NEUROSCIENCE

RESEARCH ARTICLE

F. Zhang et al. / Neuroscience 491 (2022) 200–214

A Single Model Deep Learning Approach for Alzheimer's Disease Diagnosis

Fan Zhang, [a,b] **Bo Pan,** [d] **Pengfei Shao,** [a,b] **Peng Liu,** [c]**, Alzheimer's Disease Neuroimaging Initiative** [†]**the Australian Imaging Biomarkers Lifestyle flagship study of ageing**[,‡]**Shuwei Shen,** [c*] **Peng Yao** [b*] **and Ronald X. Xu** [a,c*]

[a] *Department of Precision Machinery and Precision Instrumentation, University of Science and Technology of China, Hefei, Anhui 230026, China*

[b] *Key Laboratory of Precision Scientific Instrumentation of Anhui Higher Education Institutes, University of Science and Technology of China, Hefei, Anhui 230026, China*

[d] *First Affiliated Hospital, University of Science and Technology of China, Hefei, Anhui 230031, China*

[c] *Suzhou Advanced Research Institute, University of Science and Technology of China, Suzhou, Jiangsu 215000, China*

Abstract—Early and accurate diagnosis of Alzheimer's disease (AD) and its prodromal period mild cognitive impairment (MCI) is essential for the delayed disease progression and the improved quality of patients' life. The emerging computer-aided diagnostic methods that combine deep learning with structural magnetic resonance imaging (sMRI) have achieved encouraging results, but some of them are limit of issues such as data leakage, overfitting, and unexplainable diagnosis. In this research, we propose a novel end-to-end deep learning approach for automated diagnosis of AD. This approach has the following differences from the current approaches: (1) Convolutional Neural Network (CNN) models of different structures and capacities are evaluated systemically and the most suitable model is adopted for AD diagnosis; (2) A data augmentation strategy named Two-stage Random RandAugment (TRRA) is proposed to alleviate the overfitting issue caused by limited training data and to improve the classification performance in AD diagnosis; (3) An explainable method of Grad-CAM++ is introduced to generate the visually explainable heatmaps to make our model more transparent. Our approach has been evaluated on two publicly accessible datasets for two classification tasks of AD vs. cognitively normal (CN) and progressive MCI (pMCI) vs. stable MCI (sMCI). The experimental results indicate that our approach outperforms the state-of-the-art approaches, including those using multi-model and three-dimensional (3D) CNN methods. The resultant heatmaps from our approach also highlight the lateral ventricle and some regions of cortex, which have been proved to be affected by AD. © **2022 IBRO. Published by Elsevier Ltd. All rights reserved.**

Key words: Alzheimer's disease diagnosis, mild cognitive impairment, data augmentation, explainable deep learning model.

INTRODUCTION

Alzheimer's disease (AD) is the most common type of dementia (Tiwari et al., 2019). It is estimated that 131 million people worldwide will suffer from AD and other dementias by 2050, presenting a great healthcare challenge in the 21st century (Livingston et al., 2017). Mild cognitive impairment (MCI) represents a slight decline of mental ability along the continuum from normal cognition to AD, while over 33% of MCI subjects will progress to AD within five or more years (Ward et al., 2013; Livingston et al., 2017). Currently, there is no curative treatment for AD. However, the progression of the disease can be slowed down through medications, exercise and memory training (Anonymous, 2020). In this regard, early detection of AD and accurate diagnosis of MCI are critical for delaying the disease progress and improving the patient's quality of life (Liu et al., 2020). Structural magnetic resonance imaging (sMRI) has been increasingly used for clinical diagnosis of AD and MCI because

[*] Corresponding authors. Address: Department of Precision Machinery and Precision Instrumentation, University of Science and Technology of China, Hefei, Anhui 230026, China (R.X. Xu).

E-mail addresses: swshen@ustc.edu.cn (S. Shen), yaopeng@ustc.edu.cn (P. Yao), xux@ustc.edu.cn (R. X. Xu).

[†] Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (http://www.loni.ucla.edu/ADNI). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. ADNI investigators include (complete listing available at http://adni.loni.ucla.edu/wpcontent/uploads/how_to_apply/ADNI_Authorship_List.pdf).

[‡] Data used in the preparation of this article was obtained from the Australian Imaging Biomarkers and Lifestyle flagship study of ageing (AIBL) funded by the Commonwealth Scientific and Industrial Research Organisation (CSIRO) which was made available at the ADNI database (www.loni.usc.edu/ADNI). The AIBL researchers contributed data but did not participate in analysis or writing of this report. AIBL researchers are listed at www.aibl.csiro.au.

Abbreviations: AD, Alzheimer's disease; BIDS, Brain Imaging Data Structure; CN, cognitively normal; CNN, Convolutional Neural Network; MCI, mild cognitive impairment; PET, positron emission tomography; pMCI, progressive mild cognitive impairment; sMCI, stable mild cognitive impairment; sMRI, structural magnetic resonance imaging; TRRA, Two-stage Random RandAugment.

https://doi.org/10.1016/j.neuroscience.2022.03.026

0306-4522/© 2022 IBRO. Published by Elsevier Ltd. All rights reserved.

it can help differentiate neuropathological alterations associated with these diseases (Serrano-Pozo et al., 2011), and it does not involve ionizing radiation and is cheaper compared with positron emission tomography (PET) (Spasov et al., 2019).

In recent years, many researchers have developed computer-aided diagnostic systems by combining machine learning methods and sMRI data to identify the progression of AD (Beheshti and Demirel, 2015; Christian et al., 2015; Liu et al., 2015; Moller et al., 2016; Rathore et al., 2017; Cao et al., 2020; Kang et al., 2021; Prakash et al., 2021). Herein, the primary research tasks include the classification of AD versus cognitively normal (CN) (Wen et al., 2020) and the prediction of conversion from MCI toward AD (stable MCI (sMCI) versus progressive MCI (pMCI)) (Anonymous, 2020). In these studies, the predefined features are first obtained from image preprocessing procedures, and then different types of classifiers are applied for classification tasks (Beheshti and Demirel, 2015; Liu et al., 2015; Moller et al., 2016). Since the feature selection and the classification algorithms are executed independently in traditional machine learning methods (LeCun et al., 2015), this may lead to the potential loss of information associated with the classification tasks (Nguyen and de la Torre, 2010).

Deep learning is a state-of-the-art machine learning technique capable of extracting low-to-high level feature representations automatically from large and high-dimensional data sets, superior to the traditional machine learning methods (Jo et al., 2019). As one of the most popular deep learning architectures, Convolutional Neural Network (CNN) has recently been explored for AD diagnosis (Farooq et al., 2017; Vu et al., 2018; Wang et al., 2019; Lian et al., 2020; Liu et al., 2020). Lian et al. proposed a hierarchical fully convolutional network to construct the hierarchical classifier for AD diagnosis (Lian et al., 2020). Liu et al. proposed a multi-model deep learning method for hippocampal segmentation and AD diagnosis (Liu et al., 2020). Despite these encouraging results, the credibility of some studies in CNN-assisted AD diagnosis is hindered by data leakage issues (Wen et al., 2020). Wen et al. analyzed the reasons that cause data leakage and pointed that a subject simultaneously appearing in training, validation and test sets may virtually increase the performance of the CNN models (Wen et al., 2020). Backstrom et al. also verified that the diagnostic accuracy of the unbiased splitting (at the subject level) is 8% lower than that of the biased splitting (at the slice level) (Backstrom et al., 2018). Two-dimensional (2D) CNN models, such as DenseNet (Huang et al., 2017) and EfficientNet (Mingxing and Quoc, 2019), have been successfully implemented in natural image classification and are also explored in AD diagnosis (Wen et al., 2020). 2D models pre-trained on ImageNet (Deng et al., 2009) are readily applicable to small-scale medical image datasets by transfer learning to achieve better performance (Liu et al., 2021). In addition, many slices can be extracted from a single 3D image to increase the amount of training data in 2D models (Wen et al., 2020). However, it was also reported that the AD classification accuracy for 2D CNN models is 10% lower than that of three-dimensional (3D) CNN models (Wen et al., 2020). We will focus on 2D CNN models with the hypothesis that they will yield the classification performance comparable to a 3D model after algorithm optimization.

This research aims at addressing several unsolved problems associated with CNN-assisted AD diagnosis. First of all, there is no systematic comparison of the classification performance for different CNN models in AD diagnosis. For example, Wen et al. (Wen et al., 2020) and Valliani et al. (Valliani and Soni, 2017) both used ResNet-18 in their studies but discarded other ResNet models (He et al., 2016). Second, automated augmentation strategies have not been introduced in CNN-assisted AD diagnosis despite their demonstrated effectiveness in alleviating the overfitting issue caused by limited training data. Finally, many CNN models for AD diagnosis cannot provide the explanations of their predictions due to the "black box" nature of deep learning.

When performing classification tasks on large-scale image datasets, ameliorating model structure from initial AlexNet (Krizhevsky et al., 2017) to EfficientNet (Mingxing and Quoc, 2019) or increasing the capacity of the similar model structures can always achieve better performance (He et al., 2016; Mingxing and Quoc, 2019). However, this is not always correct on small-scale image datasets because the increased capacity may cause the model to transition from an under-fitting area to an over-fitting area (Belkin et al., 2019). Considering that even Alzheimer's Disease Neuroimaging Initiative (ADNI), one of the largest public datasets for AD diagnosis, has limited amount of data, the first question we focus on is: which model structure yields the best performance and what capacity of models in similar structures is most suitable for AD diagnosis? In this research, we try to identify the most suitable model by assessing the performance of CNNs with different structures and capacities.

At the same time, we need to further alleviate the overfitting issue caused by the limited amount of data. Data augmentation is one of the effective methods to alleviate the overfitting issue and finally improve the generalization of models. Since the conventional data augmentation strategies are problem specific, it is difficult to extend the same strategies to different applications and fields. Automated augmentation strategies are expected to overcome this shortcoming (Cubuk et al., 2019; Lim et al., 2019; Cubuk et al., 2020) and various automated augmentation strategies, such as AutoAugment (Cubuk et al., 2019) and RandAugment (Cubuk et al., 2020), have proven their effectiveness in alleviating overfitting and improving model robustness for natural image classification. Considering the difference between natural image datasets and sMRI datasets, direct use of data augmentation strategies developed for the former may not be the best choice. In this research, we propose a Two-stage Random RandAugment (TRRA) for improved classification performance in AD diagnosis.

Recently, visual explanations of CNN models on large-scale image dataset for enhanced transparency has attracted more and more research attention. Gradient-weighted Class Activation Mapping (Grad-CAM) introduces the gradients of the predicted target with respect to the final convolutional layer to generate a heatmap highlighting the areas that are important to the predicted target in the image (Selvaraju et al., 2017). Fan et al. introduced 3D Grad-CAM to their approach and found that their model focused on the ventricles, hippocampus, and some regions of cortex when classifying AD and NC (Fan et al., 2021). As an improved version of Grad-CAM, Grad-CAM++ generates better visual explanations of model predictions to improve the model transparency (Chattopadhay et al., 2018).

In this paper, we propose a novel end-to-end deep learning approach for automated diagnosis of AD from the sMRI data. The main contributions of this research are summarized as follows:

(1) CNN models of different structures and capacities are evaluated systemically, and the experimental results indicate that models in advanced structure with moderate capacity rather than the largest one can achieve better performance. To the best of our knowledge, this is the first report of using EfficientNet for AD diagnosis.
(2) A TRRA data augmentation strategy is proposed to alleviate the overfitting issue caused by limited training data and to improve the classification performance in AD diagnosis.
(3) An explainable method of Grad-CAM++ is introduced to generate the visually explainable heatmaps to make our model more transparent.

EXPERIMENTAL PROCEDURES

Participants and data preprocessing

Data used in this research were obtained from Alzheimer's Disease Neuroimaging Initiative (ADNI) database (https://adni.loni.usc.edu/) and Australian Imaging Biomarkers and Lifestyle flagship study of ageing (AIBL) database (https://aibl.csiro.au/). ADNI dataset is one of the largest publicly accessible datasets used for AD diagnosis and has been widely used in scientific research. AIBL dataset has the similar inclusion criteria and image acquisition procedures with ADNI dataset and is commonly used to further evaluate the generalization ability of the models. The ADNI was launched in 2003 as a public–private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial MRI, PET, other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of MCI and early AD. For up-to-date information, see www.adni-info.org. Data in AIBL database was collected by the AIBL study group. AIBL study methodology has been reported previously (Ellis et al., 2009). Informed consent was acquired from all participants, and the ethics committee of the leading institution of each dataset approved their research. Baseline images in two datasets are used in this study, and images in ADNI dataset are from four phases (ADNI-1, GO, 2 and 3). The MCI subjects in ADNI dataset are specified as sMCI subjects that are diagnosed as MCI at all available time points over 36 months, or pMCI subjects that convert to AD within 36 months after the baseline time. The 36-month conversion time is consistent with the time in the literature (Liu et al., 2017; Lian et al., 2020; Wen et al., 2020).

Considering that CNN can extract low-to-high level features automatically, in order to provide fair evaluation results, we use the "minimal" preprocessing procedure suggested by Wen et al. (Wen et al., 2020). First, all the data are converted into the Brain Imaging Data Structure (BIDS) format (Gorgolewski et al., 2016). Second, the N4ITK method is used for the bias field correction (Tustison et al., 2010). Third, the SyN algorithm (Avants et al., 2008) from ANTs (Avants et al., 2014) is used for affine registration that aligns each image to the MNI space with the ICBM 2009c nonlinear symmetric template (Fonov et al., 2009; Fonov et al., 2011). Finally, the registered images are cropped to remove the background, resulting in the images of size 169 × 208 × 179, with 1 mm isotropic voxels. For each subject, we obtain 129 slices of RGB images by discarding the first twenty and last twenty slices along the sagittal direction and copying each of the remaining slices to the R, G, and B channels. All the preprocessing procedures are performed using the Clinica (Routier et al., 2018; Samper-Gonzalez et al., 2018; Wen et al., 2020) and the ANTs (Avants et al., 2010; Avants et al., 2011) software packages. Some subjects are excluded by the preprocessing procedures for the following reasons: AD and CN subjects whose label change over time; MCI subjects who have two or more label changes (for example, progressing to AD and then reverting back to MCI); MCI subjects who do not convert to AD and are followed for less than 36 months; Subjects who do not pass quality check (Fonov et al., 2018). Table 1 and Table 2 summarize the demographics, the mini-mental state examination (MMSE) scores, and the global clinical dementia rating (CDR) scores of the ADNI and AIBL participants.

Overview of the proposed deep learning approach

Fig. 1 shows the flowchart of our proposed approach that includes the sequential stages of training, validation/testing and visual explanation. The pre-processed images are firstly resized from 208 × 179 to 297 × 256 in all the stages. During the training stage, the TRRA data augmentation strategy is applied to each image in the training set and the resultant image is randomly cropped to match the size of 224 × 224 required by the CNN models. For the AD classification task, we use the model pre-trained on the ImageNet dataset and fine-tune it on the ADNI training set. For the MCI conversion prediction task, we also investigate the possibility of transferring a CNN model pre-trained on AD classification task to this task. For each classification task, the model generates two prediction outputs per image and the cross-entropy loss function expressed as Equation1 is adopted:

$$loss(x, class) = -\log\left(\frac{\exp\left(x[class]\right)}{\sum_j \exp\left(x[j]\right)}\right) \tag{1}$$

where $class \in \{0, 1\}$ specifies the ground-truth class and x is the values predicted by the model. No data

Table 1. Summary of participant demographics, MMSE and CDR scores at baseline for ADNI

	Subjects	Age	Gender	MMSE	CDR
AD	333	75.0 ± 7.8 [55.1, 90.9]	150 F/183 M	23.2 ± 2.1 [18, 27]	0.5: 156; 1: 176; 2: 1
CN	338	74.4 ± 5.7 [59.8, 89.6]	174 F/164 M	29.1 ± 1.1 [24, 30]	0: 338
sMCI	296	72.2 ± 7.44 [55.0, 88.6]	119 F/177 M	28.0 ± 1.7 [23, 30]	0.5: 296
pMCI	302	74.3 ± 7.1 [55.2, 91.7]	123 F/179 M	26.8 ± 1.9 [19, 30]	0.5: 300; 1: 2

Values are presented as Means ± S.D. [range]. M: male, F: female.

Table 2. Summary of participant demographics, MMSE and CDR scores at baseline for AIBL

	Subjects	Age	Gender	MMSE	CDR
AD	77	75.0 ± 7.7 [55.5, 93.4]	43F / 34 M	20.6 ± 5.3 [6, 29]	0.5: 29; 1: 40; 2: 6; 3: 2
CN	450	73.1 ± 6.2 [60.3, 92.1]	263F / 187 M	28.8 ± 1.2 [25, 30]	0: 425; 0.5: 25

Values are presented as Means ± S.D. [range]. M: male, F: female.

augmentation strategy is applied during the validation/test stage, and the input image is only center cropped to ensure the repeatability of each test. For each subject, soft voting is used to generate the subject-level decision (Raschka, 2015). First, SoftMax normalization is carried out on the output of all slices from the same patient to obtain the predicted probability p. Second, the number of correct predictions for the j-th slice of all subjects on the validation set is divided by all the number of correct predictions for 129 slices in order to obtain the weight of the j-th slice w_j. Finally, the subject-level decision is made based on the following formula:

$$\hat{y} = \arg\max_i \sum_{j=1}^{129} w_j p_{ij} \tag{2}$$

where $\hat{y}$ is the class of a subject that is finally predicted and $i \in \{0,1\}$ contains all the possible classes. For the AD classification task, the subject will be predicted as AD (CN) if $\hat{y}=\text{i} = 1(0)$. For MCI conversion prediction task, the subject will be predicted as pMCI (sMCI) if $\hat{y}=\text{i} = 1(0)$. The weight w_j reflects the importance of each slice and w_j calculated on the validation set will be retained and used when evaluating on the test set (Wen et al., 2020).

For the visual explanation stage, the gradient weights α^{ij}_{kc} for the predicted class c and the feature map A^k is firstly calculated using the following formula:

$$\alpha^{ij}_{kc} = \frac{\frac{\partial^2 Y^c}{(\partial A^k_{ij})^2}}{2\frac{\partial^2 Y^c}{(\partial A^k_{ij})^2} + \sum_a \sum_b A^k_{ab} \left\{ \frac{\partial^3 Y^c}{(\partial A^k_{ij})^3} \right\}} \tag{3}$$

where Y^c is the predicted class score, and A^k is the k-th feature map of the last convolutional layer. (i,j) and (a,b) are the position of the feature mapA^k. Then, the gradient of Y^c with respect to the position (i,j) of the feature map A^k is calculated. Then, the weights w^c_k is calculated as:

$$w^c_k = \sum_i \sum_j \alpha^{ij}_{kc} \cdot relu\left(\frac{\partial Y^c}{\partial A^k_{ij}}\right) \tag{4}$$

where *relu* function is used to get positive gradients. Finally, the visually explainable heatmap is generated by combining the weights w^c_k and all K feature maps:

$$L^c_{ij} = relu\left(\sum_k w^c_k \cdot A^k_{ij}\right) \tag{5}$$

Convolutional Neural Network (CNN) models

To systemically evaluate different CNN models, five CNN structures from classic VGG series (Simonyan and Zisserman, 2014) to the latest EfficientNet series (Mingxing and Quoc, 2019) are adopted in this research and their detailed information is list in Table3. For all the models, the last fully connected (FC) layer is replaced with a new FC layer with 2 output nodes.

Compared with the conventional convolution, the depth-wise separable convolution used in EfficientNet (Mingxing and Quoc, 2019) can reduce the number of parameters and reduce the issue of overfitting. The main building block used in EfficientNet named mobile inverted bottleneck (Sandler et al., 2018; Tan et al., 2019) is shown in Fig. 2.

Data augmentation strategy

Inspired by RA, we propose a novel automated data augmentation strategy called Two-stage Random RandAugment (TRRA). TRRA consists of 23 available transformations and all available transformations and corresponding range of magnitude are listed in Table 4 (7 newly added transformations compared with RandAugment are bolded). The 23 transformations are further divided into two categories of [*color*] and[*shape*]. The magnitude M for all the transformations is an integer randomly sampled between the preset two values. TRRA contains three interpretable integer hyperparameters N_{color}, N_{shape} and P. $N_{color(shape)}$ is used to control the number of transformations that are randomly selected from the color (shape) category and sequentially applied to the training image. The probability parameter P is used to control whether the selected transformation should be executed or not so

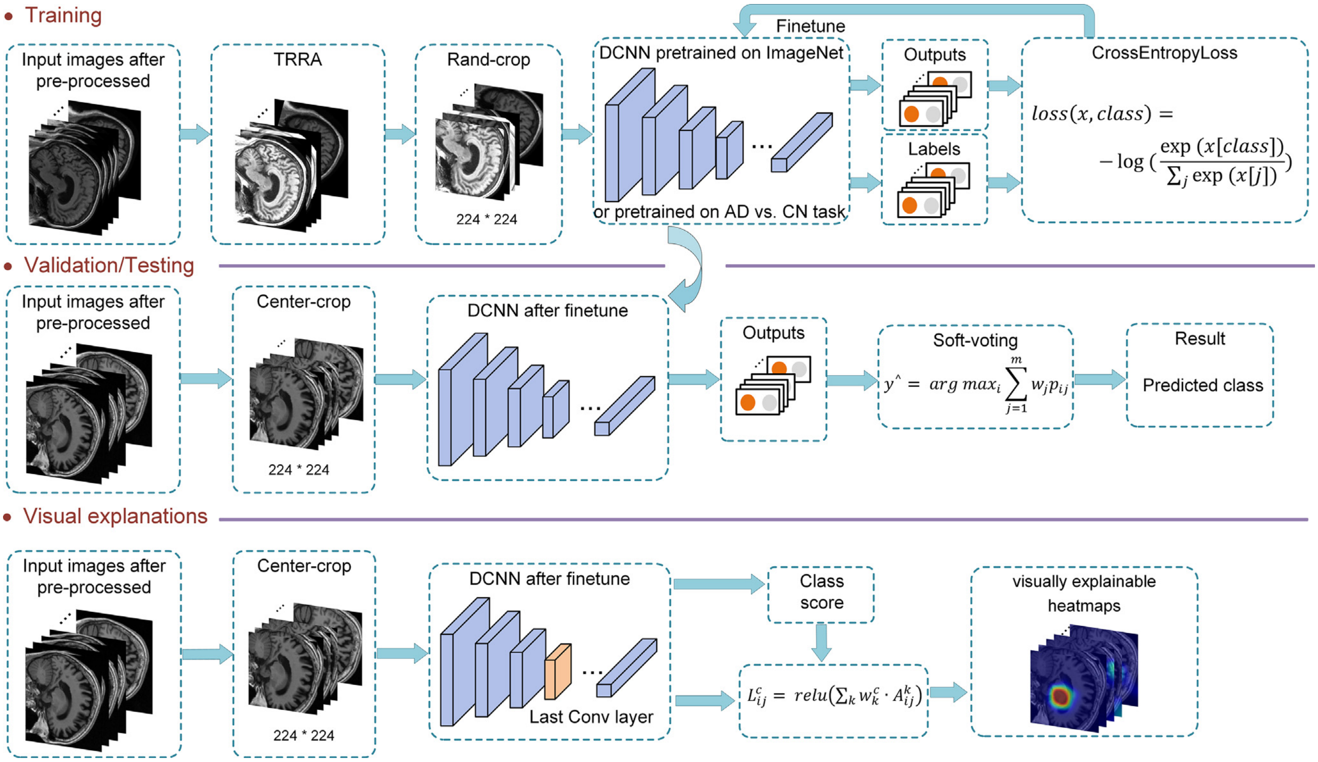
• Training
Input images after pre-processed
TRRA
Rand-crop
224 * 224
Finetune
DCNN pretrained on ImageNet
or pretrained on AD vs. CN task
Outputs
Labels
CrossEntropyLoss
$loss(x, class) = -\log\left(\frac{\exp(x[class])}{\sum_j \exp(x[j])}\right)$
• Validation/Testing
Input images after pre-processed
Center-crop
224 * 224
DCNN after finetune
Outputs
Soft-voting
$\hat{y} = arg\ max_i \sum_{j=1}^{m} w_j p_{ij}$
Result
Predicted class
• Visual explanations
Input images after pre-processed
Center-crop
224 * 224
DCNN after finetune
Last Conv layer
Class score
$L_{ij}^{c} = relu(\sum_k w_k^c \cdot A_{ij}^k)$
visually explainable heatmaps

Fig. 1. The flowchart of the proposed deep learning approach.

that each transformation has the probability of $1-P$ to remain the input image unchanged.

When performing data augmentation, TRRA first selects N_{color} transformations in the color category and apply them to the image according to the preset magnitude M. Each transformation has the probability of $1-P$ not to be superimposed on the input image. Then, TRRA select N_{shape} transformations in the shape category and apply them to the image according to the preset magnitude M. Each transformation has the probability of $1-P$ not to be superimposed on the input image. The workflow of TRRA when N_{color} and N_{shape} are both equal to 1 is shown in Fig. 3. The input image is processed by TRRA to generate an augmented image.

The rationale of TRRA design lies on the following three aspects. First, we think that adding 7 kinds of transformations and setting M randomly sampling between two values can further increase the diversity and quantity of training data. Second, we believe that color attributes related transformations and shape attributes related transformations contribute differently to the classification performance. In RA, each operation is randomly selected from all the transformations without differentiating categories. Therefore, it is likely that most of the operations are selected from the category with relatively small contributions in the case of $N > 1$. So, we use two hyperparameters N_{color} and N_{shape} to explicitly specify the number of transformations selected from the two categories. Finally, we believe that superimposing too many transformations on the input image will destroy its inherent characteristics despite the increased diversity of training data by data augmentation. The probability parameter P and $N_{color(shape)}$ can limit the data augmentation process to a suitable range. So, we introduce third hyperparameter P to control the probability of execution of each operation.

The following ablation experiments are designed to verify the contribution of each improvement of TRRA to classification performance.

(1) To investigate the contribution of 7 newly added transformations: We expand the search space of RA by adding the 7 kinds of transformations so we can get RandAugment-23 (RA-23). RA-23 uses a fixed magnitude M as RA.
(2) To investigate the contribution of a random M: We change the magnitude M of RA-23 from a fixed value to an integer randomly sampled between $[5, X]$ $(X \in [10, 30])$ to get Random-RandAugment-23 (RRA-23).
(3) To investigate the contribution of dividing all transformations into two categories of [*color*] and [*shape*], we set the probability parameter P in TRRA to 1, and then compare TRRA with RRA-23.
(4) To investigate the contribution of the probability parameter P, we compare the performance of TRRA under different probability parameter P.

For RA, RA-23 and RRA-23, we perform a grid search to get their optimal performance. Specifically, hyperparameter N is sampled from 1 to 8 in a step size

Table 3. Detailed information of CNNs with different structures and different parameters

Model	Params (M)	FLOPs (B)	Model	Params (M)	FLOPs (B)	Model	Params (M)	FLOPs (B)
VGG-11	132.9	7.6	SE-ResNet-50	28.1	3.9	EfficientNet-B1	7.8	0.7
VGG-13	133.1	11.3	SE-ResNet-101	49.3	7.6	EfficientNet-B2	9.1	1.0
VGG-16	138.4	15.5	SE-ResNet-152	66.8	11.4	EfficientNet-B3	12.2	1.8
VGG-19	143.7	19.7	SENet-154	115.1	20.8	EfficientNet-B4	19.3	4.2
ResNet-18	11.7	1.8	DenseNet-121	8.0	2.9	EfficientNet-B5	30.4	9.9
ResNet-34	21.8	3.7	DenseNet-169	14.2	3.4	EfficientNet-B6	43.0	19
ResNet-50	25.6	4.1	DenseNet-201	20.0	4.4	EfficientNet-B7	66.4	37
ResNet-101	44.6	7.9	DenseNet-161	28.7	7.8			
ResNet-152	60.2	11.6	EfficientNet-B0	5.3	0.39			

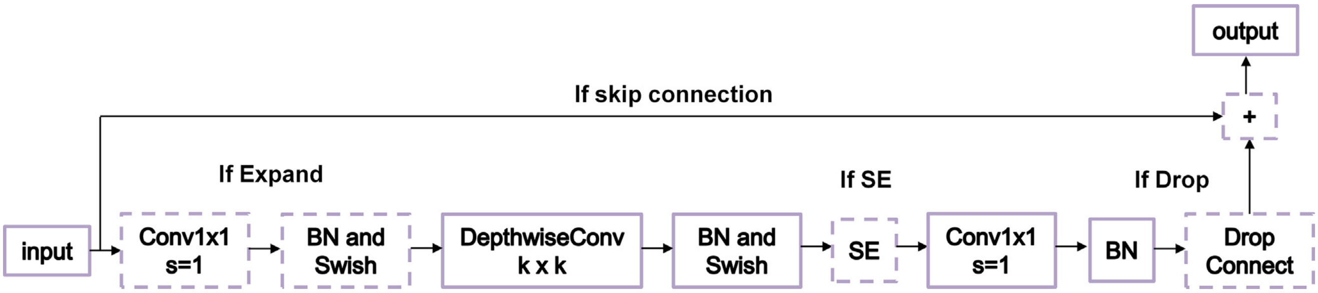

Fig. 2. Schematic diagram of mobile inverted bottleneck used in EfficientNet.

of 1 for each strategy. Hyperparameter M is sampled from 5 to 30 in a step size of 5 for RA and RA-23. For RRA-23, M is an integer value randomly sampled between $[5, X]$, and X is sampled from 10 to 30 in a step size of 5.

Evaluation metrics

The following commonly used metrics are chosen to evaluate the classification performance for AD diagnosis (Lian et al., 2020): accuracy, sensitivity, specificity, and area under the receiver operating characteristic curve (AUC), where accuracy is used as the main evaluation metric.

Implementation

The performances of the proposed approach are evaluated using two binary tasks of AD classification (AD vs. CN) and MCI conversion prediction (sMCI vs. pMCI). The AD classification task is used as a baseline for evaluating the performance of different models and data augmentation strategies, and the best model is used for the MCI conversion prediction task.

To avoid data leakage, we adopt a previously reported method (Wen et al., 2020) to split the ADNI dataset and carefully check the results. Specifically, the ADNI dataset is split into the training/validation/test sets at the subject-level. The training and the validation sets are used for the selection of the model capacity of the five structures and the grid search of the hyperparameters of four data augmentation strategies. The test set only tests the best-performing model of each structure and the best hyperparameter combination of each data augmentation strategy. We ensure that age and sex distributions between training, validation and test sets are not significantly different. To avoid the influence caused by a single split, we carry out a total of three splits following the same ratio of training/validation/test sets (6:2:2) as Backstrom et al (Backstrom et al., 2018). All experiments are performed using these three splits so that the mean and standard deviation of the metrics can be obtained.

All the training and the testing tasks are performed on 2 NVIDIA GeForce GTX 2080Ti graphics cards using Pytorch. To prevent overfitting, we adopt an early stopping strategy: when the validation accuracy doesn't improve for a continuous 20 epochs, the training process will stop, otherwise, the training will continue to the end of the predefined periods. The selected model is the one which obtain the highest validation accuracy during training. Batch size for model training in this study is 128, but it is reduced for some of the large models to match the memory capacity of the graphic cards.

RESULTS

Comparison study of different CNN models

In this part of experiments, we first compare different CNN models on ADNI validation set to determine the best-performing model of each structure, and then test them on ADNI test set. The detailed experimental results of different CNN models on ADNI validation set can be found in Appendix A.

The AD classification performance on ADNI test set of CNN models in different structures is presented in Table 5. ResNet-18 is also selected for comparing with the results in the literature (Valliani and Soni, 2017;

Table 4. List of all transformations can be selected during the search using TRRA

	Operation Name	Description	Range of magnitude
[*color*]	Auto Contrast	Maximize (normalize) image contrast.	–
	Equalize	Equalize the image histogram.	–
	Invert	Invert (negate) the image.	–
	Posterize	Reduce the number of bits for each color channel.	[0, 4]
	Solarize	Invert all pixel values above a threshold.	[0, 256]
	Solarize Add	Add a value to the image and do solarize.	[0, 100]
	Color	Adjust image color balance.	[0.1, 1.9]
	Contrast	Adjust image contrast.	[0.1, 1.9]
	Brightness	Adjust image brightness.	[0.1, 1.9]
	Sharpness	Adjust image sharpness.	[0.1, 1.9]
	Random noise	Add a noise randomly sampled from a uniform distribution.	[0, 0.4]
	Gaussian noise	Add a noise randomly sampled from the Gaussian distribution.	[0, 0.4]
	Gaussian blur	Gaussian blur filter.	[0, 2.0]
[*shape*]	**Horizontal flip**	Flip the image Horizontally (left to right).	–
	Vertical flip	Flip the image vertically (top to bottom).	–
	Rotate	Rotate the image according to magnitude.	[0, 30]
	Shear X	Shear the image along the horizontal axis.	[0, 0.3]
	Shear Y	Shear the image along the vertical axis.	[0, 0.3]
	Cutout	Set a random square patch with a side length of magnitude, pixels inside turn gray.	[0, 40]
	Translate X	Move the image along the horizontal axis.	[0, 100]
	Translate Y	Move the image along the vertical axis.	[0, 100]
	Scale	Scale the image horizontally and vertically with equal magnitude degrees.	[0.9, 1.4]
	Scale XY	Scale the image horizontally and vertically with different magnitude degrees.	[0.9, 1.4]

Wen et al., 2020). As the data in Table 5 show: (1) Accuracy of ResNet-18 without applying data augmentation is 0.774. This is very similar to Wen et al. (0.760) (Wen et al., 2020) and Valliani et al. (0.788) (Valliani and Soni, 2017), which indicates no data leakage in our evaluation. (2) The models of different CNN structures in Table 5 are all in the moderate capacity rather than the maximum capacity, which indicates that the models with moderate capacity instead of maximum capacity achieve the best performance. (3) The classification performance of each model applying TRRA show similar significant improvement. The general improvement of more than 10% in them indicate that the performance of the model trained with the proposed data augmentation strategy is better than that of the model trained with unenhanced data in AD classification task. 4) EfficientNet-B1 and DenseNet-169 both achieve the highest accuracy (0.932) on the ADNI test set. Combining the above observations and data in Appendix A, we can see that more advanced model structures can achieve better performance, and models in similar structure with moderate capacity rather than the largest one can achieve better performance. Considering that EfficientNet-B1 has the highest accuracy on both ADNI validation set and ADNI test set, it is used in the following experiments.

Comparison study of different data augmentation strategies

In this part of experiments, we first perform a grid search on ADNI validation set to determine the optimal hyperparameter combination of each data augmentation strategy, and then test them on ADNI test set. The detailed experimental results of EfficientNet-B1 with different data augmentation strategies on ADNI validation set can be found in Appendix B.

The AD classification performance on ADNI test set of each data augmentation strategy is presented in Fig. 4. Observations from Fig. 4 show that: (1) RA-23 performs better than RA, which indicates that adding 7 kinds of transformations in the search space helps to improve classification performance. (2) RRA-23 performs better than RA-23, which indicates that compared with the fixed magnitude, a magnitude randomly sampled between two values helps to improve classification performance.

As shown in Fig. 4, RRA-23 helps to get the best accuracy of 0.917 on ADNI test set when the hyperparameters N is 7 and M is randomly sampled from [5, 30]. So, we set the sum of N_{color} and N_{shape} as 7, and M randomly sample between [5, 30] in searching the optimal hyperparameters for TRRA. The detailed experimental results of EfficientNet-B1 with TRRA under different hyperparameters on ADNI validation set refer to Appendix C. From the results we observe that randomly selecting five transformations from the [*color*] category and randomly selecting two transformations from the [*shape*] category achieve the best performance, which indicates that [*color*] category contributes more to classification performance than [*shape*] category.

The AD classification performance on ADNI test set of TRRA is presented in Fig. 5. Observations from Fig. 4 and Fig. 5 show that: (1) TRRA performs better than RRA-23. The accuracy of TRRA is 0.930 when P is equal to 1 and is further improved compared with 0.917 achieved by RRA-23, which indicates dividing all transformations into two categories of [*color*] and [*shape*] is better for

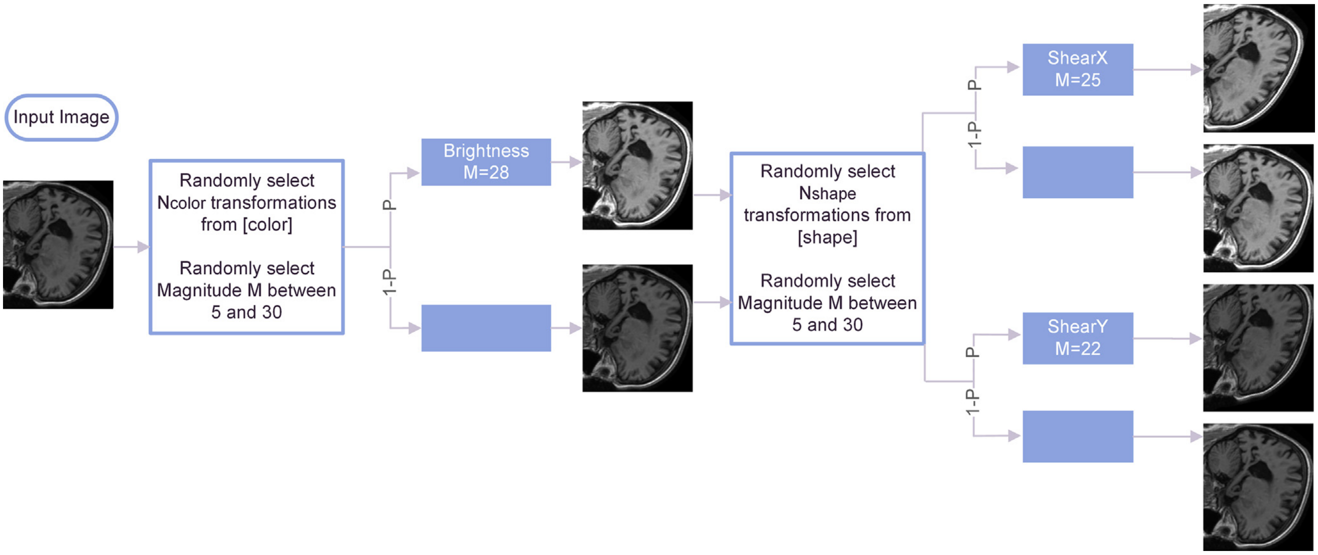
Input Image
Randomly select Ncolor transformations from [color]
Randomly select Magnitude M between 5 and 30
P
1-P
Brightness M=28
Randomly select Nshape transformations from [shape]
Randomly select Magnitude M between 5 and 30
P
1-P
ShearX M=25
P
1-P
ShearY M=22

Fig. 3. Workflow of TRRA when N_{color} and N_{shape} are both equal to 1. The input image is processed by TRRA to generate an augmented image.

Table 5. AD classification performance on ADNI test set of best-performing model of each CNN structure

Model	Performance with TRRA				Performance without data augmentation			
	Accuracy	Sensitivity	Specificity	AUC	Accuracy	Sensitivity	Specificity	AUC
VGG-13	0.912 ± 0.009	0.904 ± 0.026	0.920 ± 0.043	0.962 ± 0.009	0.789 ± 0.018	0.788 ± 0.062	0.791 ± 0.024	0.872 ± 0.022
ResNet-18	0.912 ± 0.004	0.874 ± 0.007	0.950 ± 0.007	0.957 ± 0.008	0.774 ± 0.025	0.753 ± 0.014	0.796 ± 0.037	0.853 ± 0.030
ResNet-50	0.920 ± 0.014	0.904 ± 0.031	0.935 ± 0.019	0.961 ± 0.011	0.784 ± 0.014	0.727 ± 0.025	0.841 ± 0.051	0.865 ± 0.027
SE-ResNet-101	0.922 ± 0.015	0.889 ± 0.019	0.955 ± 0.012	0.960 ± 0.009	0.794 ± 0.014	0.722 ± 0.038	0.866 ± 0.064	0.875 ± 0.029
DenseNet-169	0.932 ± 0.006	0.904 ± 0.014	0.960 ± 0.019	0.961 ± 0.009	0.800 ± 0.018	0.778 ± 0.040	0.821 ± 0.074	0.869 ± 0.026
EfficientNet-B1	**0.932 ± 0.006**	**0.924 ± 0.000**	**0.940 ± 0.012**	**0.961 ± 0.012**	0.777 ± 0.019	0.692 ± 0.038	0.861 ± 0.070	0.870 ± 0.027

Values are presented as Means ± S.D.

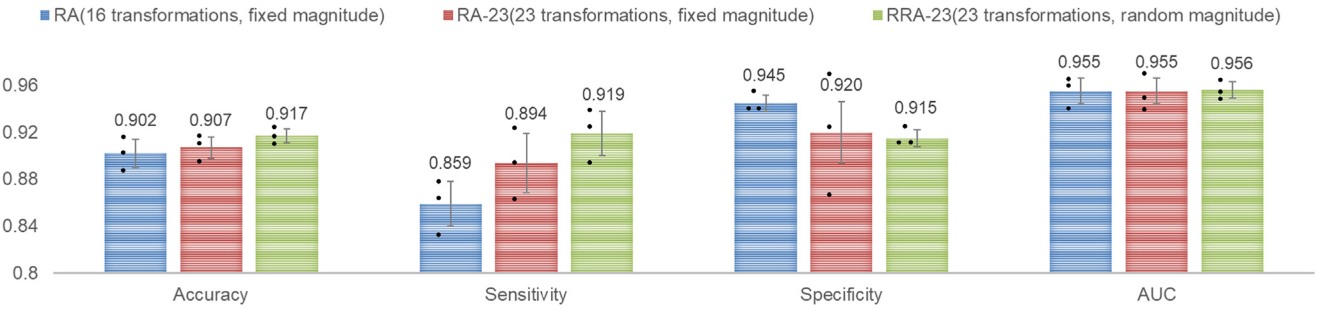

Fig. 4. AD classification performance on ADNI test set of the of RA, RA-23, and RRA-23 in optimal hyperparameters. Black dots superimposed on the bar are data points, and numbers above error bar are mean values.

classification performance. (2) The accuracy and AUC of P is 0.9 are improved by 0.2% and 0.3% compared that when P is 1, which indicates that P can help improve classification performance.

Classification performance on MCI conversion prediction task

Fig. 6 shows the performance of MCI conversion prediction task using different pre-training methods and data

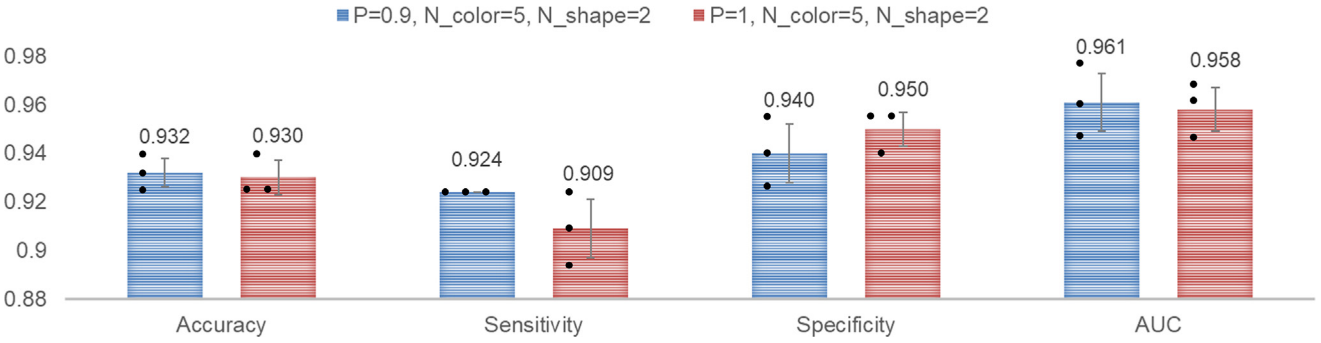

Fig. 5. AD classification performance on ADNI test set of TRRA. Black dots superimposed on the bar are data points, and numbers above error bar are mean values.

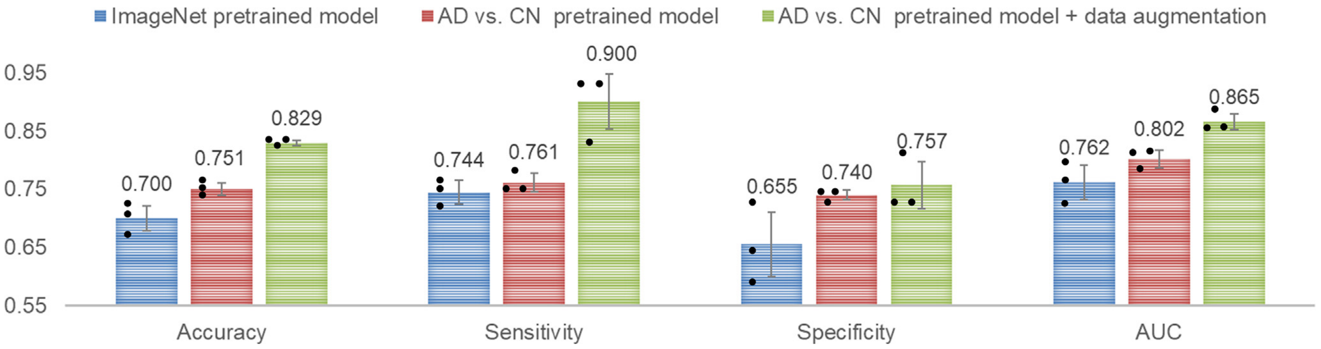

Fig. 6. The classification performance on MCI conversion prediction task. Black dots superimposed on the bar are data points, and numbers above error bar are mean values.

augmentation strategies. We first train ImageNet pre-trained EfficientNet-B1 on ADNI training set of MCI conversion prediction task without using data augmentation and achieve accuracy of 0.700 on test set. Then, we use the EfficientNet-B1 model performing best on AD classification task, and fine-tune it without data augmentation on the ADNI training set of MCI conversion prediction task, and accuracy on ADNI test set achieves 0.751. Compared with the ImageNet pre-trained model, using AD classification task for pre-training improves the accuracy of the MCI conversion prediction task by 5.1%. This proves the effectiveness of using AD classification task for pre-training. Finally, we use TRRA to perform data augmentation during the training process on MCI conversion prediction task, accuracy on ADNI test set is further improved to 0.829, which is increased by 7.8% in the comparison with no data augmentation. This indicates that the performance of the model trained with the proposed data augmentation strategy is better than that of the model trained with unenhanced data in MCI conversion prediction task.

Classification performance on AIBL dataset

AIBL dataset is used to further evaluate the generalization of our proposed approach. Specifically, we choose the EfficientNet-B1 models performing best on ADNI dataset (accuracy is 0.932 ± 0.006 on ADNI test set) and take the AIBL dataset containing 77 AD subjects and 450 CN subjects for testing. Accuracy on AIBL dataset is 0.920 ± 0.006. Noticeably, we do not further fine-tune the model on AIBL dataset, and use all the data as a test set, which is a more difficult choice. To the best of our knowledge, only Wen et al. (Wen et al., 2020) test the ADNI trained model on the AIBL dataset, and our performance is better than theirs. The results verify that our approach generalizes well not only on the dataset from the same research, but also on the dataset from a similar study. Table 6 presents the details experimental results.

Visually explainable heatmaps

Grad-CAM++ has been previously introduced to generate visually explainable heatmaps helping to highlight the brain regions related with predicted target. The visually explainable heatmaps generated from different CNN models are presented in Fig. 7. The regions highlighted on the heatmaps are slightly different due to different network structures. Overall, five models pay more attention to dilation of the lateral ventricle and cortical atrophy.

The dilation of the lateral ventricle and cortical atrophy are the macroscopic features of neuropathological alterations in AD brain (Dickerson et al., 2009; Serrano-Pozo et al., 2011). The models comprehensively consider these regions to make the final subject-level diagnosis, which brings good classification performance.

Comparison with other methods

In this part, we provide a performance comparison table to further compare with most recent and state-of-the-art methods reported in the literature. Table 7 summarizes

Table 6. AD vs. CN classification performance on AIBL dataset

Approach	Accuracy	Sensitivity	Specificity	AUC
Our	0.920 ± 0.006	0.818 ± 0.021	0.937 ± 0.004	0.939 ± 0.003
(Wen et al., 2020)	0.896 ± 0.011	0.771 ± 0.051	0.918 ± 0.020	-

Values are presented as Means ± S.D.

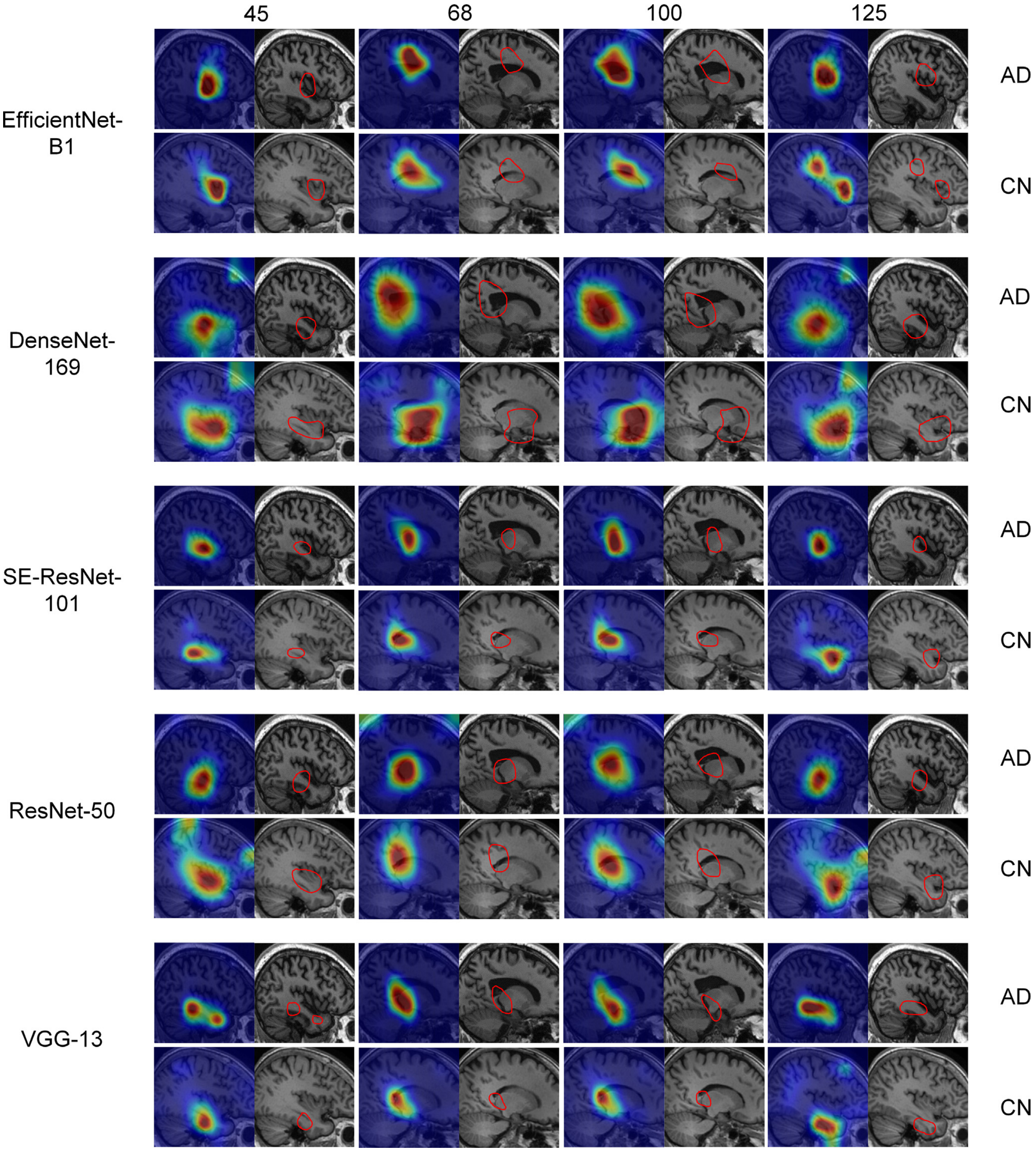

Fig. 7. The visual explanation results of different CNN models on AD classification task. The highlighted regions on heatmaps are of higher correlation with the predicted class, and the boundary of the most important red area is drawn on the original image for easy observation. The numbers on the top indicate the slice positions.

Table 7. A comparative table of methodologies on both AD vs. CN task and pMCI vs. sMCI task using structural MRI data from the ADNI dataset

Study	AD vs. CN				pMCI vs. sMCI				Approach
	Accuracy	Sensitivity	Specificity	AUC	Accuracy	Sensitivity	Specificity	AUC	
Aderghal et al., 2018	0.90	0.94	0.86		-				3D ROI-based
Backstrom et al., 2018	0.90				-				3D subject-level
Cheng et al., 2017	0.87	0.86	0.86	0.92	-				3D patch-level
Danni and Liu, 2017	0.86	0.84	0.90	0.91	-				3D subject-level
Li et al., 2017	0.88	0.91	0.84	0.93	-				3D subject-level
Fan et al., 2018	0.90	0.88	0.91	0.92	-				3D patch-level
Lian et al., 2020	0.90	0.82	0.97	0.95	0.81	0.53	0.85	0.78	3D patch-level
Liu et al., 2017	0.91	0.88	0.94	0.96	0.78	0.42	0.82	0.78	3D patch-level
Liu et al., 2018	0.91	0.87	0.93	0.96	-				3D patch-level
Shmulev and Belyaev, 2018	-				0.62	0.75	0.54	0.70	3D subject-level
Valliani and Soni, 2017	0.81				-				2D slice-level
Spasov et al., 2019	-				0.72	0.63	0.81	0.79	3D subject-level
Liu et al., 2020	0.89	0.87	0.91	0.93	-				3D ROI-based
Kang et al., 2021	0.90				-				2D slice-level
Wen et al., 2020	0.89	0.87	0.90		0.74	0.80	0.68		3D ROI-based
Our method	**0.93**	0.92	0.94	**0.96**	**0.83**	**0.90**	0.76	**0.87**	**2D slice-level**

the methods using sMRI data from the ADNI dataset for AD diagnosis (no data leakage in all methods). As the data indicates, we rank first in accuracy and AUC on both classification tasks. The performance of the proposed 2D single model approach on the two classification tasks both outperforms the existing state-of-the-art methods including those using multi-model and 3D CNN.

DISCUSSION

As introduced previously, despite the existing research is encouraging, deep learning based diagnostic methods for AD and its prodromal period MCI still have some limitations. In this research, we propose a novel end-to-end deep learning approach for the automated diagnosis of AD. The proposed approach outperforms the state-of-the-art approaches, including those using multi-model and 3D CNN methods.

For the AD diagnosis, our approach achieves the accuracy of 0.93, 0.83 for AD classification, MCI conversion prediction on the ADNI dataset respectively, and achieves an accuracy of 0.92 for AD classification on the AIBL dataset. For the first time, we systematically assessed CNN models with different structures and capacities for AD diagnosis. The results in Table 5 and Table A.1 indicate that more advanced model structures like EfficientNet and DenseNet can achieve better performance, and models in similar structure with moderate capacity rather than the largest one can achieve better performance.

Limited by lack of large-scale sMRI dataset, it is not easy to train a model of good classification performance for AD diagnosis. To alleviate this problem, we propose TRRA which is more suitable for AD diagnosis task than RA. The results of the ablation experiments in Fig. 4 and Fig. 5 presents the contribution of each improved elements of TRRA to classification performance. In addition, the experimental results in Fig. 6 also proves that pre-training on AD classification task can improve the classification performance of the MCI conversion prediction task.

Unbiased evaluation of performance is an essential task of deep learning, and the test set should not be used for hyperparameter selection. We therefore choose a rigorous evaluation strategy: Training and validation sets are used for the selection of the model capacity of the five structures and the grid search of the hyperparameters of four data augmentation strategies, and the test set is only adopted for evaluation of the final classifier.

Meanwhile, we introduce Grad-CAM++ to understand how the model makes the classification decision. The heatmaps in Fig. 7 show that our approach pays more attention to the lateral ventricle and some regions of cortex, which have been proved to be affected by AD.

Our approach greatly improves the classification performance of 2D CNN for AD diagnosis and the increases transparency of the model. The systematic evaluation of various CNN models provides a reference for subsequent studies. The proposed data augmentation strategy can greatly improve the diagnostic performance by alleviating the overfitting problem caused by the limited data in medical datasets, and it is also flexible to expend in other imaging modalities and medical datasets. Considering the potential scarcity of data in the medical field, we only use less invasive and cheaper sMRI data that can be obtained in non-tertiary medical center and medium hospitals, which can make our method applicable to a wider clinical environment.

In summary, we propose a novel end-to-end deep learning approach for automated diagnosis of AD from sMRI data. First, CNN models of different structures and capacities are evaluated systemically, and the most suitable model is adopted for AD diagnosis. Then, a data augmentation strategy called TRRA able to alleviate overfitting is proposed to improve classification performance. Meanwhile, to understand how the model makes decisions and increase transparency of our approach, Grad-CAM++ is introduced to generate visually explainable heatmaps. The effectiveness of our proposed approach has been extensively evaluated on two publicly accessible datasets. The experimental results indicate that our approach outperforms the state-of-the-art approaches, including those using multi-model and three-dimensional (3D) CNN methods. The resultant heatmaps from our approach also highlight the lateral ventricle and some regions of cortex, which have been proved to be affected by AD.

DECLARATION OF INTEREST

The authors declare that there are no conflicts of interest.

CRediT authorship contribution statement

Fan Zhang: Methodology, Software, Validation, Formal analysis, Data curation, Writing – original draft, Writing – review & editing. **Bo Pan:** Conceptualization, Formal analysis, Visualization. **Pengfei Shao:** Conceptualization, Methodology, Resources. **Peng Liu:** Methodology, Formal analysis. **Shuwei Shen:** Conceptualization, Methodology, Formal analysis, Writing – review & editing. **Peng Yao:** Methodology, Software, Formal analysis, Writing – review & editing, Supervision, Project administration. **Ronald X. Xu:** Conceptualization, Methodology, Formal analysis, Writing – review & editing, Visualization, Supervision, Project administration.

ACKNOWLEDGEMENTS

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California. This study was supported by National Key R&D Program of China (Grant Number: 2021YFC2401402) and The Strategic Priority Research Program of the Chinese Academy of Sciences (Grant Number: XDA16021303).

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: 31st IEEE International Symposium on Computer-Based Medical Systems (CBMS). p. 345–350. https://doi.org/10.1109/cbms.2018.00067.

Anonymous (2020) 2020 Alzheimer's disease facts and figures. Alzheimers Dement 16(3):391–460. https://doi.org/10.1002/alz.12068.

Avants BB, Epstein CL, Grossman M, Gee JC (2008) Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. Med Image Anal 12(1):26–41. https://doi.org/10.1016/j.media.2007.06.004.

Avants BB, Tustison NJ, Song G, Cook PA, Klein A, Gee JC (2011) A reproducible evaluation of ANTs similarity metric performance in brain image registration. Neuroimage 54(3):2033–2044. https://doi.org/10.1016/j.neuroimage.2010.09.025.

Avants BB, Tustison NJ, Stauffer M, Song G, Wu BH, Gee JC (2014) The Insight ToolKit image registration framework. Front Neuroinformatics 8:13. https://doi.org/10.3389/fninf.2014.00044.

Avants BB, Yushkevich P, Pluta J, Minkoff D, Korczykowski M, Detre J, Gee JC (2010) The optimal template effect in hippocampus studies of diseased populations. Neuroimage 49(3):2457–2466. https://doi.org/10.1016/j.neuroimage.2009.09.062.

Backstrom K, Nazari M, Gu YH, Jakola AS. (2018). An efficient 3D deep convolutional network for Alzheimer's disease diagnosis using MR images. 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018). 149-153, 10.1109/isbi.2018.8363543.

Beheshti I, Demirel H (2015) Probability distribution function-based classification of structural MRI for the detection of Alzheimer's disease. Comput Biol Med 64:208–216. https://doi.org/10.1016/j.compbiomed.2015.07.006.

Belkin M, Hsu D, Ma SY, Mandal S (2019) Reconciling modern machine-learning practice and the classical bias-variance trade-off. Proc Natl Acad Sci U S A 116(32):15849–15854. https://doi.org/10.1073/pnas.1903070116.

Cao P, Gao J, Zhang ZP (2020) Multi-view based multi-model learning for MCI diagnosis. Brain Sci 10(3). https://doi.org/10.3390/brainsci10030181.

Chattopadhay A, Sarkar A, Howlader P, Balasubramanian VN, Ieee (2018). Grad-CAM plus plus: generalized gradient-based visual explanations for deep convolutional networks. 2018 Ieee Winter Conference on Applications of Computer Vision. New York, Ieee: 839-847.

Cheng D, Liu M, Fu J, Wang Y (2017) Classification of MR brain images by combination of multi-CNNs for AD diagnosis. In: Ninth International Conference on Digital Image Processing (ICDIP 2017). p. 10420. https://doi.org/10.1117/12.2281808.

Christian S, Antonio C, Petronilla B, Maria C (2015) Magnetic resonance imaging biomarkers for the early diagnosis of Alzheimer's disease: a machine learning approach. Front Neurosci 9:307. https://doi.org/10.3389/fnins.2015.00307.

Cubuk ED, Zoph B, Mane D, Vasudevan V, Le QV, Soc IC (2019). AutoAugment: learning augmentation strategies from data. 2019 Ieee/Cvf Conference on Computer Vision and Pattern Recognition. Los Alamitos, Ieee Computer Soc: 113-123.

Cubuk ED, Zoph B, Shlens J, Le QV (2020) Randaugment: Practical automated data augmentation with a reduced search space. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. p. 3008–3017. https://doi.org/10.1109/cvprw50498.2020.00359 2020.

Danni C, Liu M. (2017). CNNs based multi-modality classification for AD diagnosis. 2017 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics: 5 pp. Doi: 10.1109/cisp-bmei.2017.8302281.

Deng J, Dong W, Socher R, Li K, Li FF, Li and Ieee (2009). ImageNet: a large-scale hierarchical image database. Cvpr: 2009 Ieee Conference on Computer Vision and Pattern Recognition, Vols 1-4. New York, Ieee: 248-255.

Dickerson BC, Bakkour A, Salat DH, Feczko E, Pacheco J, Greve DN, Grodstein F, Wright CI, et al. (2009) The cortical signature of Alzheimer's disease: regionally specific cortical thinning relates to symptom severity in very mild to mild AD dementia and is detectable in asymptomatic amyloid-positive individuals. Cereb Cortex 19(3):497–510. https://doi.org/10.1093/cercor/bhn113.

Ellis KA, Bush AI, Darby D, De Fazio D, Foster J, Hudson P, Lautenschlager NT, Lenzo N, et al. (2009) The Australian Imaging, Biomarkers and Lifestyle (AIBL) study of aging: methodology and baseline characteristics of 1112 individuals recruited for a longitudinal study of Alzheimer's disease. Int Psychogeriatr 21 (4):672–687. https://doi.org/10.1017/s1041610209009405.

Fan, Liu M, A. S. D. N. Initiative (2018) Alzheimer's disease diagnosis based on multiple cluster dense convolutional networks. Comput Med Imag Graph. https://doi.org/10.1109/IST.2017.8261460.

Fan ZH, Li J, Zhang L, Zhu GM, Li P, Lu XY, Shen PY, Shah SAA, et al. (2021) U-net based analysis of MRI for Alzheimer's disease diagnosis. Neural Comput Appl 13. https://doi.org/10.1007/s00521-021-05983-y.

Farooq A, Anwar SM, Awais M, Rehman S, Ieee (2017) A deep CNN based multi-class classification of Alzheimer's disease using MRI. In: 2017 Ieee International Conference on Imaging Systems and Techniques. New York: Ieee. p. 111–116.

Fonov V, Evans AC, Botteron K, Almli CR, McKinstry RC, Collins DL, G. Brain Dev Cooperative (2011) Unbiased average age-appropriate atlases for pediatric studies. Neuroimage 54 (1):313–327. https://doi.org/10.1016/j.neuroimage.2010.07.033.

Fonov VS, Dadar M, Collins DL (2018). ''Deep learning of quality control for stereotaxic registration of human brain MRI.'' bioRxiv, DOI: https://doi.org/10.1101/303487.

Fonov VS, Evans A, Mckinstry RC, Almli CR, Collins DL (2009) Unbiased nonlinear average age-appropriate brain templates from birth to adulthood. Neuroimage 47:S102. https://doi.org/10.1016/s1053-8119(09)70884-5.

Gorgolewski KJ, Auer T, Calhoun VD, Craddock RC, Das S, Duff EP, Flandin G, Ghosh SS, et al. (2016) The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. Sci Data 3:9. https://doi.org/10.1038/sdata.2016.44.

He KM, Zhang XY, Ren SQ, Sun J, Ieee (2016) Deep residual learning for image recognition. In: Ieee Conference on Computer Vision and Pattern Recognition. New York: Ieee. p. 770–778.

Huang G, Liu Z, van der Maaten L, Weinberger KQ, Ieee (2017) Densely connected convolutional networks. In: 30th Ieee Conference on Computer Vision and Pattern Recognition. New York: Ieee. p. 2261–2269.

Jo T, Nho K, Saykin AJ (2019) Deep learning in Alzheimer's disease: diagnostic classification and prognostic prediction using neuroimaging data. Front Aging Neurosci 11:14. https://doi.org/10.3389/fnagi.2019.00220.

Kang W, Lin L, Zhang B, Shen X, Wu S (2021) Multi-model and multi-slice ensemble learning architecture based on 2D convolutional neural networks for Alzheimer's disease diagnosis. Comput Biol Med 136 104678.

Krizhevsky A, Sutskever I, Hinton GE (2017) ImageNet classification with deep convolutional neural networks. Commun ACM 60 (6):84–90. https://doi.org/10.1145/3065386.

LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521 (7553):436–444. https://doi.org/10.1038/nature14539.

Li F, Cheng D, Liu M (2017) Alzheimer's disease classification based on combination of multi-model convolutional networks. In: 2017 IEEE International Conference on Imaging Systems and Techniques (IST). https://doi.org/10.1109/IST.2017.8261566.

Lian CF, Liu MX, Zhang J, Shen DG (2020) Hierarchical fully convolutional network for joint atrophy localization and Alzheimer's disease diagnosis using structural MRI. IEEE Trans Pattern Anal Mach Intell 42(4):880–893. https://doi.org/10.1109/tpami.2018.2889096.

Lim, S., I. Kim, T. Kim, C. Kim and S. Kim (2019). Fast AutoAugment. Advances in neural information processing systems 32. H. Wallach, H. Larochelle, A. Beygelzimer et al. La Jolla, Neural Information Processing Systems (Nips). 32.

Liu JX, Li MX, Luo YL, Yang S, Li W, Bi YF (2021) Alzheimer's disease detection using depthwise separable convolutional neural networks. Comput Meth Programs Biomed 203:10. https://doi.org/10.1016/j.cmpb.2021.106032.

Liu M, Zhang J, Adeli E, Shen D (2017) Landmark-based deep multi-instance learning for brain disease diagnosis. Med Image Anal 43:157. https://doi.org/10.1016/j.media.2017.10.005.

Liu M, Zhang J, Nie D, Yap P-T, Shen D (2018) Anatomical landmark based deep feature representation for MR images in brain disease diagnosis. IEEE J Biomed Health Inf 22(5):1476–1485. https://doi.org/10.1109/JBHI.2018.2791863.

Liu MH, Li F, Yan H, Wang KD, Ma YX, Shen L, Xu MQ, I. Alzheimers Dis Neuroimaging (2020) A multi-model deep convolutional neural network for automatic hippocampus segmentation and classification in Alzheimer's disease. Neuroimage 208:15. https://doi.org/10.1016/j.neuroimage.2019.116459.

Liu MX, Zhang DQ, Shen D, Alzheimer's Dis N (2015) View-centralized multi-atlas classification for Alzheimer's disease diagnosis. Hum Brain Mapp 36(5):1847–1865. https://doi.org/10.1002/hbm.22741.

Livingston G, Sommerlad A, Orgeta V, Costafreda SG, Huntley J, Ames D, Ballard C, Banerjee S, et al. (2017) Dementia prevention, intervention, and care. Lancet 390(10113):2673–2734. https://doi.org/10.1016/s0140-6736(17)31363-6.

Mingxing T, Quoc L (2019) EfficientNet: rethinking model scaling for convolutional neural networks. In: Kamalika C, Ruslan S, editors. Proceedings of the 36th International Conference on Machine Learning. PMLR. p. 6105–6114.

Möller C, Pijnenburg YAL, van der Flier WM, Versteeg A, Tijms B, de Munck JC, Hafkemeijer A, Rombouts S, et al. (2016) Alzheimer disease and behavioral variant frontotemporal dementia: automatic classification based on cortical atrophy for single-subject diagnosis. Radiology 279(3):838–848. https://doi.org/10.1148/radiol.2015150220.

Nguyen MH, de la Torre F (2010) Optimal feature selection for support vector machines. Pattern Recogn 43(3):584–591. https://doi.org/10.1016/j.patcog.2009.09.003.

Prakash D, Madusanka N, Bhattacharjee S, Kim CH, Park HG, Choi HK (2021) Diagnosing Alzheimer's disease based on multiclass mri scans using transfer learning techniques. Curr Med Imaging 17(12):1460–1472. https://doi.org/10.2174/1573405617666210127161812.

Raschka S (2015) Python machine learning. Packt publishing ltd.

Rathore S, Habes M, Iftikhar MA, 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. https://doi.org/10.1016/j.neuroimage.2017.03.057.

Routier A, Burgos N, Díaz M, Bacci M, Bottani S, El-Rifai O, Fontanella S, Gori P, Guillon J, et al. (2018) Clinica: an open source software platform for reproducible clinical neuroscience studies. Front Neuroinformatics 15. https://doi.org/10.3389/fninf.2021.689675.

Samper-González J, Burgos N, Bottani S, Fontanella S, Lu P, Marcoux A, Routier A, Guillon J, et al. (2018) Reproducible evaluation of classification methods in Alzheimer's disease:

framework and application to MRI and PET data. Neuroimage 183:504–521. https://doi.org/10.1016/j.neuroimage.2018.08.042.

Sandler M, Howard A, Zhu ML, Zhmoginov A, Chen LC, Ieee (2018) MobileNetV2: inverted residuals and linear bottlenecks. In: 2018 Ieee/Cvf Conference on Computer Vision and Pattern Recognition. New York: Ieee. p. 4510–4520.

Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D, Ieee (2017) Grad-CAM: Visual explanations from deep networks via gradient-based localization. In: 2017 Ieee International Conference on Computer Vision. New York: Ieee. p. 618–626.

Serrano-Pozo A, Frosch MP, Masliah E, Hyman BT (2011) Neuropathological Alterations in Alzheimer Disease. Cold Spring Harb Perspect Med 1(1):23. https://doi.org/10.1101/cshperspect.a006189.

Shmulev Y, Belyaev M (2018). Predicting conversion of mild cognitive impairments to Alzheimer's disease and exploring impact of neuroimaging. 83-91, 10.1007/978-3-030-00689-1_9.

Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. Comput Sci.

Spasov S, Passamonti L, Duggento A, Lio P, Toschi N, I., Alzheimers, Dis, Neuroimaging, (2019) A parameter-efficient deep learning approach to predict conversion from mild cognitive impairment to Alzheimer's disease. Neuroimage 189:276–287. https://doi.org/10.1016/j.neuroimage.2019.01.031.

Tan MX, Chen B, Pang RM, Vasudevan V, Sandier M, Howard A, Le QV, I. C. Soc (2019) MnasNet: Platform-aware neural architecture search for mobile. In: 2019 Ieee/Cvf Conference on Computer Vision and Pattern Recognition. Los Alamitos: Ieee Computer Soc. p. 2815–2823.

Tiwari S, Atluri V, Kaushik A, Yndart A, Nair M (2019) Alzheimer's disease: pathogenesis, diagnostics, and therapeutics. Int J Nanomed 14:5541–5554. https://doi.org/10.2147/ijn.S200490.

Tustison NJ, Avants BB, Cook PA, Zheng YJ, Egan A, Yushkevich PA, Gee JC (2010) N4ITK: improved N3 bias correction. IEEE Trans Med Imaging 29(6):1310–1320. https://doi.org/10.1109/tmi.2010.2046908.

Valliani A, Soni A (2017) Deep residual nets for improved Alzheimer's diagnosis. Acm Int Conf Bioinf ACM 615. https://doi.org/10.1145/3107411.3108224.

Vu TD, Ho NH, Yang HJ, Kim J, Song HC (2018) Non-white matter tissue extraction and deep convolutional neural network for Alzheimer's disease detection. Soft Comput 22(20):6825–6833. https://doi.org/10.1007/s00500-018-3421-5.

Wang HF, Shen YY, Wang SQ, Xiao TF, Deng LM, Wang XY, Zhao XY (2019) Ensemble of 3D densely connected convolutional network for diagnosis of mild cognitive impairment and Alzheimer's disease. Neurocomputing 333:145–156. https://doi.org/10.1016/j.neucom.2018.12.018.

Ward A, Tardiff S, Dye C, Arrighi HM (2013) Rate of conversion from prodromal Alzheimer's disease to Alzheimer's dementia: a systematic review of the literature. Dementia Geriatric Cognitive Disord Extra 3(1):320–332. https://doi.org/10.1159/000354370.

Wen J, Thibeau-Sutre E, Diaz-Melo M, Samper-González J, Routier A, Bottani S, Dormont D, Durrleman S, et al. (2020) Convolutional neural networks for classification of Alzheimer's disease: overview and reproducible evaluation. Med Image Anal 63:101694.

APPENDIX A. CLASSIFICATION PERFORMANCE OF DIFFERENT CNN MODELS

Table A.1 shows AD classification accuracy of different CNN models on ADNI validation set. We observe that regardless of whether TRRA is implemented during the training process, accuracy of models in similar structure increases to the maximum value and then decreases as the model capacity increases. This indicates that the models with moderate capacity instead of maximum capacity can achieve the best performance.

APPENDIX B. CLASSIFICATION PERFORMANCE OF DIFFERENT DATA AUGMENTATION STRATEGIES

Table B.1 shows AD classification accuracy of EfficientNet-B1 with different data augmentation strategies on ADNI validation set. We can observe that accuracy first increases to the maximum value and then decreases as the value of *N* increases for all data augmentation strategies. The decrease in accuracy is most likely because too many transformations are superimposed on the input image, which I leads to an inherent characteristics gap between the augmented image and the original image.

Table A.1. AD classification performance of different CNN models on ADNI validation set

Model	Accuracy without DA	Accuracy with DA	Model	Accuracy without DA	Accuracy with DA	Model	Accuracy without DA	Accuracy with DA
VGG-11	0.797 ± 0.028	0.900 ± 0.009	SE-ResNet-50	0.762 ± 0.019	0.902 ± 0.012	EfficientNet-B1	**0.797 ± 0.016**	**0.915 ± 0.018**
VGG-13	**0.805 ± 0.025**	**0.907 ± 0.013**	SE-ResNet-101	**0.792 ± 0.035**	**0.910 ± 0.018**	EfficientNet-B2	0.764 ± 0.038	0.912 ± 0.015
VGG-16	0.792 ± 0.013	0.902 ± 0.012	SE-ResNet-152	0.787 ± 0.007	0.900 ± 0.004	EfficientNet-B3	0.767 ± 0.034	0.907 ± 0.018
VGG-19	0.782 ± 0.016	0.882 ± 0.031	SENet-154	0.782 ± 0.040	0.892 ± 0.004	EfficientNet-B4	0.779 ± 0.030	0.905 ± 0.022
ResNet-18	0.774 ± 0.037	0.900 ± 0.007	DenseNet-121	0.784 ± 0.009	0.900 ± 0.004	EfficientNet-B5	0.769 ± 0.018	0.905 ± 0.020
ResNet-34	0.790 ± 0.025	0.910 ± 0.018	DenseNet-169	**0.795 ± 0.009**	**0.905 ± 0.020**	EfficientNet-B6	0.767 ± 0.037	0.890 ± 0.010
ResNet-50	**0.792 ± 0.014**	**0.910 ± 0.016**	DenseNet-201	0.777 ± 0.020	0.902 ± 0.021	EfficientNet-B7	0.794 ± 0.033	0.877 ± 0.009
ResNet-101	0.779 ± 0.025	0.905 ± 0.030	DenseNet-161	0.772 ± 0.028	0.902 ± 0.021			
ResNet-152	0.754 ± 0.052	0.905 ± 0.012	EfficientNet-B0	0.792 ± 0.009	0.912 ± 0.007			

Values are presented as Means ± S.D. Accuracy with DA: AD vs. CN classification using TRRA, Accuracy without DA: AD vs. CN classification without applying data augmentation.

Table B.1. AD classification accuracy of EfficientNet-B1 with different data augmentation strategies on ADNI validation set

Method	*M*	*N*							
		1	2	3	4	5	6	7	8
RA	5	0.882 ± 0.026	0.895 ± 0.032	0.892 ± 0.023	0.892 ± 0.025	0.900 ± 0.025	0.890 ± 0.015	**0.902 ± 0.021**	0.900 ± 0.023
	10	0.867 ± 0.032	0.895 ± 0.027	0.892 ± 0.019	0.887 ± 0.027	0.897 ± 0.030	0.897 ± 0.026	0.900 ± 0.023	0.900 ± 0.023
	15	0.872 ± 0.028	0.885 ± 0.029	0.895 ± 0.021	0.895 ± 0.027	0.900 ± 0.028	0.895 ± 0.021	0.897 ± 0.030	0.900 ± 0.028
	20	0.875 ± 0.026	0.892 ± 0.023	0.892 ± 0.023	0.892 ± 0.023	0.897 ± 0.030	0.895 ± 0.027	0.895 ± 0.022	0.900 ± 0.028
	25	0.877 ± 0.034	0.890 ± 0.002	0.888 ± 0.026	0.900 ± 0.023	0.897 ± 0.030	**0.902 ± 0.021**	0.895 ± 0.022	0.900 ± 0.023
	30	0.867 ± 0.034	0.885 ± 0.029	0.892 ± 0.023	0.892 ± 0.023	0.895 ± 0.032	0.890 ± 0.025	0.890 ± 0.020	0.892 ± 0.023
RA-23	5	0.880 ± 0.032	0.882 ± 0.026	0.887 ± 0.021	0.890 ± 0.025	0.897 ± 0.025	0.902 ± 0.021	0.895 ± 0.021	0.902 ± 0.021
	10	0.875 ± 0.040	0.890 ± 0.022	0.892 ± 0.028	0.897 ± 0.025	0.902 ± 0.028	0.902 ± 0.022	**0.905 ± 0.020**	0.902 ± 0.021
	15	0.884 ± 0.032	0.880 ± 0.022	0.882 ± 0.013	0.887 ± 0.021	0.900 ± 0.018	0.902 ± 0.022	0.902 ± 0.021	0.902 ± 0.022
	20	0.872 ± 0.022	0.877 ± 0.023	0.897 ± 0.031	0.895 ± 0.021	0.900 ± 0.013	**0.905 ± 0.015**	0.902 ± 0.012	0.902 ± 0.012
	25	0.872 ± 0.032	0.882 ± 0.025	0.892 ± 0.023	0.900 ± 0.018	0.902 ± 0.012	0.902 ± 0.016	0.895 ± 0.006	0.900 ± 0.015
	30	0.877 ± 0.033	0.882 ± 0.026	0.895 ± 0.022	0.880 ± 0.018	0.900 ± 0.015	0.897 ± 0.013	0.900 ± 0.009	0.885 ± 0.020
RRA-23	[5, 10]	0.872 ± 0.022	0.875 ± 0.032	0.890 ± 0.030	0.895 ± 0.027	0.895 ± 0.011	0.900 ± 0.019	0.902 ± 0.012	0.905 ± 0.022
	[5, 15]	0.870 ± 0.025	0.875 ± 0.040	0.882 ± 0.022	0.885 ± 0.012	0.892 ± 0.019	0.897 ± 0.021	0.897 ± 0.020	0.897 ± 0.094
	[5, 20]	0.872 ± 0.021	0.889 ± 0.022	0.895 ± 0.032	0.887 ± 0.021	0.897 ± 0.020	0.892 ± 0.015	0.902 ± 0.016	0.900 ± 0.094
	[5, 25]	0.865 ± 0.030	0.882 ± 0.020	0.882 ± 0.020	0.887 ± 0.021	0.897 ± 0.020	0.897 ± 0.021	0.895 ± 0.021	0.900 ± 0.007
	[5, 30]	0.872 ± 0.021	0.880 ± 0.016	0.880 ± 0.021	0.887 ± 0.021	0.897 ± 0.020	0.895 ± 0.012	**0.907 ± 0.013**	0.900 ± 0.023

Values are presented as Means ± S.D.

Table C.1. AD classification accuracy on ADNI validation set of TRRA

N_{color}	N_{shape}	*P*					
		0.1	0.3	0.5	0.7	0.9	1
1	6	0.870 ± 0.031	0.882 ± 0.037	0.872 ± 0.022	0.885 ± 0.025	0.882 ± 0.015	0.885 ± 0.023
2	5	0.862 ± 0.034	0.880 ± 0.028	0.892 ± 0.029	0.888 ± 0.028	0.887 ± 0.022	0.900 ± 0.029
3	4	0.880 ± 0.031	0.887 ± 0.028	0.892 ± 0.023	0.902 ± 0.021	0.905 ± 0.020	0.897 ± 0.009
4	3	0.880 ± 0.034	0.890 ± 0.025	0.902 ± 0.022	0.900 ± 0.018	0.910 ± 0.018	0.912 ± 0.009
5	2	0.852 ± 0.023	0.890 ± 0.026	0.902 ± 0.016	0.912 ± 0.022	**0.915 ± 0.018**	**0.915 ± 0.018**
6	1	0.870 ± 0.020	0.900 ± 0.022	0.902 ± 0.016	0.900 ± 0.009	0.910 ± 0.018	0.910 ± 0.022

Values are presented as Means ± S.D.

APPENDIX C. CLASSIFICATION PERFORMANCE OF TRRA

Table C.1 shows AD classification accuracy of EfficientNet-B1 with TRRA under different hyperparameters on ADNI validation set. From the results we observe that randomly selecting five transformations from the [*color*] category and randomly selecting two transformations from the [*shape*] category achieve the best performance, which indicates that [*color*] category contributes more to classification performance than [*shape*] category.

(Received 8 November 2021, Accepted 21 March 2022)
(Available online 7 April 2022)