU) activation, where the convolutional layer has 3×3 kernel size, unit stride with unit zero padding. Several 3×3 max pooling layers and an adaptive average pooling layer are inserted in the specific positions of the model, which can downsample the number or depth of the intermediate feature maps. An entropy-based MIL pooling layer is inserted between the average pooling layer and FC layers. At the end, two FC layers with 1024 and 2 nodes respectively as classification layer are adopted to map distributed features into the sample label space. The input images of base classifier 1 have higher resolutions than those of base classifier 2 and base classifier 3, and the intermediate feature maps also have higher resolutions. With high resolutions, base classifier 1 tends to be more sensitive to fine-grained patterns, which can better focus on subtle pathological changes in slices.

II.C.3. Base classifier 2

The base classifier 2 with deeper depth is designed to characterize complex nonlinearities. Scaling up the depth of networks may bring gradient instability and network degradation, therefore, base classifier 2 draws on the idea of residual learning, which adopts Conv layers and residual (Res) blocks as main components. Specifically, it consists of three Conv layers, six Res blocks, an entropy-based MIL pooling layer, and two FC layers. At the beginning of the model, three Conv layers with the same composition as in base classifier 1 are used to extract shallow feature maps, where the number of channels for Conv layers is 32, 32 and 64, respectively. Then a max pooling layer merges the features and reduces their dimensions, followed by six Res blocks. As shown in Fig. 3, each Res block contains two serial Conv layers, and the output of the second Conv layer adds the input of the Res block through a shortcut connection, the result of the addition is used as the output of the Res block. The number of channels for Res blocks is 64, 128, 128, 256, 256 and 512, respectively. In order to achieve the effect of downsampling, the stride of the first Conv layer in the third, fifth and sixth Res block is respectively set to 2, other Conv layers in Res blocks have the same settings as the Conv layers in the base classifier 1. After that, the average pooling layer, MIL pooling layer, and classification layer that same as base classifier 1 are adopted. Base classifier 2 with deeper depth is designed to characterize complex nonlinearities. The Res blocks can transfer shallow feature information extracted by three Conv layers to deeper layers, thereby enhancing feature representations and strengthening their learning. Benefiting from network depth, base classifier 2 has better nonlinear representation ability, which can learn to fit more complex features and generalize well on diagnostic tasks.

II.C.4. Base classifier 3

The base classifier 3 is designed as a network with wider architecture. The suitable network width can ensure that the layers learn rich features, such as texture features in different frequencies and different directions. Base classifier 3 consists of five Conv layers, six Inception blocks, an entropy-based MIL pooling layer, and two FC layers. To maintain the proper size of feature maps, the stride of the first Conv layer is set to 2, followed by four Conv layers, where 1×1 Conv layer allows the model to control the depth of the feature more flexibly as needed. The number of channels for Conv layers is 32, 32, 64, 128 and 192, respectively. After serial Conv layers, the Inception blocks further process the extracted features. As shown in Fig. 3, each Inception block has four paths to perform convolution operations on the input and concatenates to generate the output of the block, it contains several 1×1, 3×3, and 5×5 Conv layers. The number of input channels for Inception blocks is 256, 480, 512, 512, 768 and 1024, respectively. Similar to the base classifier 1, the 3×3 max pooling layers and an adaptive average pooling layer are inserted in the specific positions to downsample the feature maps, the MIL pooling layer and classification layer are inserted at the end of the model. In base classifier 3, the maximum number of channels for blocks reaches 1024, which is twice the maximum number of the other base CNN classifiers. More channels characterize richer feature information of images, which can endow the model with better representational ability. Thus, base classifier 3 with wide architecture can potentially better learn and characterize rich tissue information in slices.

II.C.5. Base classifier 4

As summarized in Table 1, the clinical information data including age, gender, cognitive test, etc. were collected from the subjects. Since these data are not as complicated as images, shallow neural networks are enough to mine information in these clinical data. For this reason, MLP is chosen as the base classifier 4. In more detail, the MLP is composed of three layers, including an input layer, a hidden layer, and an output layer. The number of nodes for three layers is 15, 20 and 2, respectively. All layers contain one FC layer, followed by BN and ReLU activations. Since the MLP is simple in structure and with few parameters, it is suitable for clinical information data analysis.

The loss function in the proposed base classifiers for classification can be formulated as:

$$\mathcal{L}(X, Y, P, \omega_c) = -log(P(Y|X), Y) \tag{4}$$

where X denotes the input data of the base classifiers (i.e., sMRI for base CNN classifiers, clinical information data for MLP), Y denotes the corresponding true label, P denotes the predicted results, and ω_c is the learnable parameters of these classifiers.

II.D. Ensemble approaches for classifiers

The predictions from these trained base classifiers are combined by different ensemble approaches. Specifically, common voting approaches (i.e., majority voting, weighted voting, SVM-based voting) and proposed multi-head self-attention (MHSA) voting have been performed on classifiers of ensemble framework and compared. In common voting approaches, the fixed weight is assigned to each classifier in the ensemble for the aggregation of the classification results. The major drawback of these approaches is that the aggregation is not data-adaptive and ignores the interactions among base classifiers, which potentially brings bias to the final classification, especially in the presence of weak base classifiers.

Considering that common voting approaches ignore the interactions among base classifiers and potentially introduce bias resulting in unreliable predictions, an MHSA voting scheme is proposed to aggregate the results of base classifiers, which can calculate and exploit the interactions among base classifiers during their fusion. The MHSA voting is to calculate the correlation and importance among the base classifiers, and then use these interactions to aggregate the results and obtain the final classification results. It is defined as linear transformation, interaction calculation, and aggregation & final decision three stages. The proposed MHSA voting scheme is shown in Fig. 4

1) *Linear Transformation:* In this stage, the outputs of each base classifier are linearly transformed into three vectors q, k, and v, and the distribution spaces of these vectors are basically the same. Formally, an embedded representation is constructed to represent the outputs of all base classifiers. Denote the embedded representation as Φ, where $\Phi = [\phi_1, \phi_2, ..., \phi_n, ..., \phi_N]^T \in \mathbb{R}^{N\times C}$. Here, $\phi_n \in \mathbb{R}^{1\times C}(n = 1, 2, ..., N)$ indicates the outputs of the n-th base classifier, N and C are the number of base classifiers and the output dimension of each base classifier, respectively. Define Q, K and V as the set of q, k and v, respectively, where $Q = [q_1, q_2, ..., q_N]^T = \Phi \cdot W^Q \in \mathbb{R}^{N\times C}$, $K = [k_1, k_2, ..., k_N]^T = \Phi \cdot W^K \in \mathbb{R}^{N\times C}$, and $V = [v_1, v_2, ..., v_N]^T = \Phi \cdot W^V \in \mathbb{R}^{N\times C}$. Here, $W^Q \in \mathbb{R}^{C\times C}$, $W^K \in \mathbb{R}^{C\times C}$, and $W^V \in \mathbb{R}^{C\times C}$ are the weight of the linear transformation matrix.

2) *Interaction Calculation:* In the second stage, we need to score all the base classifiers based on the results of a certain classifier, and this score determines the degree of interactions

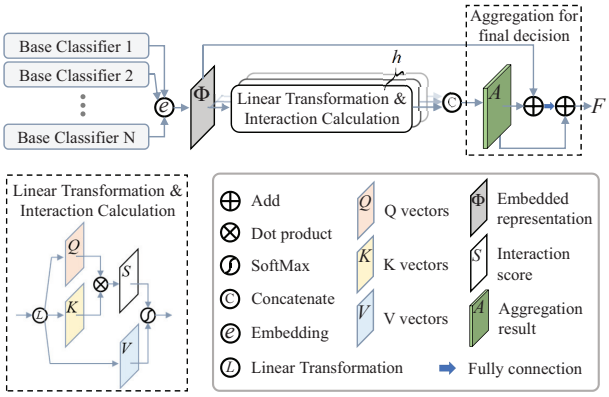


Figure 4: Illustration of the proposed MHSA voting scheme. It includes linear transformation, interaction calculation, and aggregation for the final decision three parts. The linear transformation part transforms the outputs of based classifiers into three vectors Q, K, and V. The interactions among base classifiers are calculated based on Q, K, and V by the interaction calculation part. Then, these interactions are adopted to enhance the representation and generate the final decision.

among this classifier and other base classifiers. The similarity between each pair of base classifiers is calculated by the dot product of K and Q, namely QK^T. Then a SoftMax function is used to normalize the similarity QK^T, and get an interaction score $S \in \mathbb{R}^{N\times N}$ which can reflect the interactions among the base classifiers.

$$S = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1N} \\ s_{21} & s_{22} & \cdots & s_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ s_{N1} & s_{N2} & \cdots & s_{NN} \end{bmatrix} = SoftMax(\frac{QK^T}{\sqrt{d}}) \tag{5}$$

where s_{ij} represents the interaction between the q_i and k_j, $\sqrt{d}$ can make the MHSA voting scheme have a more stable gradient flow during the training process. After that, the V is multiplied by S, which means maintaining the relationship among the associated base classifiers and reducing the impact of the less-correlated classifiers.

3) *Aggregation & Final Decision:* To learn interaction information in different representation subspaces, the above two stages are performed several times, and the results of these times are concatenated and linearly transformed.

$$A = Aggregation(Q, K, V) = Concat(S_1V, ...S_hV) \cdot W^A \tag{6}$$

where A is the aggregation result, $S_k(k = 1, ..., h)$ indicates interaction score in different representation subspaces, $W^A \in \mathbb{R}^{C\times C}$ is the linear transformation matrix, $Concat$ is channel

concatenation. Then the aggregation result is passed through the residual connection and the FC layer to enhance the representation, and get the final decision, which can be described by the following equation:

$$F = FC(A + \Phi) + A \tag{7}$$

where F is the final decision generated by the MHSA voting. The MHSA voting can achieve the modeling of the interactions among the base classifiers and fuse the outputs of each base classifier based on these interactions.

II.E. Implementations

The proposed ensemble framework is implemented based on the PyTorch deep learning library. The framework is trained on a PC with an NVIDIA GTX 1080Ti graphics card. The loss function in Eq. (4) is adopted to supervise the learning of the base classifiers parameters, which are optimized by the Adam optimizer with a low learning rate of 0.0001.

To validate the proposed framework, a series of comparison and ablation experiments are conducted. In the comparison experiments, several ML-based and DL-based methods were compared with the proposed framework to demonstrate the superiority of our framework. Since all results acquired by different methods are measured based on the same ADNI cohort, and most of these methods have similar pre-processing pipeline and implementation details to that in the proposed method, we compare our results with the reported results by the compared methods. In the ablation experiments, the effectiveness of the entropy-based MIL pooling layer and MHSA voting scheme, several studies are conducted to evaluate the influence of transfer learning and clinical information, and the indispensability of four base classifiers. More details about the implementations are as follows.

II.E.1. Data split

20% samples (154 samples) of the dataset are selected as the test samples and the remaining 80% samples (617 samples) as the training samples. A five-fold cross-validation strategy is adopted to verify the reliability of the proposed framework, in which four folds of the training samples are used for training and one fold for validation. To make sure that no significant difference in the age and gender distributions among the training, validation, and test samples, the Chi-square test is used to verify the distributions.

II.E.2. Training strategy

For task 1 (i.e., AD *vs.* CN), the base CNN classifiers are trained from scratch directly, and the parameters of them are initialized randomly. For task 2 (i.e., pMCI *vs.* sMCI), transfer learning is adopted to train the base CNN classifiers. MCI is a preclinical stage of AD, the structural changes of brains caused by MCI may be more subtle than those caused by AD, which means task 2 is more challenging than task 1. According to the development of AD, the two tasks are highly correlated, and the information learned from AD and CN subjects can be employed as a supplement to enrich the information for task 2[28,29]. Thus, the parameters of base CNN classifiers trained on task 1 are transferred to initialize the training for task 2. Early stopping is applied for all training processes, the training process is terminated when the validation loss exceeds the lower threshold in 10 continuous epochs.

II.E.3. Evaluation metrics

In two classification tasks, four evaluation metrics, namely, classification accuracy (ACC), sensitivity (SEN), specificity (SPE), and the area under the receiver operating characteristic curve (AUC) are adopted to evaluate the classification performance. These metrics are respectively defined as:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{8}$$

$$SEN = \frac{TP}{TP + FN} \tag{9}$$

$$SPE = \frac{TN}{TN + FP} \tag{10}$$

where TP denotes true positive, TN denotes true negative, FP denotes false positive, and FN denotes false negative. The ROC curve is generated according to the (SEN, 1−SPE) pairs. The AUC characterizes the classification performance of the methods, the performance is better when AUC is closer to 1.

III. RESULTS

III.A. Comparison with other methods

To demonstrate the superiority of the proposed ensemble framework, we compare the results on two tasks of our method and other methods. The classification results on ADNI dataset are summarized in Table 2.

Table 2: Comparison of the proposed method with the existing state-of-the-art methods reported in the literature.

	Methods	Data	AD *vs.* CN				pMCI *vs.* sMCI			
			ACC	SEN	SPE	AUC	ACC	SEN	SPE	AUC
ML-based	Moradi et al.[20]	sMRI + Clinical info	-	-	-	-	82.00%	87.00%	74.00%	**90.00**%
	Beheshti et al.[21]	sMRI	93.01%	89.13%	96.80%	93.51%	75.00%	76.92%	73.23%	75.08%
	Calvini et al.[23]	sMRI	-	74.00%	85.00%	86.30%	-	-	-	-
	Koikkalainen et al.[24]	sMRI	86.00%	81.00%	91.00%	-	72.10%	77.00%	71.00%	-
	Liu et al.[25]	sMRI	93.06%	94.85%	90.49%	95.79%	79.25%	87.92%	75.54%	83.44%
	Shi et al.[26]	sMRI + PET + CSF	95.00%	95.30%	94.70%	93.20%	-	-	-	-
	Tong et al.[27]	sMRI	90.00%	86.00%	93.00%	-	72.00%	69.00%	74.00%	-
	Coupe et al.[28]	sMRI	91.00%	87.00%	94.00%	-	74.00%	73.00%	74.00%	-
DL-based	Suk et al.[18]	sMRI + PET	95.35%	94.65%	95.22%	98.77%	75.92%	48.04%	**95.23**%	74.66%
	Shi et al.[43]	sMRI + PET	97.13%	95.93%	98.53%	97.20%	78.88%	68.04%	86.81%	80.10%
	Liu et al.[44]	sMRI + PET	91.40%	92.32%	90.42%	-	-	-	-	-
	Cui et al.[45]	sMRI	92.29%	90.63%	93.72%	96.95%	75.00%	73.33%	76.19%	79.70%
	Liu et al.[29]	sMRI	91.09%	88.05%	93.50%	95.86%	76.90%	42.11%	82.43%	77.64%
	Kang et al.[34]	sMRI	90.40%	-	-	-	66.70%	-	-	-
	Lian et al.[46]	sMRI	90.30%	82.40%	96.50%	95.10%	80.9%	52.60%	85.40%	78.10%
	Chen et al.[47]	sMRI	95.32%	91.18%	93.94%	-	77.60%	71.62%	75.85%	-
	Zhang et al.[48]	sMRI	93.20%	92.40%	94.00%	96.10%	82.90%	**90.00**%	75.70%	86.50%
	Basaia et al.[22]	sMRI + PET + CSF	93.20%	93.00%	93.30%	-	-	-	-	-
	Proposed	sMRI + Clinical info	**98.61**%	**98.54**%	**98.67**%	**99.08**%	**84.49**%	83.50%	81.48%	85.69%

In the task of AD *vs.* CN, the best ACC, SEN, SPE, and AUC values implemented by previous works are respectively 97.13%, 95.93%, 98.53%, and 98.77%, which are realized by the works of Shi et al.[43] and Suk et al.[18] The proposed method has the ACC of 98.61%, the SEN of 98.54%, the SPE of 98.67%, and the AUC of 99.08%, which are respectively 1.48%, 2.61%, 0.14%, and 0.31% higher than the best metrics achieved by other methods. In the task of pMCI *vs.* sMCI, the values of ACC, SEN, SPE, and AUC obtained by the proposed framework are respectively 84.49%, 83.50%, 81.48%, and 85.69%. Our method achieves the best prediction accuracy, which is 1.59% higher than the best ACC obtained by Zhang et al.[48] These results show that the proposed framework can indeed yield a more accurate diagnosis, and have satisfactory performance on other evaluation metrics.

III.B. Effectiveness of entropy-based MIL pooling

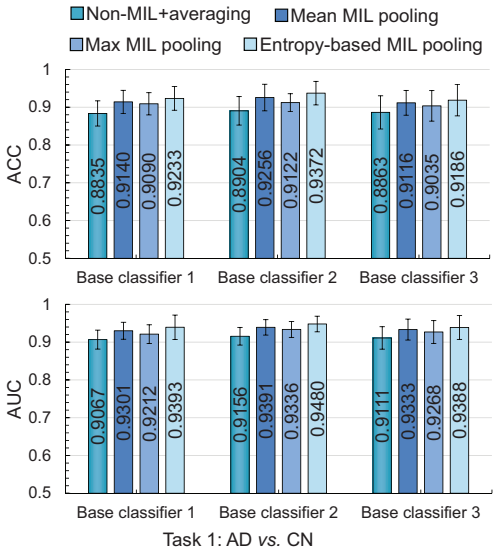
To evaluate the effectiveness of entropy-based MIL pooling, we compare the results of base classifiers without MIL pooling and with different MIL pooling layers. The compared methods include non-MIL+averaging method, Mean MIL pooling method, and Maximum pooling method. The non-MIL+averaging method has the same architectures as base CNN classifiers except no MIL pooling, and performs classification through averaging the slice-level results. Both mean MIL pooling method and maximum pooling method also have the same architectures as base CNN classifiers, only replacing the entropy-based MIL pooling layer. The classification results in terms of ACC and AUC for two tasks are shown in Fig. 5.

From Fig. 5, it can be learned that MIL methods (i.e., mean MIL pooling, max MIL pooling, and entropy-based MIL pooling) yield better results in terms of ACC and AUC. Taking the base classifier 1 as an example, the ACC and AUC achieved by MIL methods are on average higher 0.0319 and 0.0247 than non-MIL method in task 1, and higher 0.0199 and 0.0249 in task 2. Compared with mean MIL pooling and max MIL pooling methods, the proposed entropy-based MIL pooling achieves the best results on both tasks, which can reach 0.9372 ACC and 0.9480 AUC on task 1 (achieved by base classifier 2), 0.7959 ACC and 0.8081 AUC on task 2 (achieved by base classifier 1). The above results reflect that the MIL methods can improve the classification performance than the non-MIL method, and confirm that the entropy-based MIL pooling method is more effective than the normal MIL methods, which shows the effectiveness of entropy-based MIL pooling.

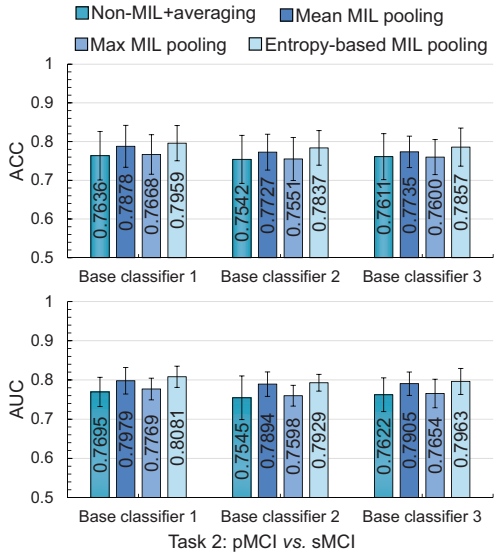
III.C. Effectiveness of MHSA voting

A key component of the proposed ensemble framework is the ensemble approaches to fuse the base classifiers. We conduct the experiments to verify the effectiveness of MHSA voting. Specifically, base classifier 1, base classifier 2, and base classifier 3 are fused via different voting approaches including majority voting (MV), weighted voting (WV), SVM-based voting (SVM), and the proposed MHSA voting. Table 3 reports the corresponding results of different ensemble approaches.

From Table 3, it can be observed that two learnable ensemble approaches (i.e., SVM-based voting, and MHSA voting) yield better classification performance on two tasks than unlearnable approaches (i.e., majority voting, and weighted voting). In the task of AD *vs.* CN, the results obtained by majority voting and weighted voting are lower than the maximum values of ACC and AUC (achieved by base classifier 2) before fusion. And the



(a) Classification results in terms of ACC and AUC for task 1



(b) Classification results in terms of ACC and AUC for task 2

Figure 5: Classification results in terms of ACC and AUC achieved by three base CNN classifiers with different MIL pooling layers for two tasks, i.e., AD *vs.* CN, and pMCI *vs.* sMCI. The error bars denote the standard deviations of the results.

results obtained by SVM-based voting are basically consistent with the maximum values before fusion. Only the MHSA voting achieves an improvement in results, with the ACC of 0.9419, and the AUC of 0.9545, which is at least 0.0047 higher than the metrics generated by base classifiers. In the task of pMCI *vs.* sMCI, all ensemble approaches can obtain better results than that before fusion. The results obtained via majority voting have the minimum improvement, with the ACC of 0.8061, and the AUC of 0.8165. The maximum improvement

Table 3: Classification results of different ensemble approaches on two tasks.

Ensemble Members	Ensemble Approach	AD *vs.* CN		pMCI *vs.* sMCI	
		ACC	AUC	ACC	AUC
Base classifier 1	-	0.9233 ± 0.0314	0.9393 ± 0.0323	0.7959 ± 0.0454	0.8081 ± 0.0271
Base classifier 2	-	0.9372 ± 0.0311	0.9480 ± 0.0207	0.7837 ± 0.0447	0.7929 ± 0.0251
Base classifier 3	-	0.9186 ± 0.0416	0.9388 ± 0.0315	0.7857 ± 0.0492	0.7963 ± 0.0332
Base classifier 1, 2, 3	MV	0.9279 ± 0.0283	0.9306 ± 0.0301	0.8061 ± 0.0366	0.8165 ± 0.0318
	WV	0.9302 ± 0.0245	0.9415 ± 0.0202	0.8265 ± 0.0409	0.8316 ± 0.0431
	SVM	0.9349 ± 0.0209	0.9478 ± 0.0199	0.8286 ± 0.0422	0.8367 ± 0.0395
	MHSA	**0.9419 ± 0.0232**	**0.9545 ± 0.0205**	**0.8408 ± 0.0350**	**0.8535 ± 0.0283**

Data are mean ± standard deviation.
MV: majority voting; WV: weighted voting; SVM: SVM-based voting; MHSA: MHSA voting.

on results is achieved by MHSA voting, which is at least 0.0449 higher than the metrics generated by base classifiers. Compared with these common ensemble approaches, MHSA voting can further improve the effects of fusion. These results confirm the effectiveness of using MHSA voting.

III.D. Influence of transfer learning

To demonstrate the impact of transfer learning, we compare the experimental results with and without transfer learning. In this group of experiments, we train base classifiers from scratch for task 2 without adopting transfer learning strategy, and compare their classification performance with that obtained by base classifiers trained with transfer learning strategy. Fig. 6 shows the classification results in terms of ACC and AUC for task 2.

As shown in Fig. 6, it can be seen that transfer learning strategy significantly improves the classification performance. Take base classifier 1, 2, 3, 4 fused via MHSA voting as an example, with the aid of transfer learning, it improves the ACC from 0.8106 to 0.8449, the AUC from 0.8196 to 0.8569, which has at least a 4.23% boost. Meanwhile, other methods trained with transfer learning have higher gain percentages, the ACC has an average gain of 5.65%, and the AUC has an average gain of 6.13%. These results indicate that the use of transfer learning strategy can indeed improve the classification performance on task 2.

III.E. Influence of clinical information

As introduced in Section II.C.5., base classifier 4 (i.e., MLP) is chosen for clinical information analysis. Base classifier 4 is fused with other base classifiers via MHSA voting to construct a multi-model ensemble framework. To investigate the influence of clinical information, we compare the classification performance achieved by *Only Clin info* (base classifier 4), *Without*

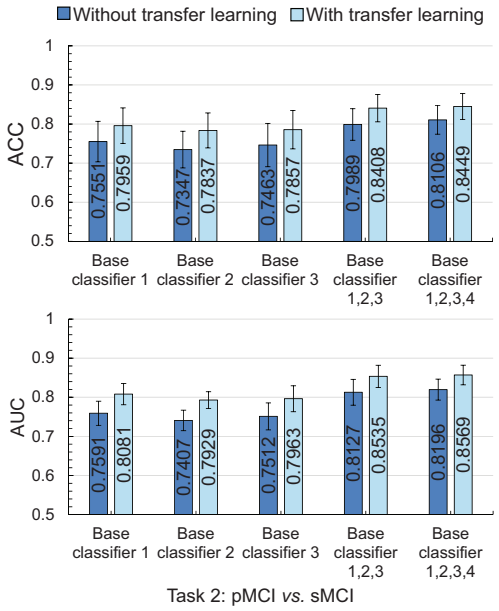


Figure 6: Classification results in terms of ACC and AUC achieved by base classifiers trained without and with transfer learning for task 2. Base classifier 1, 2, 3 and base classifier 1, 2, 3, 4 are fused via MHSA voting. The error bars denote the standard deviations of the results.

Clin info (base classifier 1, 2, 3 fused via MHSA voting), and *With Clin info* (base classifier 1, 2, 3, 4 fused via MHSA voting). The corresponding results are as demonstrated in Table 4.

Table 4: Classification results of different methods with and without clinical information.

Task	Method	ACC	SEN	SPE	AUC
AD *vs.* CN	*Only clin info.*	0.9837 ± 0.0204	0.9756 ± 0.0209	0.9778 ± 0.0298	0.9848 ± 0.0167
	Without clin info.	0.9419 ± 0.0232	0.9268 ± 0.0311	0.9556 ± 0.0199	0.9545 ± 0.0205
	With clin info.	**0.9861 ± 0.0182**	**0.9854 ± 0.0233**	**0.9867 ± 0.0221**	**0.9908 ± 0.0143**
pMCI *vs.* sMCI	*Only clin info.*	0.6980 ± 0.0870	0.7255 ± 0.0744	0.7037 ± 0.0661	0.6987 ± 0.0598
	Without clin info.	0.8408 ± 0.0350	0.8273 ± 0.0422	**0.8234 ± 0.0406**	0.8535 ± 0.0283
	With clin info.	**0.8449 ± 0.0332**	**0.8350 ± 0.0405**	0.8148 ± 0.0341	**0.8569 ± 0.0214**

Data are mean ± standard deviation.

As shown in Table 4, for the task of AD *vs.* CN, the use of clinical information can significantly improve the diagnosis performance. Compared with the results achieved by *Without Clin info*, *With Clin info* improves the ACC from 0.9419 to 0.9861, the SEN from 0.9268 to 0.9854, the SPE from 0.9556 to 0.9867, and the AUC from 0.9545 to 0.9908. And the quantification biases of ACC, SEN, and AUC obtained by *With Clin info* are smaller than that of *Without Clin info*. *Only Clin info* can obtain similar performance to *With Clin*

info in terms of ACC. However, the SEN, SPE, AUC achieved by *Only Clin info* are lower than that achieved by *With Clin info*, and the quantification biases of these metrics are also higher. For the task of pMCI *vs*. sMCI, the use of clinical information also improves diagnosis performance, but not as significantly as the task of AD *vs*. CN. *With Clin info* yields better results, with the ACC of 0.8449, the AUC of 0.8569, which are higher than that obtained by the other two methods. Though *Without Clin info* yields similar performance to *With Clin info*, the quantification biases of all metrics are higher than that obtained by *With Clin info*. The above results reveal that the use of clinical information can provide better classification performance, and reduce the quantification bias of diagnosis.

III.F. Indispensability of four base classifiers

To prove the indispensability of the four types of base classifiers, we summarize and compare the classification performance of fused different types of base classifiers. Specifically, base classifier 1, 2, 3, and 4 are randomly fused by MHSA voting. The corresponding results for task 1 and task 2 are reported in Table 5, and some of the ROC curves for the two tasks are respectively represented in Fig. 7.

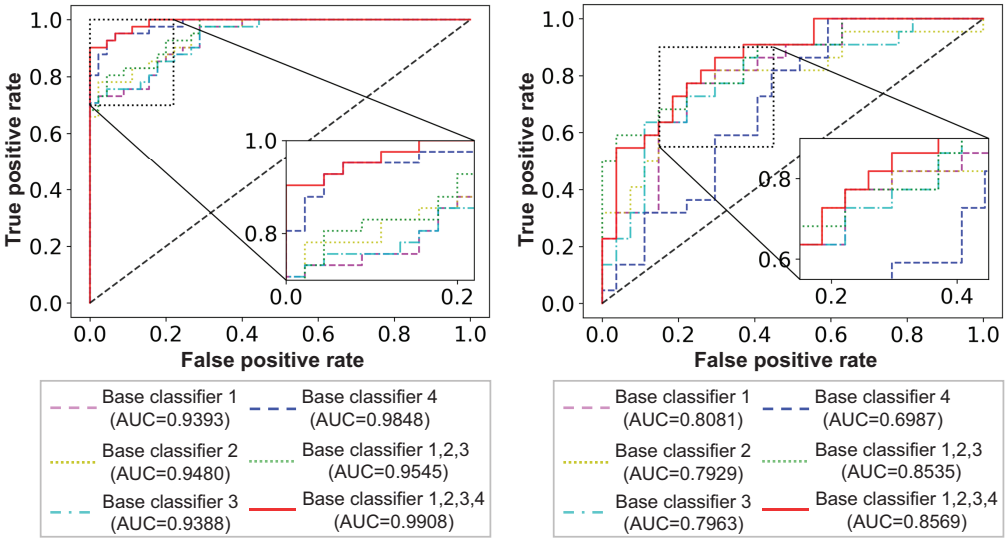
Table 5: Classification results of fused different types of base classifiers.

No. of Cls	Members	AD *vs*. CN		pMCI *vs*. sMCI	
		ACC	AUC	ACC	AUC
1	Base classifier 1	0.9233 ± 0.0314	0.9393 ± 0.0323	0.7959 ± 0.0454	0.8081 ± 0.0271
	Base classifier 2	0.9372 ± 0.0311	0.9480 ± 0.0207	0.7837 ± 0.0447	0.7929 ± 0.0251
	Base classifier 3	0.9186 ± 0.0416	0.9388 ± 0.0315	0.7857 ± 0.0492	0.7963 ± 0.0332
	Base classifier 4	0.9837 ± 0.0204	0.9848 ± 0.0167	0.6980 ± 0.0870	0.6987 ± 0.0598
2	Base classifier 1, 2	0.9396 ± 0.0276	0.9539 ± 0.0191	0.7999 ± 0.0371	0.8102 ± 0.0298
	Base classifier 1, 3	0.9253 ± 0.0291	0.9405 ± 0.0198	0.8018 ± 0.0466	0.8143 ± 0.0248
	Base classifier 2, 3	0.9380 ± 0.0323	0.9485 ± 0.0212	0.7993 ± 0.0322	0.8036 ± 0.0231
	Base classifier 1, 4	0.9847 ± 0.0197	0.9863 ± 0.0155	0.7967 ± 0.0507	0.8098 ± 0.0336
	Base classifier 2, 4	0.9856 ± 0.0184	0.9902 ± 0.0152	0.7901 ± 0.0581	0.8003 ± 0.0364
	Base classifier 3, 4	0.9847 ± 0.0187	0.9866 ± 0.0152	0.7896 ± 0.0482	0.7998 ± 0.0342
3	Base classifier 1,2,3	0.9419 ± 0.0232	0.9545 ± 0.0205	0.8408 ± 0.0350	0.8535 ± 0.0283
	Base classifier 1,2,4	0.9855 ± 0.0265	0.9902 ± 0.0144	0.8059 ± 0.0382	0.8154 ± 0.0350
	Base classifier 1,3,4	0.9841 ± 0.0227	0.9862 ± 0.0156	0.8122 ± 0.0394	0.8205 ± 0.0312
	Base classifier 2,3,4	0.9852 ± 0.0197	0.9900 ± 0.0137	0.8041 ± 0.0435	0.8181 ± 0.0344
4	Base classifier 1, 2, 3, 4	**0.9861 ± 0.0182**	**0.9908 ± 0.0143**	**0.8449 ± 0.0332**	**0.8569 ± 0.0214**

Data are mean ± standard deviation.

From Table 5, when four base classifiers are fused, the best classification results can be obtained, the values of ACC for task 1 and task 2 are respectively 0.9861 and 0.8449, the values of AUC are respectively 0.9908 and 0.8569. And the quantification bias is also satisfactory. Base classifier 1, base classifier 2, and base classifier 3 have similar performance

on both tasks. Base classifier 4 (i.e., MLP) achieves great performance on task 1, while it performs not good on task 2. When two base CNN classifiers are randomly fused, the classification results are similar to that achieved by a single CNN classifier, and the quantification biases are lower. Due to the influence of clinical information, any base CNN classifier (i.e., base classifier 1, 2, and 3) fused with base classifier 4 could further boost the diagnosis performance, especially in the task of AD *vs*. CN. Though one base CNN classifier fused with base classifier 4 can improve the ACC and AUC, the quantification biases of them are higher than that before fused with base classifier 4. When three base classifiers are randomly fused, the fusions that include base classifier 4 can yield satisfactory results in the task of AD *vs*. CN, which are better than the fusions only including base CNN classifiers. In the task of pMCI *vs*. sMCI, we can see that the fusions only including base CNN classifiers have the better performance than that fusions including base classifier 4.



(a) ROC curves of AD *vs*. CN (b) ROC curves of pMCI *vs*. sMCI

Figure 7: Comparison of the ROC curves. The ROC curves are obtained by base classifier 1, base classifier 2, base classifier 3, base classifier 4, the fusion of base classifier 1, 2, 3, and the fusion of base classifier 1, 2, 3, 4. The upper left area of ROC curve is zoomed for clarity.

From Fig. 7, it can be learned that the fusion of four base classifiers has better ROC curves than others. The results in Table 5 and Fig. 7 illustrate that the fusion of these base classifiers can achieve better diagnosis performance than a single classifier, and each base classifier could play an important part in the ensemble framework.

IV. DISCUSSIONS

This work presents a reliable ensemble framework to diagnose AD and MCI using neural networks. MHSA voting improves the fusion of base classifiers in the ensemble, and entropy-based MIL strategy could use more effective information contained in sMRI. Overall, the proposed method provides the reliable diagnosis of AD and prediction of MCI conversion. We built our method based on ensemble learning for several reasons. First, though DL-based methods have been shown to surpass human experts in predictive accuracy, they tend to exhibit higher variance, especially when only a single DL model is adopted. However, reliable diagnosis is needed in the clinic, high variance makes it hard for a single model to generate convincing judgments. In contrast to a single DL model, ensemble learning that combines the outputs of multiple DL models has been proven to achieve better outcomes and generalizability[49], which is more applicable in clinical settings. Second, because the characteristics of AD are concealed, slow, and non-lethal, the collection of samples is difficult, often resulting in the limited number of samples. The limited number of samples may lead to over-fitting or inadequate training of a model, and limit the identification of complex AD patterns. Ensemble learning has the power in dealing with these challenges[32].

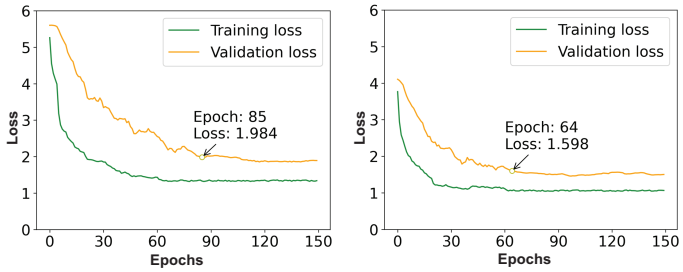
We compared the performance of the proposed method against several ML-based and DL-based methods. In all compared methods, MRI images were preprocessed through a similar pipeline to this work, including motion correction, intensity correction, skull stripping, and normalization. Following this basic pipeline, different methods then performed some specific operations (e.g., tissue segmentation and slices sampled in this work) to generate slices, regions, or patches of the brain according to the needs of these methods. In addition, cross-validation and corresponding data split were also adopted in most of the compared methods[18,20,21,25–28,34,43–45,47], and they took the average of the cross-validation results as the final performance. These means such as preprocessing procedures or cross-validation are a part of the compared methods and have no impact on demonstrating the effectiveness of the proposed method. As the results shown in Section III.A., our method significantly outperformed the compared methods in classification accuracy for both tasks (AD *vs.* CN, pMCI *vs.* sMCI). Noting that some compared methods[18,20,24,25,27,29,43,46,48] had quite unbalanced SEN and SPE, the imbalance of SEN and SPE indicates that the missed diagnosis or misdiagnosis rate of these methods was high. A previous work[29] achieved SEN of 42.11%, and the SPE of 82.43% in the task of pMCI *vs.* sMCI, which means only 42.11% pMCI patients were correctly diagnosed and 17.57% sMCI patients were misdiagnosed. The proposed method achieved balanced and satisfactory SEN and SPE for both tasks, which demonstrates that

our method can conduct a reliable diagnosis. Furthermore, the five-fold cross-validation approach has been performed in this work. The mean values and standard deviation of ACC and AUC are as demonstrated in Table 5. The proposed method achieved the best results on both tasks, which had the ACC of 0.9861 ± 0.0182 and the AUC of 0.9908 ± 0.0143 on AD *vs*. CN task, the ACC of 0.8449 ± 0.0332 and the AUC of 0.8569 ± 0.0214 on pMCI *vs*. sMCI task. The quantification biases of these metrics were effectively reduced by the use of ensemble learning, which was lower than that of each base classifier. The results with low quantification bias generated by our method indicate that the proposed method is able to generate a robust diagnosis, which is also in good agreement with the effect of ensemble learning.

In this work, MHSA voting is proposed to aggregate the outputs of base classifiers as previous studies[34,36] typically adopted the common voting approaches which ignore the interactions among base classifiers. The majority voting, weighted voting, and SVM-based voting are commonly used for the aggregation in the ensemble. However, these common voting approaches sometimes may cause a decrease or stay flat on results after fusion. The reason for this is that majority voting and weighted voting are not data-adaptive, they assign the fixed weight to each base classifier. Though SVM-based voting is a learnable ensemble approach, it leaves the interactions among the ensemble members out of consideration. MHSA voting has been shown to have an improvement on results after fusion. This implies that the interactions among the base classifiers can play a role during their fusion, and MHSA voting can exploit the interactions to generate better classification results during the fusion of base classifiers.

While MIL strategy has been applied in the diagnosis of different diseases, to our knowledge, rare studies have explored it in the diagnosis of AD based on slice-level. We incorporated entropy-based MIL strategy into base CNN classifiers to use more effective information contained in sMRI. As shown in Fig. 5, MIL strategy can indeed further improve the performance of both tasks in contrast to non-MIL methods, in which the proposed entropy-based MIL strategy has been shown to achieve the best classification results. Due to AD-related pathological areas having the uneven distribution in sMRI, non-MIL methods are easily affected, thereby resulting in sub-optimal performance in two tasks. Compared with non-MIL methods, MIL methods consider the relationships between slices, which is beneficial to improving the utilization of information contained in sMRI. The normal MIL methods (i.e., mean MIL method, and maximum MIL method) consider that the relationships between slices have no difference, and the slices have similar feature expression abilities. Nevertheless, the slices with abundant tissue information are generally getting more attention in

clinical diagnosis, and radiologists also focus on these slices. Similar to the habit of radiologists' review of MRI, the proposed entropy-based MIL method measures the feature expression abilities of different slices according to their information entropy, which can generate more reasonable embedding-level features for further classification. And therefore, the entropy-based MIL method has better performance than normal MIL methods.



(a) Loss function curves without transfer learning

(b) Loss function curves with transfer learning

Figure 8: Comparison of the loss function curves achieved by training without and with transfer learning for task 2.

Transfer learning improved the classification results in terms of ACC and AUC by ∼4% across two tasks. This situation is consistent with existing studies[28,29]. The results demonstrate that the two tasks are correlated, and the supplementary information from AD and CN subjects implicitly enriches the features in the task of pMCI *vs*. sMCI during training. In addition, we also analyzed the influence of transfer learning on training duration. Here, we trained the proposed method for task 2 without early stopping and set the epochs to 150. Fig. 8 shows the loss function curves with and without transfer learning during training. As observed in Fig. 8, the training loss has a faster downward trend than the validation loss, and after the convergence of training, the validation loss is slightly more than the training loss. With transfer learning, the initial values of training and validation losses (epoch 1) were lower than that without transfer learning, and the validation loss converged to about 1.6 after epoch 64. The validation loss converged to about 2.0 after epoch 85 when transfer learning strategy was not adopted. These results show that the model can fit the data better and faster when using transfer learning. In this work, early stopping was adopted with the patience of 10 epochs on the validation loss, and the training time was five minutes per epoch. For the task of pMCI *vs*. sMCI, the training lasted about 6 hours, which can save about 1.7 hours in contrast to that training without transfer learning. For the task of AD *vs*. CN, the training time was about 7.5 hours.

As different imaging modalities and clinical data can provide various information about AD patients, we adopted multimodal data (sMRI and clinical information) to develop an ensemble framework. In this work, the use of multimodal data led to an overall improvement

in both tasks, which improved the diagnosis performance and reduced the quantification bias. From Table 2, it can be observed that most studies using multimodal data have better performance than the studies using single modal data. Moreover, we analyzed the sensitivity of clinical information to two tasks. For the task of AD *vs*. CN, the use of clinical information only can also obtain satisfactory performance, while for the task of pMCI *vs*. sMCI, the use of clinical information only cannot achieve good results. These results show that the clinical information is more sensitive to the task of AD *vs*. CN than that to the task of pMCI *vs*. sMCI. It can be also learned that cognitive and neuropsychological measures in clinical information change greatly from normal cognition to dementia, and these measures have no significant change in CN or MCI stages. This inference is consistent with previous research[50,51].

AD is an irreversible neurodegenerative disease with concealed, slow, and non-lethal characteristics, which is also a serious social problem. The dementia symptoms caused by AD gradually worsen over several years. In general, a person with AD lives 4 to 8 years after diagnosis but can live as long as 20 years, depending on other factors (e.g., earlier diagnosis or intervention). At present, AD has no cure, some treatments can only temporarily slow the worsening of dementia symptoms and improve the quality of life for AD patients and their caregivers. Earlier diagnosis of AD is crucial for prolonging the lifespan and improving the quality of life for those with AD. Our proposed ensemble framework is able to generate reliable and robust results for the diagnosis of AD and the prediction of MCI conversion, which has great practical significance for the earlier diagnosis of AD. The detailed analyses of the results give an important indication that the proposed ensemble framework can potentially be employed in the reliable diagnosis of AD and prediction of MCI conversion. Furthermore, due to the characteristics of AD, the collection of AD samples is difficult in clinical settings. With ensemble learning, the dilemma caused by the limited number of samples can be solved to some extent[32]. The proposed method is based on ensemble learning, which makes our method potential to perform reliable diagnoses under limited data, thereby reducing the burden of physicians collecting data. In many clinical settings, because it is difficult to identify the exact cause of dementia, multiple diagnostic tests are typically adopted to determine if a person has AD, including brain imaging, mental cognitive status tests, etc. To closer to practical clinical application and obtain a more reliable diagnosis, we also adopted the multimodal data in this work. It is worth noting that our method can also achieve satisfactory results only using sMRI.

This current work has some limitations despite its successful performance in AD diagnosis and MCI conversion prediction. The black-box nature is a common limitation of deep

learning methods, which is also the main reason that limits the widespread application of medical artificial intelligence (AI). In clinical settings, to determine whether a person suffers from a certain disease, it needs to undergo a detailed clinical examination, and the physicians confirm the condition of this person according to the clinical test results. In this process, the basis for the diagnosis is detailed and clear. For medical AI, the details of algorithmic decision-making should also be exposed like clinical diagnosis, which is currently difficult. Note that conceptual understanding and experiences owned by physicians are impossible for AI to fully learn. To deploy an explainable AI in medical practices, it still requires the necessary human oversight[52]. The interactive deep learning with the "human in the loop" can be potentially considered as a robust way to handle explainability. This human-in-the-loop deep learning combines the conceptual understanding and experiences owned by physicians with the effectiveness of deep learning, which can ensure that decision-making is controllable and clinically justified. As a high level of accountability is required in the medical field, machined decisions and predictions need to be explained clearly, our future work will include exploring human-in-the-loop deep learning.

V. CONCLUSION

In this paper, a robust ensemble framework is proposed for reliable diagnosis of AD and prediction of MCI conversion. Specifically, three base CNN classifiers with different scales of network width, depth, and resolution are designed to capture detailed features in sMRI. To better use effective information contained in sMRI, we incorporate entropy-based MIL strategy into base CNN classifiers, which can take the information densities of slices into account to generate more reasonable features for classification. Additionally, one shallow classifier (i.e., MLP) is employed to analyze the clinical information. The final diagnosis is achieved by MHSA voting approach that aggregates the predictions of base classifiers while considering the interactions among them. Extensive experimental results on ADNI database show that the proposed ensemble framework has reliable and competitive performance in both tasks.

VI. ACKNOWLEDGMENTS

This work was supported by the Young Scientist Studio of Harbin Institute of Technology. Data used in preparation of this article were obtained from the Alzheimer's Disease Neu-

roimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

VII. CONFLICT OF INTEREST DISCLOSURE

The authors have no conflicts of interest to declare.

References

[1] D. E. Barnes and K. Yaffe, The projected effect of risk factor reduction on Alzheimer's disease prevalence, The Lancet Neurology **10**, 819–828 (2011).

[2] R. Wang et al., Generation of synthetic PET images of synaptic density and amyloid from 18F-FDG images using deep learning, Medical physics **48**, 5115–5129 (2021).

[3] M. Grundman et al., Mild cognitive impairment can be distinguished from Alzheimer disease and normal aging for clinical trials, Archives of neurology **61**, 59–66 (2004).

[4] R. C. Petersen, R. O. Roberts, D. S. Knopman, B. F. Boeve, Y. E. Geda, R. J. Ivnik, G. E. Smith, and C. R. Jack, Mild cognitive impairment: ten years later, Archives of neurology **66**, 1447–1455 (2009).

[5] N. I. Bradfield and D. Ames, Mild cognitive impairment: narrative review of taxonomies and systematic review of their prediction of incident Alzheimer's disease dementia, BJPsych bulletin **44**, 67–74 (2020).

[6] A. Nordberg, PET imaging of amyloid in Alzheimer's disease, The lancet neurology **3**, 519–527 (2004).

[7] S. Lehéricy, M. Marjanska, L. Mesrob, M. Sarazin, and S. Kinkingnehun, Magnetic resonance imaging of Alzheimer's disease, European radiology **17**, 347–362 (2007).

[8] K. Blennow and H. Hampel, CSF markers for incipient Alzheimer's disease, The Lancet Neurology **2**, 605–613 (2003).

[9] C. R. Jack Jr, D. S. Knopman, W. J. Jagust, L. M. Shaw, P. S. Aisen, M. W. Weiner, R. C. Petersen, and J. Q. Trojanowski, Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade, The Lancet Neurology **9**, 119–128 (2010).

[10] G. B. Frisoni, N. C. Fox, C. R. Jack, P. Scheltens, and P. M. Thompson, The clinical use of structural MRI in Alzheimer disease, Nature Reviews Neurology **6**, 67–77 (2010).

[11] I. Arevalo-Rodriguez, N. Smailagic, M. R. i Figuls, A. Ciapponi, E. Sanchez-Perez, A. Giannakou, O. L. Pedraza, X. B. Cosp, and S. Cullum, Mini-Mental State Examination (MMSE) for the detection of Alzheimer's disease and other dementias in people with mild cognitive impairment (MCI), Cochrane Database of Systematic Reviews (2015).

[12] S. E. O'Bryant et al., Staging dementia using Clinical Dementia Rating Scale Sum of Boxes scores: a Texas Alzheimer's research consortium study, Archives of neurology **65**, 1091–1095 (2008).

[13] P. Doraiswamy, F. Bieber, L. Kaiser, K. Krishnan, J. Reuning-Scherer, and B. Gulanski, The Alzheimer's Disease Assessment Scale: patterns and predictors of baseline cognitive performance in multicenter Alzheimer's disease trials, Neurology **48**, 1511–1517 (1997).

[14] A. Estévez-González, J. Kulisevsky, A. Boltes, P. Otermín, and C. García-Sánchez, Rey verbal learning test is a useful tool for differential diagnosis in the preclinical phase of Alzheimer's disease: comparison with mild cognitive impairment and normal aging, International journal of geriatric psychiatry **18**, 1021–1028 (2003).

[15] M. Lin, S. Momin, Y. Lei, H. Wang, W. J. Curran, T. Liu, and X. Yang, Fully automated segmentation of brain tumor from multiparametric MRI using 3D context deep supervised U-Net, Medical Physics **48**, 4365–4374 (2021).

[16] R. Haweel, A. Shalaby, A. Mahmoud, N. Seada, S. Ghoniemy, M. Ghazal, M. F. Casanova, G. N. Barnes, and A. El-Baz, A robust DWT–CNN-based CAD system for early diagnosis of autism using task-based fMRI, Medical physics **48**, 2315–2326 (2021).

[17] M. Jo and S.-H. Oh, A preliminary attempt to visualize nigrosome 1 in the substantia nigra for Parkinson's disease at 3T: An efficient susceptibility map-weighted imaging (SMWI) with quantitative susceptibility mapping using deep neural network (QSMnet), Medical Physics **47**, 1151–1160 (2020).

[18] H.-I. Suk et al., Hierarchical feature representation and multimodal fusion with deep learning for AD/MCI diagnosis, NeuroImage **101**, 569–582 (2014).

[19] E.-J. Hwang, H.-G. Kim, D. Kim, H. Y. Rhee, C.-W. Ryu, T. Liu, Y. Wang, and G.-H. Jahng, Texture analyses of quantitative susceptibility maps to differentiate Alzheimer's disease from cognitive normal and mild cognitive impairment, Medical physics **43**, 4718–4728 (2016).

[20] E. Moradi et al., Machine learning framework for early MRI-based Alzheimer's conversion prediction in MCI subjects, Neuroimage **104**, 398–412 (2015).

[21] I. Beheshti et al., Classification of Alzheimer's disease and prediction of mild cognitive impairment-to-Alzheimer's conversion from structural magnetic resource imaging using feature ranking and a genetic algorithm, Computers in biology and medicine **83**, 109–119 (2017).

[22] S. Basaia et al., Automated classification of Alzheimer's disease and mild cognitive impairment using a single MRI and deep neural networks, NeuroImage: Clinical **21**, 101645 (2019).

[23] P. Calvini et al., Automatic analysis of medial temporal lobe atrophy from structural MRIs for the early assessment of Alzheimer disease, Medical physics **36**, 3737–3747 (2009).

[24] J. Koikkalainen et al., Multi-template tensor-based morphometry: application to analysis of Alzheimer's disease, NeuroImage **56**, 1134–1144 (2011).

[25] M. Liu, D. Zhang, and D. Shen, Relationship induced multi-template learning for diagnosis of Alzheimer's disease and mild cognitive impairment, IEEE transactions on medical imaging **35**, 1463–1474 (2016).

[26] Y. Shi, H.-I. Suk, Y. Gao, S.-W. Lee, and D. Shen, Leveraging coupled interaction for multimodal Alzheimer's disease diagnosis, IEEE transactions on neural networks and learning systems **31**, 186–200 (2019).

[27] T. Tong et al., Multiple instance learning for classification of dementia in brain MRI, Medical image analysis **18**, 808–818 (2014).

[28] P. Coupé et al., Scoring by nonlocal image patch estimator for early detection of Alzheimer's disease, NeuroImage: clinical **1**, 141–152 (2012).

[29] M. Liu, J. Zhang, E. Adeli, and D. Shen, Landmark-based deep multi-instance learning for brain disease diagnosis, Medical image analysis **43**, 157–168 (2018).

[30] S.-H. Wang, P. Phillips, Y. Sui, B. Liu, M. Yang, and H. Cheng, Classification of Alzheimer's disease based on eight-layer convolutional neural network with leaky rectified linear unit and max pooling, Journal of medical systems **42**, 1–11 (2018).

[31] C. Wu et al., Discrimination and conversion prediction of mild cognitive impairment using convolutional neural networks, Quantitative imaging in medicine and surgery **8**, 992 (2018).

[32] Y. Cao, T. A. Geddes, J. Y. H. Yang, and P. Yang, Ensemble deep learning in bioinformatics, Nature Machine Intelligence **2**, 500–508 (2020).

[33] A. Loddo, S. Buttau, and C. Di Ruberto, Deep learning based pipelines for Alzheimer's disease diagnosis: a comparative study and a novel deep-ensemble method, Computers in biology and medicine **141**, 105032 (2022).

[34] W. Kang et al., Multi-model and multi-slice ensemble learning architecture based on 2D convolutional neural networks for Alzheimer's disease diagnosis, Computers in Biology and Medicine **136**, 104678 (2021).

[35] J. Y. Choi and B. Lee, Combining of Multiple Deep Networks via Ensemble Generalization Loss, Based on MRI Images, for Alzheimer's Disease Classification, IEEE Signal processing letters **27**, 206–210 (2020).

[36] H. Wang, Y. Shen, S. Wang, T. Xiao, L. Deng, X. Wang, and X. Zhao, Ensemble of 3D densely connected convolutional network for diagnosis of mild cognitive impairment and Alzheimer's disease, Neurocomputing **333**, 145–156 (2019).

[37] T. G. Dietterich, Ensemble methods in machine learning, in *International workshop on multiple classifier systems*, pages 1–15, Springer, 2000.

[38] C. R. Jack Jr et al., The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods, Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine **27**, 685–691 (2008).

[39] J. G. Sled, A. P. Zijdenbos, and A. C. Evans, A nonparametric method for automatic correction of intensity nonuniformity in MRI data, IEEE transactions on medical imaging **17**, 87–97 (1998).

[40] V. Fonov, A. Evans, R. McKinstry, C. Almli, and D. Collins, Unbiased nonlinear average age-appropriate brain templates from birth to adulthood, NeuroImage **47**, S102 (2009), Organization for Human Brain Mapping 2009 Annual Meeting.

[41] D. Holland et al., Subregional neuroanatomical change as a biomarker for Alzheimer's disease, Proceedings of the National Academy of Sciences **106**, 20954–20959 (2009).

[42] K. He, X. Zhang, S. Ren, and J. Sun, Deep residual learning for image recognition, in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[43] J. Shi, X. Zheng, Y. Li, Q. Zhang, and S. Ying, Multimodal neuroimaging feature learning with multimodal stacked deep polynomial networks for diagnosis of Alzheimer's disease, IEEE journal of biomedical and health informatics **22**, 173–183 (2017).

[44] S. Liu et al., Multimodal neuroimaging feature learning for multiclass diagnosis of Alzheimer's disease, IEEE Transactions on Biomedical Engineering **62**, 1132–1140 (2014).

[45] R. Cui and M. Liu, Hippocampus Analysis by Combination of 3-D DenseNet and Shapes for Alzheimer's Disease Diagnosis, IEEE journal of biomedical and health informatics **23**, 2099–2107 (2018).

[46] C. Lian, M. Liu, J. Zhang, and D. Shen, Hierarchical Fully Convolutional Network for Joint Atrophy Localization and Alzheimer's Disease Diagnosis Using Structural MRI, IEEE Transactions on Pattern Analysis and Machine Intelligence **42**, 880–893 (2020).

[47] Y. Chen and Y. Xia, Iterative sparse and deep learning for accurate diagnosis of Alzheimer's disease, Pattern Recognition **116**, 107944 (2021).

[48] J. Zhang, B. Zheng, A. Gao, X. Feng, D. Liang, and X. Long, A 3D densely connected convolution neural network with connection-wise attention mechanism for Alzheimer's disease classification, Magnetic Resonance Imaging **78**, 119–126 (2021).

[49] C. Ju, A. Bibaut, and M. van der Laan, The relative performance of ensemble methods with deep convolutional neural networks for image classification, Journal of Applied Statistics **45**, 2800–2818 (2018).

[50] P. K. Crane et al., Development and assessment of a composite score for memory in the Alzheimer's Disease Neuroimaging Initiative (ADNI), Brain imaging and behavior **6**, 502–516 (2012).

[51] R. C. Petersen et al., Alzheimer's disease neuroimaging initiative (ADNI): clinical characterization, Neurology **74**, 201–209 (2010).

[52] E. Tjoa and C. Guan, A survey on explainable artificial intelligence (xai): Toward medical xai, IEEE transactions on neural networks and learning systems **32**, 4793–4813 (2020).