Neural Networks 159 (2023) 14–24

Contents lists available at ScienceDirect

Neural Networks

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

Monte Carlo Ensemble Neural Network for the diagnosis of Alzheimer's disease

Chaoqiang Liu [a], Fei Huang [c], Anqi Qiu [a,b,c,d,e,f,*], for the Alzheimer's Disease Neuroimaging Initiative [1]

[a] Department of Biomedical Engineering, National University of Singapore, Singapore
[b] NUS (Suzhou) Research Institute, National University of Singapore, China
[c] School of Computer Engineering and Science, Shanghai University, China
[d] Institute of Data Science, National University of Singapore, Singapore
[e] The N.1 Institute for Health, National University of Singapore, Singapore
[f] The Johns Hopkins University, MD, USA

ARTICLE INFO

Article history:
Received 7 June 2022
Received in revised form 13 October 2022
Accepted 31 October 2022
Available online 24 November 2022

Keywords:
Monte Carlo sampling
Convolutional neural network
ResNet
DenseNet
Structural magnetic resonance imaging
Alzheimer's disease

ABSTRACT

Convolutional neural networks (CNNs) have been increasingly used in the computer-aided diagnosis of Alzheimer's Disease (AD). This study takes the advantage of the 2D-slice CNN fast computation and ensemble approaches to develop a Monte Carlo Ensemble Neural Network (MCENN) by introducing Monte Carlo sampling and an ensemble neural network in the integration with ResNet50. Our goals are to improve the 2D-slice CNN performance and to design the MCENN model insensitive to image resolution. Unlike traditional ensemble approaches with multiple base learners, our MCENN model incorporates one neural network learner and generates a large number of possible classification decisions via Monte Carlo sampling of feature importance within the combined slices. This can overcome the main weakness of the lack of 3D brain anatomical information in 2D-slice CNNs and develop a neural network to learn the 3D relevance of the features across multiple slices. Brain images from Alzheimer's Disease Neuroimaging Initiative (ADNI, 7199 scans), the Open Access Series of Imaging Studies-3 (OASIS-3, 1992 scans), and a clinical sample (239 scans) are used to evaluate the performance of the MCENN model for the classification of cognitively normal (CN), patients with mild cognitive impairment (MCI) and AD. Our MCENN with a small number of slices and minimal image processing (rigid transformation, intensity normalization, skull stripping) achieves the AD classification accuracy of 90%, better than existing 2D-slice CNNs (accuracy: 63% $\sim$ 84%) and 3D CNNs (accuracy: 74% $\sim$ 88%). Furthermore, the MCENN is robust to be trained in the ADNI dataset and applied to the OASIS-3 dataset and the clinical sample. Our experiments show that the AD classification accuracy of the MCENN model is comparable when using high- and low-resolution brain images, suggesting the insensitivity of the MCENN to image resolution. Hence, the MCENN does not require high-resolution 3D brain structural images and comprehensive image processing, which supports its potential use in a clinical setting.

1. Introduction

Alzheimer's disease (AD) is clinically characterized by the appearance of a progressive decline in memory and cognition (Alzheimer's Association, 2015). It is the most common form of dementia and has an astounding impact at individual and societal levels (Prince et al., 2015; Rizzi, Rosset, & Roriz-Cruz, 2014; Wimo et al., 2017). Early stages of AD are windows of opportunity in reducing the incidence and symptoms of AD and hence the early diagnosis of AD can potentially mitigate disease impact (Ewers, Sperling, Klunk, Weiner, & Hampel, 2011; Pellegrini et al., 2018; Rathore, Habes, Iftikhar, Shacklett, & Davatzikos, 2017). Brain morphology is recognized as a biological marker of the AD progression from preclinical to overt stages of AD (Frisoni, Fox, Jack, Scheltens, & Thompson, 2010). Structural MRI has therefore been incorporated into the clinical assessment of AD.

* Correspondence to: National University of Singapore, 4 Engineering Drive 3, Block E4 04-08, 117583, Singapore.
E-mail address: bieqa@nus.edu.sg (A. Qiu).

[1] Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging 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.

https://doi.org/10.1016/j.neunet.2022.10.032
0893-6080/

Deep learning methods have increasingly been used in the computer-aided diagnosis of AD due to their flexibility and ability to learn brain image features that have the most discriminative power of AD diagnosis (e.g., Ansart et al., 2021; Jin et al., 2020; Tanveer et al., 2020; Wen et al., 2020). Particularly, *convolutional neural network* (CNN) on brain structural images shows its potential to distinguish normal aging and AD (see recent review in Wen et al. (2020)). Most of the existing CNN studies are applied to high-resolution structural images of the brain Bäckström, Nazari, Gu, and Jakola (2018), Basaia et al. (2019), He, Zhang, Ren, and Sun (2016). Nevertheless, in a clinical setting, high-resolution structural MRI images may not be obtained due to limited clinical acquisition time. Also, brain images vary in terms of image quality and acquisition protocols from one hospital to another (Bottani et al., 2022). It remains unclear how well existing CNNs can be applied to the diagnosis of AD in the clinical setting.

In the past ten years, a substantial body of research mainly employs CNNs on 2D slices (Aderghal et al., 2018; Cheng & Liu, 2017; He et al., 2016), 3D patches, 3D regions of interest (ROIs) (Liu, Ji and Qiu, 2021), 3D whole-brain images (Huang, Chung, & Qiu, 2021; Wee et al., 2019), or volumetric features of brain structural images (see review in Liu et al. (2018)). Most of CNN studies on 2D slices take advantage of the existing CNN architectures on natural images, such as ResNet (He et al., 2016), Inception (Cui et al., 2019), VGGNet (Nigri, Ziviani, Cappabianco, Antunes, & Veloso, 2020; Qiu et al., 2018), AlexNet (Lee, Ellahi, & Choi, 2019; Liu, Li et al., 2021), GoogLeNet (Liu, Li et al., 2021; Sarraf, DeSouza, Anderson, & Tofighi, 2016), and etc., where one or a few of 2D slices of brain structural images are taken as inputs of 2D-slice CNNs. Valliani and Soni (2017) demonstrated the usefulness of the CNNs pre-trained on natural images in the AD classification. ResNet performs better than VGGNet (Nigri et al., 2020) and a baseline CNN with one convolutional layer and two fully connected layers (Valliani & Soni, 2017). The 2D-slice CNN approaches are computationally efficient and less dependent on image resolution, increase the samples by the number of slices per scan, and may only require minimal image processing (Wen et al., 2020). But, the 2D-slice CNNs largely ignore the fact that the brain is a 3-dimensional object and hence have a low AD diagnosis accuracy. Also, the selection of slices is not straightforward.

In contrast, 3D CNN approaches on whole-brain images can incorporate the 3-dimensional spatial relevance of the brain (e.g. Dickerson et al., 2001; Qiu et al., 2018; Salvatore et al., 2015; Valliani & Soni, 2017). In 3D CNN approaches, the dimensionality of estimated parameters is high, more samples are therefore required, and the computational time can be intensive. This can be partially solved while 3D patches or 3D ROIs are considered as inputs of 3D CNNs. CNNs on 3D patches, similar to 2D-slice CNNs, reduce the computational burden, but the selection and size of patches can be tricky (Qiu et al., 2018; Valliani & Soni, 2017). On the other hand, CNNs on 3D ROIs incorporate prior knowledge of brain regions that are well-known to be affected early in AD, such as the hippocampal and medial temporal ROIs (Dickerson et al., 2001; Salvatore et al., 2015). The determination of 3D ROIs is dependent on image acquisition and needs intensive image processing, such as brain segmentation and registration. A recent review in Wen et al. (2020) implemented the CNNs on 2D slices, 3D patches, and ROIs, as well as 3D whole-brain images, and provided the most comparable classification results among these methods. The 3D CNN approaches (whole-brain images, ROIs, patches) achieved the AD classification accuracy of 74%∼88% from normal aging, while the 2D-slice CNN obtained the accuracy of 79%.

Recently, the ABCD Neurocognitive Prediction Challenge (2019 ABCD-NP-Challenge; Oxtoby et al., 2019) invited researchers to submit machine learning methods for predicting fluid intelligence from brain structural MRI. Most of the methods with top prediction performance employed ensemble approaches. For instance, Vang, Cao, and Xie (2019) incorporated a gradient boosting machine (GBM) into traditional 3D CNN, where GBM obtains a strong predictor by ensembling many weak predictors via adding a new estimator fitted to the residual of the model and true labels. Several studies trained multiple machine learning models, such as multiple regressors (Kao, Zhang, Goebel, Chen, & Manjunath, 2019) or multiple 3D ResNet (Guerdan et al., 2019), or different machine learning models (Tamez-Pena, Orozco, Sosa, Valdes, & Nezhadmoghadam, 2019), and ensembled their predictors via voting, averaging, or stacking. Building multiple base learners and assembling weak predictors to become a strong predictor is common among these successful approaches. Ganaie and Tanveer (2022) recently developed an ensemble of deep learning models that can learn highly complicated patterns from MRI scans for the detection of AD by utilizing diverse solutions. Nevertheless, the number of weak predictors is limited to the number of base learners and ensemble approaches. The more base learners are built, the more costly the computation is. To avoid such issues, a recent study proposed an ensemble deep random vector functional link network that optimizes a single network and generates an ensemble via optimization at different levels of random projections of the data (Ganaie & Tanveer, 2022). Moreover, an intuitionistic fuzzy random vector functional link network aimed to find a weighting scheme for auto-detection of outliers and noise samples (Malik, Ganaie, Tanveer, Suganthan, & Initiative, 2022).

This study takes the advantage of the 2D-slice CNN fast computation and the idea of ensemble approaches (Ganaie & Tanveer, 2022; Malik et al., 2022) to develop a Monte Carlo Ensemble Neural Network (MCENN) by introducing Monte Carlo sampling and an ensemble neural network in the integration with ResNet50. Our goals are to improve the 2D-slice CNN performance and to design the MCENN model insensitive to image resolution. Hence, the MCENN model first incorporates the existing architecture of ResNet50 that shows the best performance on the AD classification when compared to VGGNet (Nigri et al., 2020) and CNN with a few convolutional layers (Valliani & Soni, 2017). Nevertheless, ResNet50 cannot well characterize the 3-dimensional spatial relevance of the brain. We adopt the concept of the ensemble to recover 3-dimensional information from 2D slices to improve the classification performance of ResNet50. Unlike traditional ensemble approaches with multiple base learners, our study develops one base neural network learner to generate a large number of possible decisions via Monte Carlo sampling of feature importance within the combined slices to boost classification performance. In this setting, our MCENN model can overcome the main weakness of the lack of 3D brain anatomical information in the 2D-slice CNNs and develop a neural network to learn the 3D relevance of the features across multiple slices. Moreover, the MCENN model only has one base learner and is computationally efficient. Furthermore, our framework allows a large sampling rate, and hence our classification performance is stable based on the laws of large numbers. Furthermore, when a large sampling rate of image slices is used, it is equivalent to having low-resolution images in our MCENN model. Therefore, we expect that our model will not be sensitive to brain image resolution, which makes it feasible to be adopted in a clinical setting.

In our experiments, brain images from Alzheimer's Disease Neuroimaging Initiative (ADNI, 7199 scans), the Open Access Series of Imaging Studies-3 (OASIS-3, 1992 scans), and a clinical sample (239 scans) are used to evaluate the performance of the MCENN model for the classification of cognitively normal (CN),

patients with mild cognitive impairment (MCI) and AD. Our experiments demonstrate the minimal number of 2D slices needed in the MCENN model. The performance of our model is compared with the existing 2D-slice CNNs and 3D CNNs on brain images. Finally, this study designs experiments to illustrate whether our model is sensitive to image resolution.

Hence, this paper contributes to the following novelty:

- a large number of possible decisions is generated via Monte Carlo sampling of feature importance among MRI 2D slices;
- one learner achieves learning the 3D relevance of the brain anatomy via the interaction of the features of the 2D slices randomly chosen by the MCENN model;
- the MCENN model performs better than existing 2D-slice and 3D CNNs in the AD classification;
- the MCENN performance is not sensitive to the MRI image resolution;
- a new deep learning framework for the AD classification is clinically applicable.

2. Methods

This section describes our MCENN model in the integration with ResNet50. Fig. 1 shows the overall architecture employed in this study. This architecture is designed to achieve (1) the feature extraction of 2D slices; (2) the integration of 2D information; (3) the generation of a large number of possible decisions.

2.1. ResNet50

The MCENN model first adopts ResNet50 (He et al., 2016) to extract features from each 2D slice. ResNet50 is chosen because it performs well in comparison with GoogLeNet and VGG (Nigri et al., 2020). In particular, ResNet50 is a deep convolutional neural network that is made of 16 ResBlocks (see Fig. 1b). Three convolutional layers are made up of one ResBlock in which two convolutional layers have filters with a kernel size of 1×1 and one layer has filters with a kernel size of 3×3. One more convolutional layer is added in the input layer, one convolutional layer and one fully connected layer are added to the output layers of ResNet50, which is made up of the 50-layer architecture. ResNet50 is deep so that it can learn rich feature representations for a wide range of images. Hence, our study applies it to all MRI slices in the axial, coronal, and sagittal views of brain images and map each 2D slice to a feature vector with a length of 2048.

2.2. Monte Carlo Ensemble Neural Network (MCENN)

The MCENN model is designed to combine the features of 2D slices, obtain the distribution of possible decisions via one base learner, and make a final decision based on this distribution.

Denote $\mathcal{X} = \{X^i(s,d)\}_{i=1}^n$ as a set of image features obtained from ResNet50 for all n subjects. $X^i(s,d)$ represents the image features of the ith subject that correspond to slice, s, and feature dimension, d. In this study, $X^i(s,d)$ is obtained from the above ResNet50 (see Fig. 1b), where $s = 1, 2, \ldots, 368$, $d = 1, 2, \ldots, 2048$. The MCENN model first introduces two sampling functions, π_s and π_d, where π_s is used to randomly sample 2D slices with a sampling rate of r_s and π_d is used to randomly sample the feature space at a sampling rate of r_d. In particular, π_s is designed such that the higher sampling frequency is for the slices with a greater discriminative power of disease. Here, the discriminative power of each slice is evaluated via ResNet50 using training data. More detail is given in the implementation section. In contrast, the feature space is sampled via a uniform distribution because the feature space includes a wide range of feature attributes generated from ResNet50 on all slices.

Define a function f that maps the mth sampled features of the ith subject, X^i, to a probability of disease. The MCENN model can be written in the form of

$$F(X^i) = \sum_{m=1}^{M} \rho_m f\left(X^i(\pi_s^m, \pi_d^m)\right), \tag{1}$$

where M is the total number of samples drawn via π_s and π_d on X^i. ρ_m is the weight for the mth sampled features, which is used to aggregate all possible decisions obtained from f. This study designs a neural network to represent f so that it can learn the interaction of the features across the sampled slices. We exploit one fully connected (FC) layer neural network (the simplest neural network) and complicated neural networks, such as DenseNet121 (Huang, Liu, Maaten, & Weinberger, 2017), VGG (Qiu et al., 2018), GoogLeNet (Liu, Li et al., 2021), and ResNet (He et al., 2016) to explore the relationship among the slice features. We exploit the simple one FC layer neural network due to its computational efficiency. On the other hand, DenseNet121 is chose because (1) it can capture possible non-linear relations of the features across 2D slices if any; (2) the accuracy of the classification of normal aging and AD using DenseNet (82.4%) is better than VGG (51.1%), GoogLeNet (68.6%), and ResNet (70.4%). Hence, we only employ DenseNet and demonstrate whether there is a need of the complexity for the neural networks in the following experiments (see Section of Results).

2.3. Evaluation metrics

Three traditional metrics, including classification accuracy, sensitivity, and specificity, are employed to quantify the classification performance. They are defined as

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP}, \tag{2}$$

$$Sensitivity = \frac{TP}{TP + FN}, \tag{3}$$

$$Specificity = \frac{TN}{TN + FP}, \tag{4}$$

where TP, TN, FN, and FP denote the true positive, true negative, false negative, and false positive, respectively. In our study, the positive class is "AD" and the negative class is "CN". The sensitivity and specificity provide the proportion of correctly identified samples for positive and negative classes, respectively. However, these measures, including the classification accuracy, are sensitive to the ratio of the number of subjects in the positive and negative classes, and hence may provide inaccurate and misleading information on the performance of a classifier on an imbalanced dataset (López, Fernández, García, Palade, & Herrera, 2013). To overcome this issue and to take into consideration the ratio of the number of subjects in the positive and negative classes, this study uses geometric mean, defined as

$$Geometric\ Mean = \sqrt{SEN \times SPE}. \tag{5}$$

Geometric mean attempts to maximize the accuracy of each of the two classes when the number of subjects in the positive and negative classes is imbalanced (Barandela, Sánchez, García, & Rangel, 2003).

Moreover, area under the receiver operating characteristic curve (AUC) is also employed to measure the quality of the model's prediction irrespective of what classification threshold is chosen. The AUC ranges in value from 0 to 1. A model whose predictions are 100% wrong has an AUC of 0; one whose predictions are 100% correct has an AUC of 1.0.

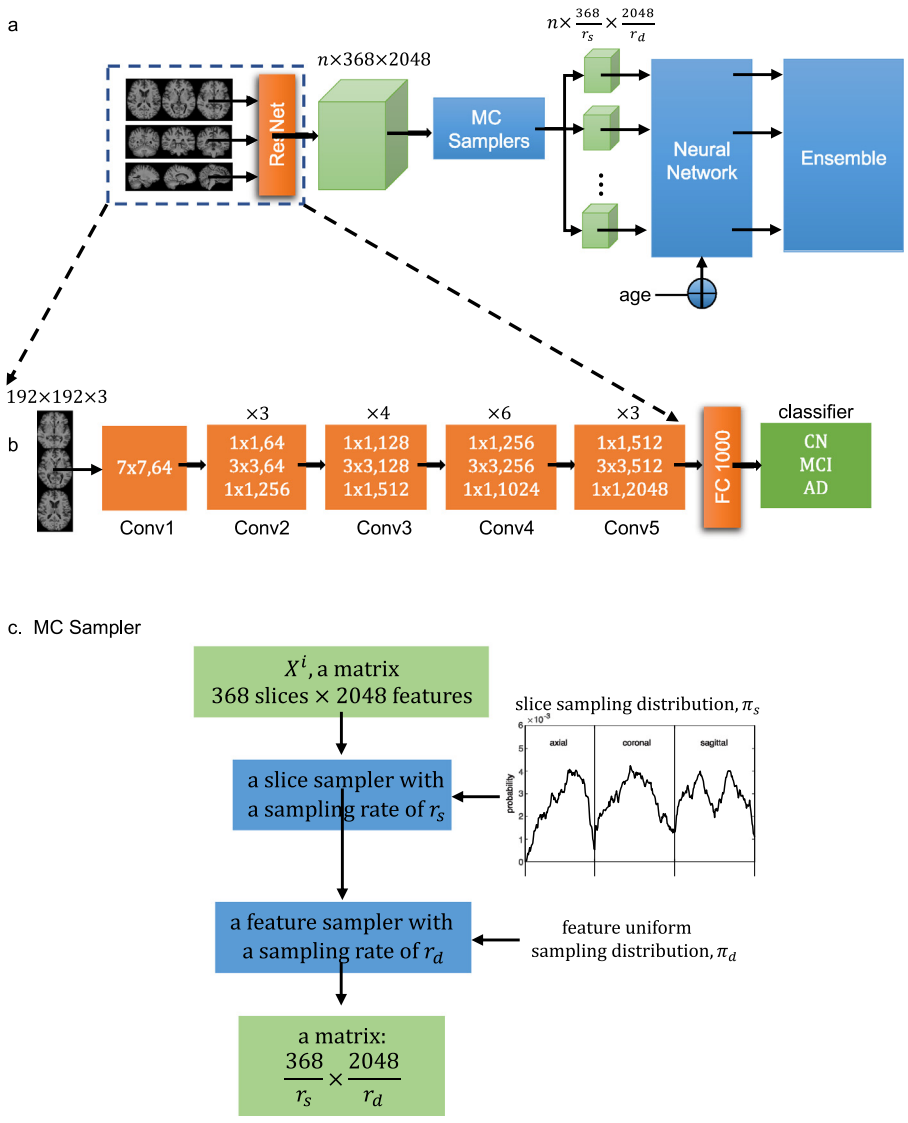

Fig. 1. The architecture of the Monte Carlo Ensemble Neural Network (MCENN) in the integration with ResNet50. Panel (a) shows the overall architecture of the MCENN model. Panel (b) illustrates the detailed architecture of ResNet50 that is adopted from He et al. (2016). Panel (c) illustrates the flowchart of the MC sampler. n represents the number of subjects. The total number of 2D slices from the axial, coronal, and sagittal views is 368 in this study. ResNet extracts 2048 features from each slice. r_s and r_d denote the sampling rates in the dimensions of 2D slices and features, respectively.

2.4. Implementation

The framework (see Fig. 1) is implemented in Python 3.7 and TensorFlow 1.13.1 library. All experiments are run using NIVIDIA Tesla V100-SXM2 GPU with 32 GB RAM and Intel Xeon Gold 5118 CPU with 2.30 GHz. This study divides the ADNI cohort into two datasets (see Fig. 2). All scans from one subject are assigned to one of the two datasets to avoid data leakage. In general, a two-step procedure is used to train our framework, one for ResNet50 and the other for the MCENN model. Nevertheless, the detailed description of training and testing data is provided for each experiment in Section 4.

ResNet50 training. This study first modifies the last fully connected layer of ResNet50 for a three-class classification problem (CN, MCI, AD). The slice of interest and the slice before and after it form an RGB image as the input of ResNet50. ResNet50 is trained in two ways (fully trained, transfer learning) based on the first ADNI dataset. First, stochastic gradient descent is employed to train the full model of ResNet50. Second, we take advantage of the ResNet50 model pre-trained on more than a million images from the ImageNet database and fine-tune the last two convolutional layers and the last fully connected layer of ResNet50. Both training approaches employ a batch size of 64, an initial learning rate of 0.01, and 55 epochs. The learning rate is gradually decayed to 0.005, 0.001, 0.0005, 0.0001 at epoch of 19, 30, 44, 53.

Moreover, the sampling distribution of slices is defined as their classification accuracies obtained from ResNet50 based on the first ADNI dataset.

MCENN training. The second ADNI dataset is employed to train and evaluate the performance of the MCENN model. 50% of subjects are used in the training and 50% are used in the evaluation. Each experiment needs to first determine the sampling rates, r_s and r_d, and the total number of samples, M. Both the neural network with one FC layer and DenseNet121 for a two-class classification problem are trained via stochastic gradient descent. We maximize the GM metric to balance the sensitivity and specificity of the neural network. The training parameters are defined as follows: a batch size is 32; learning rate values are [0.01, 0.005, 0.001, 0.0005, 0.0001] at epoch of

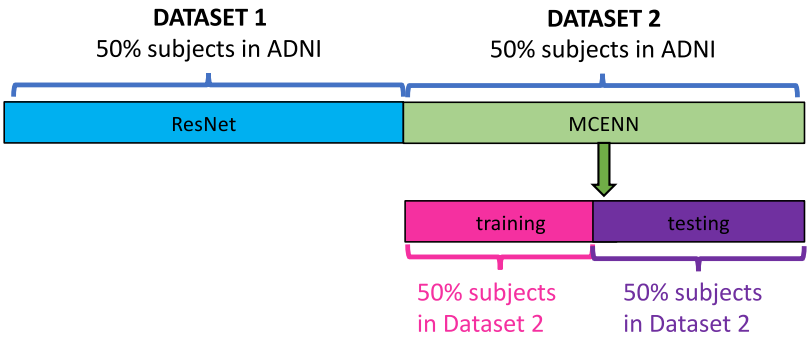

Fig. 2. The ADNI data splitting for the ResNet and MCENN training and testing.

[0, 19, 30, 44, 53], and the number of epochs is 80. Since age is an important factor for the diagnosis of AD, age with the image features ($\frac{368}{r_s} \times \frac{2048}{r_d}$) is incorporated into the feature space for the neural network.

Hyperparameters of the MCENN model. As mentioned above, several key hyperparameters, including r_s and r_d, and the total number of samples, M, determine the MCENN model. Our experiments below will discuss how to choose these parameters (see Section 4).

Code Availability. The code and demo are available at https://github.com/bieqa/Monte-Carlo-Ensemble-Neural-Network.

3. MRI data and analysis

ADNI and OASIS-3. Data used in this study were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (https://adni.loni.usc.edu) and the Open Access Series Of Imaging Studies-3 (OASIS-3; http://oasis-brains.org). Institutional review boards approved study procedures across participating institutions.

This study includes the ADNI cohort, including the ADNI-1 ($n = 811$), ADNI-GO($n = 188$) and ADNI-2($n = 1019$) studies. The number of visits per subject varied from 1 to 12. At each visit, subjects were diagnosed as cognitively normal (CN), mild cognitive impairment(MCI), or Alzheimer's disease(AD) based on the criteria described in the ADNI protocol. The total numbers of scans in individual diagnostic groups are 2190 for CN (546 subjects), 3393 for MCI (910 subjects), and 1616 for AD (592 subjects), respectively. This study excludes 648 scans of 115 subjects whose diagnosis was converted back from MCI to CN across time.

In the OASIS-3 dataset, the number of visits per subject varied from 1 to 7. There were 1531 scans for CN (712 subjects) and 335 scans for AD (274 subjects). Table 1 provides the demographic and clinical information of the subjects in the ADNI and OASIS-3 cohorts, including age, gender, mini-mental state exam (MMSE), clinical dementia rating (CDR).

Both the ADNI and OASIS-3 cohorts acquired structural T_1-weighted MRI scans using either 1.5T or 3T scanners at different study sites. Detailed acquisition information is given at https://adni.loni.usc.edu for the ADNI and https://www.oasis-brains.org for the OASIS-3. All T_1-weighted MRI scans in these two study are in the resolution of 1 mm × 1 mm × 1.2/1.25 mm.

Clinical Sample. This study also includes a clinical sample recruited from the stroke service and Memory clinics in Singapore (Thong et al., 2014, 2013). This study was approved by the Domain-Specific Review Board (DSRB) of the National Healthcare Group. The recruitment criteria were similar to those used in the ADNI cohort. This study includes a cross-sectional dataset with 104 NC, 85 MCI, and 50 AD. Table 1 lists the demographic and clinical information of these subjects.

Table 1
Demographic and clinical information of the ADNI, OASIS-3, and clinical samples.

ADNI			
	CN	MCI	AD
Number of subjects*	546	910	592
Number of MRI scans	2190	3393	1616
Female/Male	1095/1095	1376/2017	700/916
Age (mean ± SD)	76.0 ± 6.2	74.3 ± 7.7	76.0 ± 7.4
MMSE (mean ± SD)	29.0 ± 1.2	27.4 ± 2.3	21.9 ± 4.3
CDR sum of box (mean ± SD)	0.1 ± 0.3	1.6 ± 1.1	5.4 ± 2.6
OASIS-3			
Number of subjects*	712	102	274
Number of MRI scans	1531	126	335
Female/Male	923/608	58/68	147/188
Age (mean ± SD)	69.0 ± 9.3	75.0 ± 8.4	76.9 ± 8.3
MMSE (mean ± SD)	29.0 ± 1.4	27.9 ± 2.7	24.0 ± 5.1
CDR sum of box (mean ± SD)	0.1 ± 0.5	1.1 ± 1.4	4.2 ± 3.4
Clinical sample			
Number of subjects	104	85	50
Number of MRI scans	104	85	50
Female/Male	40/64	49/36	33/17
Age (mean ± SD)	66.6 ± 4.7	74.1 ± 6.4	76.7 ± 7.6
MMSE (mean ± SD)	28.0 ± 0.9	20.8 ± 3.6	16.3 ± 4.4
CDR sum of box (mean ± SD)	0.0 ± 0.0	1.1 ± 0.9	6.8 ± 2.8

All the subjects in this clinical sample underwent MRI scans that were performed on a 3T Siemens Magnetom Trio Tim scanner using a 32-channel head coil at the Clinical Imaging Research Centre of the National University of Singapore. The image protocol was T1-weighted Magnetization Prepared Rapid Gradient Recalled Echo (MPRAGE; 192 slices, 1 mm × 1 mm × 1 mm, field of view = 256×256 mm, matrix = 256 × 256, repetition time = 2300 ms, echo time = 1.9 ms, inversion time = 900 ms, flip angle = 9°).

Structural MRI Analysis. In this study, all structural images are minimally processed via bias field correction, rigid transformation (rotation and translation), intensity normalization, and skull stripping. First, non-parametric non-uniform intensity normalization (N3) is used to correct for intensity non-uniformity in structural T_1-weighted MRI images (Sled, Zijdenbos, & Evans, 1998). Second, each image is registered to the MNI space via linear transformation found using FLIRT (Jenkinson & Smith, 2001). Third, image intensity is re-scaled such that the mean intensity of the white matter is rescaled to 110. Finally, the watershed algorithm in Ségonne et al. (2004) is used to remove the brain skull.

In this study, 112 axial, 128 coronal, and 128 sagittal slices are extracted from each structural image. Each slice is zero padded to be a size of 192 at the boundary. These slices are used as inputs in our MCENN model.

Table 2
Effects of random feature sampling (r_d) on the MCENN performance of the control and Alzheimer's disease classification.

r_d	Accuracy (%)	Sensitivity (%)	Specificity (%)	Geometric mean (%)	Area under curve (%)
Neural network with one fully connected layer					
No sampling	87.6 ± 1.1	83.3 ± 1.7	92.1 ± 2.3	87.6 ± 1.1	89.4 ± 1.1
2	87.8 ± 0.5	83.5 ± 1.4	92.5 ± 0.8	87.8 ± 0.5	89.2 ± 0.1
4	88.2 ± 0.2	83.1 ± 0.6	93.7 ± 0.5	88.2 ± 0.2	89.5 ± 0.1
8	88.2 ± 0.3	83.5 ± 1.3	93.2 ± 1.2	88.2 ± 0.3	91.7 ± 0.1
16	88.2 ± 0.4	82.1 ± 1.0	94.8 ± 1.0	88.2 ± 0.4	90.9 ± 0.1
32	88.7 ± 0.5	81.8 ± 0.8	95.9 ± 0.9	88.6 ± 0.5	92.6 ± 0.1
DenseNet121					
No sampling	82.4 ± 3.0	73.3 ± 6.0	92.2 ± 3.7	82.1 ± 3.3	85.2 ± 3.2
2	85.7 ± 0.6	74.7 ± 1.5	97.5 ± 0.5	85.3 ± 0.7	91.5 ± 0.6
4	71.1 ± 10.2	44.3 ± 20.0	99.7 ± 0.4	64.1 ± 17.2	82.3 ± 0.5
8	88.1 ± 0.5	82.8 ± 2.4	93.8 ± 2.1	88.1 ± 0.5	82.6 ± 0.6
16	75.0 ± 5.6	52.3 ± 11.1	99.3 ± 0.5	71.6 ± 7.7	90.0 ± 0.2
32	80.7 ± 7.9	64.3 ± 16.4	98.1 ± 1.4	78.6 ± 10.8	89.6 ± 0.2

4. Results

This section designs experiments to explore how our framework works. The classifier of CN vs AD is employed to demonstrate the use of the MCENN models in the following aspects: (1) effects of random feature and slice sampling in the MCENN model; (2) the selection of neural networks in the MCENN model (the neural network with one FC layer and DenseNet121). Our results report the computation time for each experiment. Moreover, our model compares with existing 2D-slice and 3D CNN methods based on the AD classification performance. Furthermore, we examine the robustness of the MCENN model using the ADNI and OASIS-3 datasets as well as a clinical sample. Finally, effects of image resolution on the performance of the MCENN model are investigated to demonstrate its potential use in a clinical setting.

4.1. Effects of random feature sampling

The first experiment is designed to examine effects of random feature sampling. For this, the full ResNet50 model is trained using the first ADNI dataset for a three-class classification problem (the total number of scans: 2583; CN:918; MCI, 1160; AD:505). The second ADNI dataset (the total number of scans: 2018; CN:936; AD:1082) is used to train and evaluate the MCENN model for the classification of CN and AD. As mentioned before, all the scans from the same subject are assigned to one dataset to avoid data leakage. In this experiment, we set M as 100, ρ_m as equal weight in Eq. (1), and no slice sampling (i.e., $r_s = 1$). The MCENN model is optimized based on the procedure described in Section 2.4. At the mth trial, the MCENN model first uniformly samples $368 \times \frac{2048}{r_d}$ features from each scan and then computes the classification probability from the neural network. This is repeated 100 trials. The final classification decision for this MRI scan is made by averaging these 100 classification probabilities. To evaluate the performance of the MCENN model, it is trained using 50% of the second ADNI dataset and evaluated its performance using the rest of 50%. This process is repeated 5 times.

Table 2 lists the mean and standard deviation of the classification results among the 5 repetitions for $r_d \in \{2, 4, 8, 16, 32\}$. The neural network with one FC layer shows that the classification accuracy and geometric mean for all $r_d \in \{2, 4, 8, 16, 32\}$ are not statistically different from those without the feature sampling (Student $t-$tests: $t < 1.78$, $p > 0.11$). Nevertheless, the results from DenseNet121 show that the classification accuracy (Student $t-$test: $t = 6.63$, $p < 0.001$) and geometric mean at $r_d = 8$ (Student $t-$test: $t = 5.84$, $p < 0.001$) are statistically larger than those without the feature sampling. When $r_d \in \{4, 16\}$, the classification accuracy and geometric mean obtained from DenseNet121 are statistically smaller than those without the feature sampling (Student $t-$tests: $t < -2.33$, $p < 0.04$). The same pattern is observed in AUC. As listed in Table 2, the standard deviation of the classification accuracy, sensitivity, specificity, geometric mean, and AUC from DenseNet121 is relatively larger than that from the neural network with one FC layer.

4.2. Effects of random slice sampling

In this section, the experiments are designed to examine the effects of random slice sampling without feature sampling. Instead of uniform sampling, the slice sampling is designed based on the distribution shown in Fig. 3. Here, the sampling distribution is computed as the AD probability obtained from ResNet50 when a slice is used to distinguish CN and AD. Clearly, the slices encompassing the hippocampus in all three views give the highest probability (see Fig. 3). The experiments in this section are setup in the same way as those described in the previous section.

Table 3 lists the mean and standard deviation of the classification results among the 5 repetitions for $r_s \in \{2, 4, 8, 16, 32\}$. The neural network with one FC layer shows that the classification accuracy, geometric mean, and AUC for $r_s = \{2, 4, 8\}$ are significantly better than those without the slice sampling (Student's $t-$tests: $t > 2.47$, $p < 0.04$). The highest accuracy occurs when $r_s = 4$, while the highest AUC occurs when $r_s = 8$. This finding suggests the importance of introducing a large number of possible decisions via Monte Carlo sampling of slices to make a final classification decision.

Similarly, DenseNet121 also shows that when $r_s = 8$ the classification accuracy and geometric mean are significantly better than those without the slice sampling (Student's $t-$tests: $t > 2.60$, $p < 0.03$).

4.3. Effects of neural network with one FC layer and DenseNet121 in the MCENN model

Both Tables 2 and 3 show that the performance of the neural network with one FC layer is statistically better than that of DenseNet121 at a given r_d or r_s (Student's t-tests: all $p < 0.05$). This suggests that the non-linear relationship of image features across slices may not need to be learned through a network with a great depth, such as DenseNet121. This is partly because ResNet50 learns features from each slice through many non-linear operations. Moreover, the standard deviation of the classification accuracy obtained from DenseNet121 is larger than that obtained from the neural network with one FC layer. This indicates that DenseNet121 performance is not stable and may need more samples to train.

Fig. 4 shows the computational time per epoch for the neural network with one FC layer and DenseNet121 that are used in

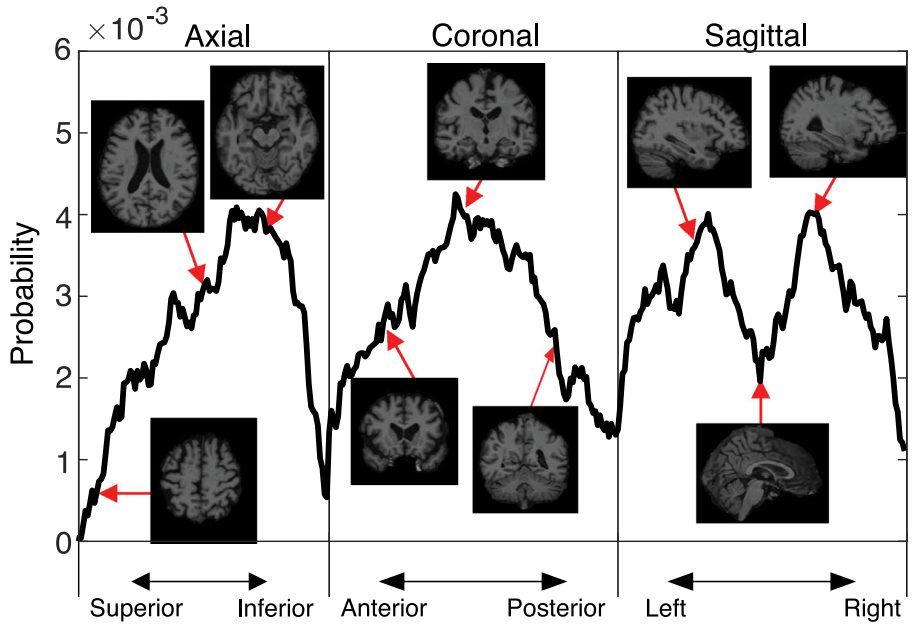

Fig. 3. The sampling distribution of the MRI slices. The panels from left to right respectively show the sampling probability for axial, coronal, and sagittal slices. Selected MRI slices illustrate the brain anatomy contributing to the discrimination between normal aging and Alzheimer's disease.

Table 3
Effects of random slice sampling (r_s) on the MCENN performance of the normal aging and Alzheimer's disease classification.

r_s	Accuracy (%)	Sensitivity (%)	Specificity (%)	Geometric mean (%)	Area under curve (%)
Neural network with one fully connected layer					
No sampling	87.6 ± 1.1	83.3 ± 1.7	92.1 ± 2.3	87.6 ± 1.1	89.4 ± 1.1
2	89.3 ± 0.7	84.3 ± 3.1	94.7 ± 2.6	89.3 ± 0.8	91.3 ± 0.9
4	90.0 ± 0.7	83.5 ± 1.4	96.9 ± 0.5	89.9 ± 0.7	91.4 ± 0.4
8	89.8 ± 0.5	83.0 ± 1.5	97.0 ± 1.4	89.7 ± 0.6	92.5 ± 0.7
16	89.0 ± 1.0	81.7 ± 2.8	96.8 ± 1.6	88.9 ± 1.0	92.3 ± 1.7
32	87.7 ± 1.5	77.5 ± 3.4	98.5 ± 0.6	87.4 ± 1.8	92.4 ± 1.0
DenseNet121					
No sampling	82.4 ± 3.0	73.3 ± 6.0	92.2 ± 3.7	82.1 ± 3.3	85.2 ± 3.2
2	82.8 ± 5.0	70.9 ± 9.1	95.5 ± 2.5	82.1 ± 5.6	87.7 ± 0.6
4	72.2 ± 14.7	50.5 ± 28.8	95.4 ± 6.8	61.2 ± 32.1	66.5 ± 1.9
8	87.4 ± 1.9	78.5 ± 4.0	96.8 ± 0.6	87.1 ± 2.1	83.7 ± 0.8
16	77.9 ± 14.4	59.5 ± 27.1	97.4 ± 1.3	72.5 ± 24.0	89.1 ± 1.4
32	79.7 ± 14.6	65.7 ± 24.6	94.7 ± 4.5	77.1 ± 19.2	88.0 ± 1.8

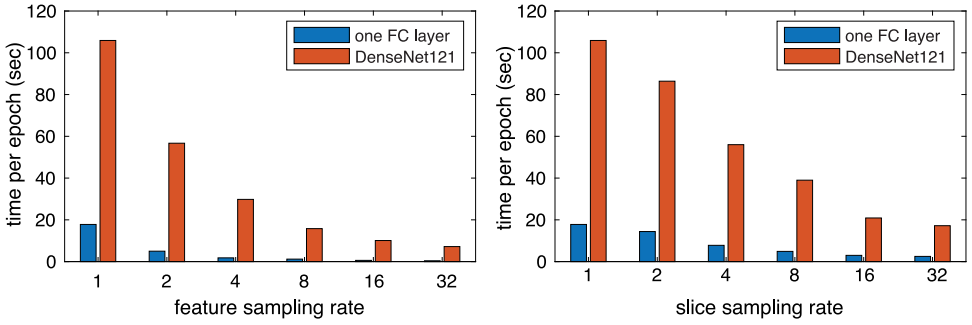

Fig. 4. Computational time per epoch for the neural network with one FC layer and DenseNet121 at different feature (left panel) and slice (right panel) sampling rates. The one FC layer performs faster than DenseNet121.

the experiments listed in Tables 2 and 3. At the same sampling rate (either feature (left panel) or slice (right panel)), the neural network with one FC layer is computationally more efficient than DenseNet121. Moreover, our MCENN with one FC layer was comparable to the 2D ResNet in terms of the computation time. Hence, we employ the MCENN with one FC layer in the following experiments.

4.4. The MCENN performance

The performance of the MCENN with one FC layer is assessed when the slice sampling rate is 4 or 8 without feature sampling or the feature sampling rate of 4. Again, the setting of the experiments is the same as those in the above sections. Table 4 shows that the MCENN can classify the CN and AD subjects at the accuracy of 88.1%∼90.0% and at the AUC of 91.4%∼93.8%. There

Table 4
The performance of the MCENN with one FC layer at the slice sampling rate of $r_s = 4, 8$ and feature sampling rate of $r_d = 1, 4$.

Neural network with one fully connected layer						
r_s	r_d	Accuracy (%)	Sensitivity (%)	Specificity (%)	Geometric mean (%)	Area under curve (%)
4	No sampling	90.0 ± 0.7	83.5 ± 1.4	96.9 ± 0.5	89.9 ± 0.7	91.4 ± 0.4
8	No sampling	89.8 ± 0.5	83.0 ± 1.5	97.0 ± 1.4	89.7 ± 0.6	92.5 ± 0.7
4	4	89.5 ± 0.4	81.8 ± 1.0	97.6 ± 0.2	89.4 ± 0.5	93.8 ± 0.1
8	4	88.1 ± 0.9	78.6 ± 2.0	98.3 ± 0.4	87.9 ± 1.0	93.2 ± 0.5

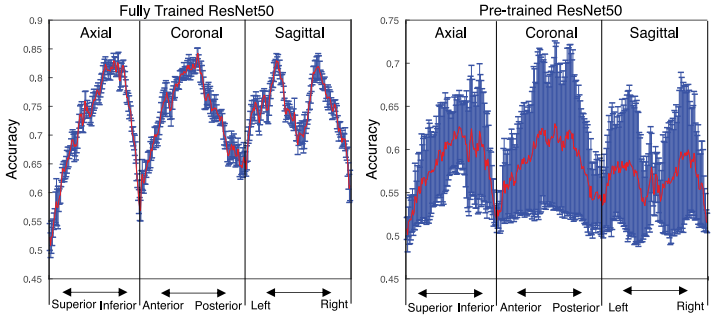

Fig. 5. The normal aging and Alzheimer's disease classification accuracy using the fully trained ResNet50 (left panel) and the pre-trained ResNet50 (right panel). The red line represents the mean classification accuracy among the 5 trials and the blue bar indicates the respective standard deviation. This figure suggests that the fully trained ResNet50 model performs better than the pre-trained ResNet50 model.

is no statistical difference in the MCENN performance when the slice sampling rate is 4 or 8 without the feature sampling or the feature sampling rate of 4. Our results suggest that 46 (368/8) slices of the MRI scan can produce the classification accuracy comparable to that when all 368 slices are used. This indicates the feasibility of adopting this setting for clinical brain scans.

Without adding age as one of features, the MCENN with one FC layer can also achieve the AUC of 93.2% ± 0.5% when the slice sampling rate is 8 without the feature sampling. This is comparable to that with age as one of features in the MCENN model (AUC: 92.5% ± 0.7%).

4.5. Comparisons with 2D-slice and 3D CNNs

In this study, our MCENN model is compared with the existing 2D-slice CNNs, 3D-patch, 3D-ROI, and 3D whole-brain CNNs on the classification of CN and AD using the ADNI dataset. We choose these existing methods as they have been extensively reviewed and evaluated fairly in the recent review (Wen et al., 2020).

For the 2D-slice CNNs, the ResNet50 is fully trained using the first ADNI dataset for the classification of CN and AD (fully trained ResNet50). In addition, the ResNet50 that is pre-trained using the ImageNet database and its last two convolutional layers and the last FC layer are fine-tuned using the first ADNI dataset (pre-trained ResNet50). Both the fully trained and pre-trained ResNet50 are fine-tuned using 50% of the second ADNI dataset and evaluated using another 50% of the second ADNI dataset. This process is repeated 5 times. Fig. 5 illustrates the mean and standard deviation of the classification accuracy of each slice. Table 5 lists the highest classification accuracies of the fully trained ResNet50 and pre-trained ResNet50 among all 368 slices. Student's t-test suggests that our MCENN model performs significantly better than the best classification results obtained from the fully trained ResNet50 (Student's $t-$test: $t = 9.95, p < 0.001$) and the pre-trained ResNet50 (Student's $t-$test: $t = 6.35, p < 0.001$).

For the 3D CNNs, we directly borrow the results from Wen et al. (2020) for the fair comparisons as the 3D CNN models have been well learned. Here, the 3D-patch CNN consists of 4 convolutional blocks and 3 FC layers. It is trained in three ways. The first one is that 36 patches of the size of 50×50 mm^3 from each image

Table 5
Comparisons of the MCENN model with the 2D-slice and 3D CNNs in the classification of normal aging and AD of the ADNI dataset. The results for 3D CNN models are borrowed from Table 6 in Wen et al. (2020).

Model	Accuracy (mean ± SD)	Accuracy of 5 repetitions of 5 repetitions
MCENN	0.90 ± 0.01	0.89, 0.90, 0.90, 0.91, 0.90
2D-slice CNNs		
Fully trained ResNet50	0.84 ± 0.01	0.83, 0.84, 0.84 0.84, 0.86
Pre-trained ResNet50	0.63 ± 0.10	0.54, 0.52, 0.72, 0.69, 0.69
3D CNNs		
3D-patch single-CNN	0.74 ± 0.08	0.75, 0.84, 0.78, 0.75, 0.59
3D-patch multi-CNN	0.81 ± 0.03	0.82, 0.84, 0.83, 0.77, 0.79
3D-ROI CNN	0.88 ± 0.03	0.84, 0.89, 0.90, 0.89, 0.85
3D whole-brain CNN	0.82 ± 0.05	0.74, 0.90, 0.83, 0.77, 0.83

are fitted into one single 3D-patch CNN (3D-patch single-CNN). The second is that one 3D-patch CNN is for one patch and there are a total 36 CNN models (3D-patch multi-CNN). For the 3D-ROI CNN model, only the ROI enclosing the hippocampus with the size of 50×50 mm^3 is fitted to the 3D-patch CNN. Last but not least, the 3D whole-brain CNN consists of 5 convolutional blocks and 3 FC layers. Table 5 lists the classification accuracies for all 5 times of the 3D-patch single-CNN, 3D-patch multi-CNN, 3D-ROI CNN, and 3D whole-brain CNN. Student's $t-$ tests demonstrate that the MCENN model performs better than the 3D-patch single-CNN (Student's $t-$test: $t = 3.81, p = 0.005$), 3D-patch multi-CNN ($t = 6.71, p < 0.001$), and 3D whole-brain CNN (Student's $t-$test: $t = 3.08, p = 0.015$) and is comparable to the 3D-ROI CNN (Student's $t-$test: $t = 2.08, p = 0.071$).

The MCENN model of the CN and AD classifier is then applied to distinguish stable MCI (n = 272) and MCI converted to AD (n = 240) in the second ADNI dataset. Table 6 shows that the MCENN model achieves the accuracy of 77% ± 5%, better than 3D whole-brain CNN and 3D-patch multi-CNN models (Student's $t-$test: $t > 4.38, p < 0.037$) and equivalently to 3D-ROI CNN (Student's $t-$test: $t = 1.11, p = 0.328$).

Table 6
Comparisons of the MCENN model with the 3D CNNs in the classification of stable MCI and MCI converted to AD of the ADNI dataset. The results for 3D CNN models are borrowed from Table 6 in Wen et al. (2020).

Model	Accuracy (mean ± SD)	Accuracy of 5 repetitions
MCENN	0.77 ± 0.05	0.80, 0.75, 0.70, 0.83, 0.78
3D CNNs		
3D-patch multi-CNN	0.70 ± 0.04	0.71,0.66, 0.66,0.71, 0.75
3D-ROI CNN	0.74 ± 0.02	0.75,0.72, 0.76,0.75, 0.75
3D whole-brain CNN	0.69 ± 0.04	0.68,0.71, 0.64,0.73, 0.67

4.6. Robustness of the MCENN model

The robustness of the MCENN with one FC layer is evaluated via the ADNI, OASIS-3, and clinical datasets. The full model of ResNet50 is trained using the first ADNI dataset (the total number of scans: 2583; CN:918; MCI:1160; AD:505) for a three-class classification problem. Then, the second ADNI dataset (the total number of scans: 3968; CN:936; MCI:1950; AD:1082) is employed to train and evaluate the MCENN with one FC layer for two-class classification problems in the setting of $M = 100$, $r_s = 4$, without the feature sampling. The MCENN model is trained using 50% of the second ADNI sample and evaluated via 50% of the second ADNI dataset. This is repeated 5 times. Table 7 lists the mean and standard deviation of the accuracy, sensitivity, specificity, geometric mean, and AUC for the two-class classifiers between CN and AD, CN and MCI, and MCI and AD for the ADNI dataset.

We employ transfer learning to examine the robustness of the MCENN with one FC layer. In this experiment, the MCENN with one FC layer is trained using the ADNI dataset. Its FC layer is fine-tuned using the first 50% of the OASIS-3 dataset, and evaluated using the second 50% of the OASIS-3 dataset. We repeat this experiment 5 times. Table 7 lists the accuracy, sensitivity, specificity, geometric mean, and AUC of the two-class classifiers between CN and AD, between CN and MCI, and between MCI and AD for the OASIS-3 dataset. The classification accuracy rates between CN and AD (Student's $t-$test: $t = 4.41, p = 0.002$) and between MCI and AD (Student's $t-$test: $t = 6.43, p < 0.001$) on the OASIS-3 dataset are lower than those on the ADNI dataset. This is partly because of the smaller sample sizes of MCI and AD patients and relatively better MMSE and CDR sum of box scores in the AD patients of the OASIS-3 dataset in comparison with those in the ADNI dataset (see Table 1). The classification accuracy between CN and MCI of the OASIS-3 dataset is slightly better than that of the ADNI dataset (Student's $t-$test: $t = -3.10, p = 0.015$) mainly because of younger CN in the OASIS-3 dataset (see Table 1).

The same transfer learning approach is applied to the clinical sample. Table 7 lists the classification accuracy, sensitivity, specificity, geometric mean, and AUC of the two-class classifiers between CN and AD, between CN and MCI, and between MCI and AD for the clinical sample. The ADNI and clinical sample show the comparable classification accuracy of between CN and AD (Student's $t-$ test: $t = 1.41, p = 0.109$) and between MCI and AD (Student's $t-$ test: $t = 0.15, p = 0.444$). But, the clinical sample shows better classification accuracy between CN and MCI (Student's $t-$ test: $t = 2.73, p = 0.021$), which may be due to lower scores of MMSE and CDR sum of box in the clinical sample (see Table 1).

Overall, this experiment demonstrates that the MCENN with one FC layer trained using the ADNI dataset is generalizable to the OASIS-3 and clinical samples. Nevertheless, the classification performance may depend on the clinical characteristics of the samples.

4.7. Simulation on clinical data with various image resolutions

This experiment aims to examine whether the performance of the MCENN with one FC layer is influenced by image resolution. For this, we downsample the clinical sample by factors of 2,4 and generate two datasets with the image resolution of 2 mm × 2 mm × 2 mm and 4 mm × 4 mm × 4 mm, respectively.

The MCENN with one FC layer is trained using the ADNI dataset. The last FC layer of the MCENN model was fine-tuned using 50% of the clinical dataset with a specific image resolution (1 mm × 1 mm × 1 mm, or 2 mm × 2 mm × 2 mm, or 4 mm × 4 mm × 4 mm). The MCENN model is evaluated using the other 50% of the data. This experiment is repeated 5 times. Table 8 lists the mean and standard deviation of the accuracy, sensitivity, specificity, geometric mean, and AUC for the two-class classifiers between CN and AD, between CN and MCI, and between MCI and AD at each image resolution. The images of 2 mm × 2 mm × 2 mm and 4 mm × 4 mm × 4 mm show the equivalent classification accuracy for all three classifiers when compared to the images of 1 mm × 1 mm × 1 mm (Student's $t-$ tests: $t < 0.90, p > 0.204$). Our results suggest that the performance of the MCENN model is not sensitive to the image resolution.

5. Discussion

This study develops the MCENN model in the integration of ResNet50 for the diagnosis of AD. Our study demonstrates the importance of random slice sampling to generate possible decisions for improving the classification performance. The MCENN model only requires the minimal processing of structural brain images (rigid transformation, intensity normalization, and brain skull stripping). Moreover, our experiments show that the MCENN model only needs a small number of slices for the AD diagnosis. Furthermore, the computational cost of the MCENN model with one FC layer takes a few seconds per epoch. Our results also suggest that the performance of MCENN model is not sensitive to image resolution.

A recent review (Wen et al., 2020) implemented the existing CNN models, such as CNN on 2D slices, 3D patches/regions of interest (ROIs), or 3D images to overcome the variations due to participant selection, image processing, sample size, or validation procedure across studies. It provided the most comparable classification results across the existing CNN models on the ADNI dataset in the literature. It demonstrated that 3D CNN approaches (3D images, 3D-ROI, 3D-patch) achieved the best performance for the classifier between CN and AD (accuracy: 74%∼88%). Our experiments show that even with a small number of slices, the MCENN model can perform better than most of the existing 2D or 3D CNNs reported in Wen et al. (2020). The performance of the MCENN model is comparable with that of the 3D CNN on the ROI encompassing the hippocampus. This may suggest that introducing prior knowledge of the AD-related image markers into the MCENN model could further improve the MCENN performance.

Our experiments evaluate the robustness of the MCENN model by applying it to the OASIS-3 dataset and a clinical sample while the MCENN model is trained using the ADNI dataset. In general, transfer learning of the MCENN model provides the comparable AD classification in both the OASIS-3 dataset and the clinical sample. Moreover, the CN vs AD classification accuracy of the OASIS-3 dataset obtained by the MCENN is better than that obtained from the 3D-ROI CNN, 3D-patch multi-CNN model, and 3D whole-brain CNN (accuracy: 64%∼67%, Table 6 in Wen et al. (2020)). Similar to the claim in Wen et al. (2020), our study shows that the classification performance is dependent on clinical characteristics. When clinical diagnosis is different from one dataset to another due to practitioner's experience and/or diagnostic tools, transfer learning can play a role of mapping the diagnosis across the two datasets.

Table 7
Robustness of the MCENN with one FC layer in the ADNI and OASIS-3 datasets as well as the clinical sample.

	Accuracy (%)	Sensitivity (%)	Specificity (%)	Geometric mean (%)	Area under curve (%)
ADNI					
CN vs. AD	90.0 ± 0.7	83.5 ± 1.4	96.9 ± 0.5	89.9 ± 0.7	91.4 ± 0.4
CN vs. MCI	68.7 ± 0.6	71.5 ± 3.4	63.3 ± 5.6	67.1 ± 1.5	69.1 ± 2.0
MCI vs. AD	73.5 ± 1.5	69.5 ± 5.8	75.7 ± 5.2	72.4 ± 1.0	77.8 ± 0.2
OASIS-3					
CN vs. AD	82.6 ± 3.3	74.3 ± 6.3	84.4 ± 5.3	79.0 ± 1.3	84.4 ± 1.2
CN vs. MCI	73.6 ± 3.1	54.3 ± 5.6	75.1 ± 3.6	63.7 ± 2.6	67.2 ± 2.3
MCI vs. AD	58.6 ± 4.4	54.1 ± 8.7	70.5 ± 8.0	61.2 ± 2.4	62.9 ± 3.8
Clinical sample					
CN vs. AD	92.2 ± 0.8	84.8 ± 3.0	95.8 ± 0.8	90.1 ± 1.4	96.7 ± 1.3
CN vs. MCI	81.5 ± 2.9	83.8 ± 7.3	79.6 ± 10.2	81.2 ± 2.6	87.4 ± 6.1
MCI vs. AD	74.0 ± 1.5	67.2 ± 5.3	78.1 ± 4.9	72.3 ± 1.3	81.4 ± 3.4

Table 8
Effects of the image resolution on the performance of the MCENN with one FC layer.

	Accuracy (%)	Sensitivity (%)	Specificity (%)	Geometric mean (%)	Area under curve (%)
Resolution of 1 mm × 1 mm × 1 mm, image size of 192 × 192 × 192					
CN vs. AD	92.2 ± 0.8	84.8 ± 3.0	95.8 ± 0.8	90.1 ± 1.4	96.7 ± 1.3
CN vs. MCI	81.5 ± 2.9	83.8 ± 7.3	79.6 ± 10.2	81.2 ± 2.6	87.4 ± 6.1
MCI vs. AD	74.0 ± 1.5	67.2 ± 5.3	78.1 ± 4.9	72.3 ± 1.3	81.4 ± 3.4
Resolution of 2 mm × 2 mm × 2 mm, image size of 96 × 96 × 96					
CN vs. AD	94.5 ± 1.5	91.2 ± 7.8	96.2 ± 2.7	93.5 ± 3.0	95.8 ± 0.5
CN vs. MCI	81.9 ± 5.0	87.1 ± 5.6	77.7 ± 12.9	81.8 ± 4.6	85.3 ± 3.8
MCI vs. AD	74.0 ± 2.2	73.6 ± 13.0	74.3 ± 11.2	72.9 ± 1.2	79.4 ± 5.1
Resolution of 4 mm × 4 mm × 4 mm, image size of 48 × 48 × 48					
CN vs. AD	93.2 ± 2.1	95.2 ± 3.0	92.3 ± 4.4	93.7 ± 1.0	98.0 ± 0.8
CN vs. MCI	76.4 ± 2.1	80.0 ± 8.6	73.5 ± 9.2	76.2 ± 2.0	82.3 ± 4.7
MCI vs. AD	72.5 ± 2.4	65.6 ± 8.2	76.7 ± 6.3	70.6 ± 2.7	75.8 ± 3.9

Even with the image resolution of 4 mm × 4 mm × 4 mm, the MCENN model provides the classification comparable to that obtained from images of 1 mm × 1 mm × 1 mm. This result demonstrates that the MCENN model may not be sensitive to image resolution. It suggests that the MCENN model may have great potential to be adopted with fast image acquisition in the clinic.

This study has some limitations that warrant consideration. The MCENN model achieves good classification results based on the ADNI dataset. Nevertheless, the numbers of MCI and AD patients in the OASIS-3 and clinical sample are small. More experiments with the large number of samples may need for the further investigation of the MCENN robustness. Moreover, incorporating prior knowledge on brain regions related to AD may increase the performance of the MCENN model. But, the cost of incorporating prior knowledge on brain anatomy related to AD requires intensive image processing (e.g., segmentation, image registration), which may not be suitable in clinical setting. Further investigation on incorporating prior knowledge and achieving fast computation is needed.

6. Conclusion

This study proposes a simple and computationally efficient neural network, MCENN, for the computer-aided diagnosis of AD. Our model takes the advantage of the Monte Carlo sampling to generate possible decisions and incorporates them as a final decision. The MCENN model outperforms the existing 2D and 3D CNNs (Wen et al., 2020). The fast computation and insensitivity to the image resolution are the advantages of the MCENN model for its potential in clinic use.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Anqi Qiu reports financial support was provided by National University of Singapore. Anqi Qiu reports a relationship with National University of Singapore that includes: funding grants.

Data availability

The data are publicly available. The code is uploaded at GitHub.

Acknowledgments

This research/project is supported by the Singapore Ministry of Education (Academic research fund Tier 1) and A*STAR (H22P0M0007). This research was also supported by the A*STAR Computational Resource Centre through the use of its high-performance computing facilities.

References

Aderghal, K., Khvostikov, A., Krylov, A., Benois-Pineau, J., Afdel, K., & Catheline, G. (2018). Classification of alzheimer disease on imaging modalities with deep CNNs using cross-modal transfer learning. In *International symposium on computer-based medical systems* (pp. 345–350).

Alzheimer's Association (2015). 2015 Alzheimer's disease facts and figures. *Alzheimer's & Dementia*, *11*, 332–384.

Ansart, M., Epelbaum, S., Bassignana, G., Bône, A., Bottani, S., Cattai, T., et al. (2021). Predicting the progression of mild cognitive impairment using machine learning: A systematic, quantitative and critical review. *Medical Image Analysis*, *67*, http://dx.doi.org/10.1016/j.media.2020.101848.

Bäckström, K., Nazari, M., Gu, I. Y.-H., & Jakola, A. S. (2018). An efficient 3D deep convolutional network for Alzheimer's disease diagnosis using MR images. In *International symposium on biomedical imaging* (pp. 149–153).

Barandela, R., Sánchez, J., García, V., & Rangel, E. (2003). Strategies for learning in class imbalance problems. *Pattern Recognition*, *36*, 849–851.

Basaia, S., Agosta, F., Wagner, L., Canu, E., Magnani, G., Santangelo, R., et al. (2019). Automated classification of alzheimer's disease and mild cognitive impairment using a single MRI and deep neural networks. *NeuroImage : Clinical*, *21*, http://dx.doi.org/10.1016/j.nicl.2018.101645.

Bottani, S., Burgos, N., Maire, A., Wild, A., Ströer, S., Dormont, D., et al. (2022). Automatic quality control of brain T1-weighted magnetic resonance images for a clinical data warehouse. *Medical Image Analysis*, *75*, Article 102219. http://dx.doi.org/10.1016/j.media.2021.102219.

Cheng, D., & Liu, M. (2017). CNNs based multi-modality classification for AD diagnosis. In *International Congress on Image and Signal Processing, BioMedical Engineering and Informatics* (pp. 1–5).

Cui, Z., Gao, Z., Leng, J., Zhang, T., Quan, P., & Zhao, W. (2019). Alzheimer's disease diagnosis using enhanced inception network based on brain magnetic resonance image. (pp. 2324–2330).

Dickerson, B. C., Goncharova, I., Sullivan, M. P., Forchetti, C. M., Wilson, R. S., Bennett, D. A., et al. (2001). MRI-derived entorhinal and hippocampal atrophy in incipient and very mild alzheimer's disease. *Neurobiology of Aging*, *22*, 747–754.

Ewers, M., Sperling, R. A., Klunk, W. E., Weiner, M. W., & Hampel, H. (2011). Neuroimaging markers for the prediction and early diagnosis of Alzheimer's disease dementia. *Trends in Neurosciences*, *34*, 430–442.

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

Ganaie, M., & Tanveer, M. (2022). Ensemble deep random vector functional link network using privileged information for Alzheimer's disease diagnosis. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 1.

Guerdan, L., Sun, P., Rowland, C., Harrison, L., Tang, Z., Wergeles, N., et al. (2019). Deep learning vs. Classical machine learning: A comparison of methods for fluid intelligence prediction. In K. M. Pohl, W. K. Thompson, E. Adeli, M. G. Linguraru (Eds.), *Adolescent brain cognitive development neurocognitive prediction* (pp. 17–25). Cham: Springer International Publishing.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *IEEE conference on computer vision and pattern recognition* (pp. 770–778).

Huang, S.-G., Chung, M. K., & Qiu, A. (2021). Fast mesh data augmentation via Chebyshev polynomial of spectral filtering. *Neural Networks*, *143*, 198–208.

Huang, G., Liu, Z., Maaten, L. V. D., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In *IEEE conference on computer vision and pattern recognition* (pp. 2261–2269).

Jenkinson, M., & Smith, S. M. (2001). A global optimisation method for robust affine registration of brain images. *Medical Image Analysis*, *52*, 143–156.

Jin, D., Zhou, B., Han, Y., Ren, J., Han, T., Liu, B., et al. (2020). Generalizable, reproducible, and neuroscientifically interpretable imaging biomarkers for Alzheimer's disease. *Advanced Science*, *7*, http://dx.doi.org/10.1002/advs.202000675.

Kao, P.-Y., Zhang, A., Goebel, M., Chen, J. W., & Manjunath, B. S. (2019). Predicting fluid intelligence of children using T1-weighted MR images and a StackNet. In K. M. Pohl, W. K. Thompson, E. Adeli, & M. G. Linguraru (Eds.), *Adolescent brain cognitive development neurocognitive prediction* (pp. 9–16). Cham: Springer International Publishing.

Lee, B., Ellahi, W., & Choi, J. Y. (2019). Using deep CNN with data permutation scheme for classification of alzheimer's disease in structural magnetic resonance imaging (sMRI). *IEICE Transactions on Information Systems*, *102-D*, 1384–1395.

Liu, C., Ji, H., & Qiu, A. (2021). Fast vertex-based graph convolutional neural network and its application to brain images. *Neurocomputing*, *434*, 1–10.

Liu, J., Li, M., Luo, Y., Yang, S., Li, W., & Bi, Y. (2021). Alzheimer's disease detection using depthwise separable convolutional neural networks. *Computer Methods and Programs in Biomedicine*, *203*, http://dx.doi.org/10.1016/j.cmpb.2021.106032.

Liu, J., Pan, Y., Li, M., Chen, Z., Tang, L.-L., Lu, C., et al. (2018). Applications of deep learning to MRI images: A survey. *Big Data Mining Analytics*, *1*, 1–18.

López, V., Fernández, A., García, S., Palade, V., & Herrera, F. (2013). An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. *Information Sciences*, *250*, 113–141.

Malik, A. K., Ganaie, M., Tanveer, M., Suganthan, P., & Initiative, A. (2022). Alzheimer's disease diagnosis via intuitionistic fuzzy random vector functional link network. *IEEE Transactions on Computational Social Systems*, 1–12. http://dx.doi.org/10.1109/TCSS.2022.3146974.

Nigri, E., Ziviani, N., Cappabianco, F. A. M., Antunes, A., & Veloso, A. (2020). Explainable deep CNNs for MRI-based diagnosis of Alzheimer's disease. In *International joint conference on neural networks* (pp. 1–8).

Oxtoby, N., Ferreira, F., Mihalik, A., Wu, T., Brudfors, M., Lin, H., et al. (2019). ABCD neurocognitive prediction challenge 2019: Predicting individual residual fluid intelligence scores from cortical grey matter morphology.

Pellegrini, E., Ballerini, L., Hernández, M. V., Chappell, F. M., González-Castro, V., Anblagan, D., et al. (2018). Machine learning of neuroimaging for assisted diagnosis of cognitive impairment and dementia: A systematic review. *Alzheimer's & Dementia : Diagnosis, Assessment & Disease Monitoring*, *10*, 519–535.

Prince, M. J., Wimo, A., Guerchet, M., Ali, G.-C., Wu, Y.-T., & Prina, M. A. (2015). *World Alzheimer report 2015 - the global impact of dementia: An analysis of prevalence, incidence, cost and trends*. Alzheimer's Disease, http://dx.doi.org/10.1111/j.0963-7214.2004.00293.x.

Qiu, S., Chang, G. H., Panagia, M., Gopal, D. M., Au, R., & Kolachalama, V. B. (2018). Fusion of deep learning models of MRI scans, Mini–Mental State Examination, and logical memory test enhances diagnosis of mild cognitive impairment. *Alzheimer's & Dementia : Diagnosis, Assessment & Disease Monitoring*, *10*, 737–749.

Rathore, S., Habes, M., Iftikhar, M. A., Shacklett, A., & Davatzikos, C. (2017). A review on neuroimaging-based classification studies and associated feature extraction methods for Alzheimer's disease and its prodromal stages. *NeuroImage*, *155*, 530–548.

Rizzi, L., Rosset, I., & Roriz-Cruz, M. (2014). Global epidemiology of dementia: Alzheimer's and vascular types. *BioMed Research International*, *2014*, http://dx.doi.org/10.1155/2014/908915.

Salvatore, C., Cerasa, A., Battista, P., Gilardi, M. C., Quattrone, A., & Castiglioni, I. (2015). Magnetic resonance imaging biomarkers for the early diagnosis of Alzheimer's disease: a machine learning approach. *Frontiers in Neuroscience*, *9*, http://dx.doi.org/10.3389/fnins.2015.00307.

Sarraf, S., DeSouza, D. D., Anderson, J. A. E., & Tofighi, G. (2016). Deepad: Alzheimer's disease classification via deep convolutional neural networks using MRI and fMRI. *BioRxiv*, http://dx.doi.org/10.1101/070441.

Ségonne, F., Dale, A. M., Busa, E., Glessner, M., Salat, D. H., Hahn, H. K., et al. (2004). A hybrid approach to the skull stripping problem in MRI. *NeuroImage*, *22*, 1060–1075.

Sled, J. G., Zijdenbos, A. P., & Evans, A. C. (1998). A non-parametric method for automatic correction of intensity non-uniformity in MRI data. *IEEE Transactions on Medical Imaging*, *17*, 87–97.

Tamez-Pena, J., Orozco, J., Sosa, P., Valdes, A., & Nezhadmoghadam, F. (2019). Ensemble of SVM, random-forest and the BSWiMS method to predict and describe structural associations with fluid intelligence scores from T1-weighed MRI. In K. M. Pohl, W. K. Thompson, E. Adeli, & M. G. Linguraru (Eds.), *Adolescent brain cognitive development neurocognitive prediction* (pp. 47–56). Cham: Springer International Publishing.

Tanveer, M., Richhariya, B., Khan, R. U., Rashid, A. H., Khanna, P., Prasad, M., et al. (2020). Machine learning techniques for the diagnosis of Alzheimer's disease: A Review, 16, 35.

Thong, J., Du, J., Ratnarajah, N., Dong, Y., Soon, H., Saini, M., et al. (2014). Abnormalities of cortical thickness, subcortical shapes, and white matter integrity in subcortical vascular cognitive impairment. *Human Brain Mapping*, *35*, 2320–2332.

Thong, J., Hilal, S., Wang, Y., Soon, H., Dong, Y., Collinson, S., et al. (2013). Association of silent lacunar infarct with brain atrophy and cognitive impairment. *Journal of Neurology, Neurosurgery and Psychiatry*, *84*, 1219–1225.

Valliani, A. A., & Soni, A. (2017). Deep residual nets for improved alzheimer's diagnosis. In *Proceedings of the 8th ACM international conference on bioinformatics, computational biology,and health informatics*. http://dx.doi.org/10.1145/3107411.3108224.

Vang, Y. S., Cao, Y., & Xie, X. (2019). A combined deep learning-gradient boosting machine framework for fluid intelligence prediction. In K. M. Pohl, W. K. Thompson, E. Adeli, & M. G. Linguraru (Eds.), *Adolescent brain cognitive development neurocognitive prediction* (pp. 1–8). Cham: Springer International Publishing.

Wee, C.-Y., Liu, C., Lee, A., Poh, J. S., Ji, H., & Qiu, A. (2019). Cortical graph neural network for AD and MCI diagnosis and transfer learning across populations. *NeuroImage: Clinical*, *23*, http://dx.doi.org/10.1016/j.nicl.2019.101929.

Wen, J., Thibeau-Sutre, E., Samper-González, J., Routier, A., Bottani, S., Durrleman, S., et al. (2020). Convolutional neural networks for classification of Alzheimer's Disease: overview and reproducible evaluation. *Medical Image Analysis*, *63*, http://dx.doi.org/10.1016/j.media.2020.101694.

Wimo, A., Guerchet, M., Ali, G.-C., Wu, Y.-T., Prina, A. M., Winblad, B., et al. (2017). The worldwide costs of dementia 2015 and comparisons with 2010. *Alzheimer's & Dementia*, *13*, 1–7.