A Two-Stage Model for Predicting Mild Cognitive Impairment to Alzheimer's Disease Conversion

Peixin Lu[1], Lianting Hu[2,3], Ning Zhang[4], Huiying Liang[2,3], Tao Tian[5]* and Long Lu[1]*

[1] School of Information Management, Wuhan University, Wuhan, China, [2] Medical Big Data Center, Guangdong Provincial People's Hospital, Guangzhou, China, [3] Guangdong Cardiovascular Institute, Guangdong Provincial People's Hospital, Guangzhou, China, [4] School of Business, Qingdao University, Qingdao, China, [5] The First Division of Psychiatry, Jingmen No. 2 People's Hospital, Jingmen, China

Early detection of Alzheimer's disease (AD), such as predicting development from mild cognitive impairment (MCI) to AD, is critical for slowing disease progression and increasing quality of life. Although deep learning is a promising technique for structural MRI-based diagnosis, the paucity of training samples limits its power, especially for three-dimensional (3D) models. To this end, we propose a two-stage model combining both transfer learning and contrastive learning that can achieve high accuracy of MRI-based early AD diagnosis even when the sample numbers are restricted. Specifically, a 3D CNN model was pretrained using publicly available medical image data to learn common medical features, and contrastive learning was further utilized to learn more specific features of MCI images. The two-stage model outperformed each benchmark method. Compared with the previous studies, we show that our model achieves superior performance in progressive MCI patients with an accuracy of 0.82 and AUC of 0.84. We further enhance the interpretability of the model by using 3D Grad-CAM, which highlights brain regions with high-predictive weights. Brain regions, including the hippocampus, temporal, and precuneus, are associated with the classification of MCI, which is supported by the various types of literature. Our model provides a novel model to avoid overfitting because of a lack of medical data and enable the early detection of AD.

Keywords: mild cognitive impairment, Alzheimer's disease, contrastive learning, transfer leaning, MRI, deep learning

INTRODUCTION

Alzheimer's disease (AD), a severe neurodegenerative disease, is the most common type of dementia (Heun et al., 1997; Association, 2019). Nowadays, at least 50 million people worldwide suffer from AD or other types of dementia, and it is expected that this number will reach 131 million in 2050 (Livingston et al., 2017). This further increases the burden of the medical care system in aging societies. Mild cognitive impairment (MCI) is a stage between normal and AD, with 10–12% of people developing AD each year (Petersen, 2000). Based on the progression toward AD, it can be classified into two categories: progressive MCI (pMCI) and stable MCI (sMCI). Although there is no effective treatment for AD at present, its progression can be slowed by medication, memory training, exercise, and diet, which necessitates the early detection of potential patients (Roberson and Mucke, 2006). Neuroimaging techniques, which can detect disease-related neuropathological changes, are valuable tools for assessing and diagnosing AD (Johnson et al., 2012). MRI is one of

the most widely studied neuroimaging techniques because it is non-invasive, generally available, affordable, and capable of distinguishing between different soft tissues (Klöppel et al., 2008).

With the rapid development, deep learning has achieved remarkable progress in a variety of fields, especially in computer vision and medical imaging, where it outperforms traditional machine learning methods (Shen et al., 2017; Bernal et al., 2019; Abrol et al., 2021). Deep learning approaches perform feature selection during model training and loss function optimization without the need for domain experts' prior knowledge. As a result, individuals with no medical expertise can use them for research and applications, especially in the field of medical image analysis (Shen et al., 2017). Notably, Convolutional Neural Network (CNN) has achieved outstanding performance in the classification tasks of AD and normal control (NC) (Liu et al., 2019) and pMCI/sMCI (Choi and Jin, 2018; Spasov et al., 2019). In general, deep neural networks require large samples for model fitting, especially 3D deep neural network models with more parameters. However, as compared with existing million-sample natural image datasets, neuroimaging datasets have a relatively small sample size (Russakovsky et al., 2015), which can possibly be explained by the following factors. At first, collecting large training sets and labeling image data are costly and time consuming (Shen et al., 2017; Irvin et al., 2019). Furthermore, technical and privacy issues also constraints obstruct medical data collection (Irvin et al., 2019). Therefore, preventing model overfitting due to the scarcity of medical samples has become one of the hottest topics in deep learning of neuroimaging.

Transfer learning is a popular method for dealing with a small number of samples. It utilizes a pretrained model with supervised learning on a large labeled dataset (source domain, e.g., ImageNet) and then fine tunes it on the task of interest (target domain). Studies have shown that knowledge transferred from large-scale annotated natural images (ImageNet) to medical images can significantly improve the effectiveness of assisted diagnosis (Tajbakhsh et al., 2016; Raghu et al., 2019). However, standard medical images, such as MRI, CT, and positron emission tomography (PET), are in three dimensional (3D), preventing ImageNet-based pretrained models from being directly transferred to MRI. Converting 3D data into two-dimensional (2D) slices is a typical method, however, this ignores the rich 3D spatial anatomical information and inevitably affects the performance. To address this issue, several studies (Yang et al., 2017; Zeng and Zheng, 2018) have used pretrained 3D models based on natural video datasets (Tran et al., 2014; Carreira and Zisserman, 2017) to transfer to medical imaging tasks, but have not yet achieved the expected results because of the vast difference between these two domains.

Recently, contrastive learning, a self-supervised learning method, has recently been demonstrated to perform superiorly in various vision tasks (Wu et al., 2018; Zhuang C. et al., 2019; Chen X. et al., 2020). Momentum Contrast (MoCo) (He et al., 2020) is a state-of-the-art method in contrastive learning, which minimizes positive pairs variability while maximizing negative pairs variability. Based on existing research concerns, we proposed a two-stage model based on MoCo (He et al., 2020)

to classify sMCI and pMCI. The main contributions of our study are as follows.

1) Systematic evaluation of 3D ResNet models with different structures and selection of the best model for sMCI and pMCI classification. Provides a reference for related studies.
2) A two-stage model is proposed to solve the problem of domain transfer between the source and target domains, which solves the problem of overfitting caused by small samples in sMCI and pMCI classification and improves the classification performance in AD diagnosis. To the best of our knowledge, we first introduce the MoCo in pMCI and sMCI classification.
3) Three-dimensional Gradient-weighted Class Activation Mapping (Grad-CAM), which is widely used for model interpretability, was introduced to get the heatmap that highlights the brain regions our model focuses on and increases the interpretability of the model.

MATERIALS AND METHODS

As indicated in **Figure 1**, our two-stage transfer learning model was divided into three main parts. In our framework, we did not directly transfer trained model on natural image sets or other medical image sets to our research such as previous studies, mainly for the following reasons: at first, the adoption of 3D CNN in this study to preserve more spatial information limits direct transfer learning from 2D natural images; second, the different components of medical image sets may harm the performance. For instance, the Med3D is composed of MRI and CT of brains, lungs, chests, and other organs (Chen et al., 2019), while our MCI data set only includes brain MRI data. The details of each step were described next.

Dataset and Data Preprocessing

Data used in our study were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database, which is the largest open-access AD database with wide popularity in AD-related research. It was launched in 2003 by the several public and private organizations to measure the progression of MCI and early AD by medical image (such as MRI, PET), genomics, biological markers, and neuropsychological assessments (Jack et al., 2008). More information can be found at http://adni.loni.usc.edu/.

As defined in this study, participants with MCI at baseline who developed or did not develop AD within 3 years were referred to as pMCI and sMCI, respectively. To prevent data leakage, only participants' baseline data were selected in this study. Finally, data from 577 MCI subjects (Means $\pm$ std age = 73.08 $\pm$ 7.25 years) were included in our study, and 297 of the MCI was classified as sMCI (51.5%) and the rest 280 were pMCI (48.5%). The demographics and the mini-mental state examination scores (MMSEs) information of subjects is summarized in **Table 1**. Differences in the median age and gender between groups were tested using ANOVA and Fisher's exact test, respectively.
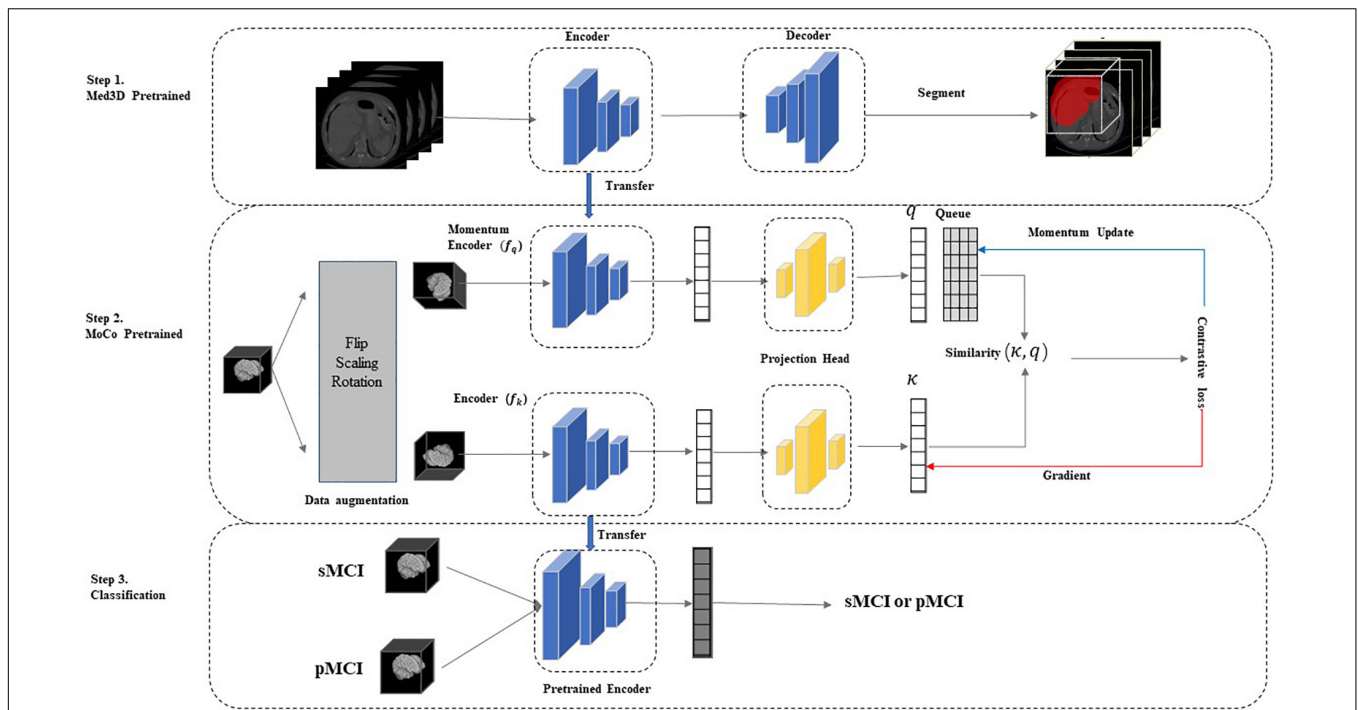
FIGURE 1 | The two-stage model. In step 1, we used the Med3D (Chen et al., 2019) pre-trained model to initialize network weights and learn the common medical image features. In step 2, we performed contrastive learning on unlabeled sMCI and pMCI samples using the improved MoCo to update the network weights, further increasing the correlation between the target and source domains and learning the specific medical features of the sMCI/pMCI classification task. Finally, the network was fine-tuned using labeled sMCI/pMCI samples to achieve sMCI and pMCI classification. MoCo, Momentum Contrast; pMCI, progressive mild cognitive impairment; sMCI, stable mild cognitive, impairment.

TABLE 1 | A summary of the demographic and clinical information of participants.

	Number	Age (years old)	Sex(M/F)	MMSE
sMCI	297	72.2 ± 7.4[55.0,88.4]	174/123	28.0 ± 1.7[23,30]
pMCI	280	73.9 ± 6.9[55.2,88.3]	172/108	26.8 ± 1.8[21,30]

Values are presented as Means ± Standard Deviation [Range]; M, Male; F, Female; MMSE, Mini-Mental State Examination.

These two interactions yielded no statistically significant results ($p > 0.05$).

All 1.5T and 3T structural MRI of the participant were downloaded. The detailed information of MRI, such as scanner protocols, can be found at http://adni.loni.usc.edu/methods/documents/mri-protocols/. Data are preprocessed using FSL[1] with three main steps: bias field correction using the N4 algorithm (Tustison et al., 2010); affine linear alignment of scans onto the MIN152 atlas; skull stripping of each image for $129 \times 145 \times 129$ voxels. **Figure 2** shows the difference before and after preprocessing of the MRI from the same sample.
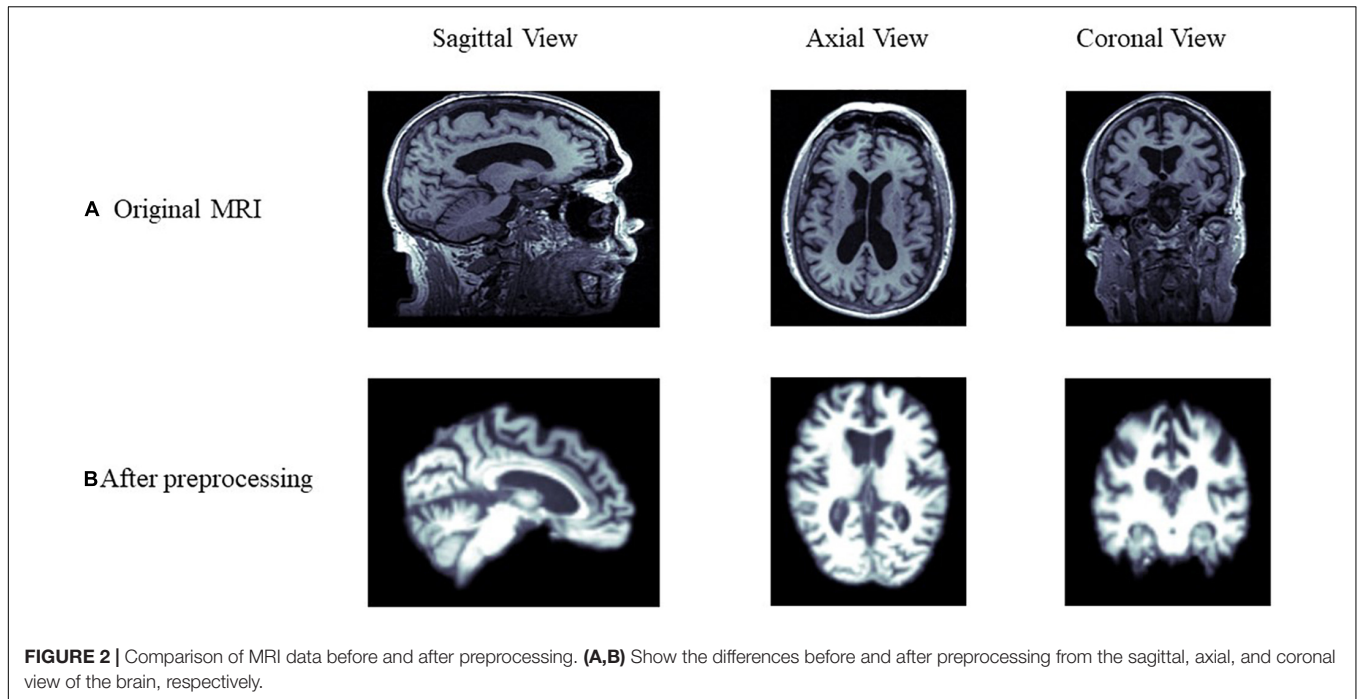
Network Weight Initialization Using Med3D

Many studies have demonstrated that using transfer learning parameter initialization can significantly improve the

[1] https://fsl.fmrib.ox.ac.uk/

performance of models compared to training from scratch (Afzal et al., 2019; Mousavian et al., 2019; Naz et al., 2021). This study selected the Med3D network and its pretrained weights on eight segmented datasets (Chen et al., 2019).

The authors of Med3D integrated data from eight medical segmentation datasets to create the 3DSeg-8 dataset, which contains multiple modalities (MRI and CT), target organs (brain, heart, pancreas), and pathological conditions (CodaLab, 2021 Competition; Menze et al., 2015; Tobon-Gomez et al., 2015; Medical Segmentation Decathlon, 2021). Med3D uses a standard encode–decode partition structure, where the encoder uses the ResNet family. The main idea of ResNet is to introduce the residual block in the network, as illustrated in **Figure 3**, where F(x) is the residual mapping and X is the identity mapping, also called "shortcut." This helps train a deeper network to achieve higher accuracy without vanishing or explosion of the gradient. Notably, Med3D uses a parallel format for model training, which means the same encoder is used for eight datasets, and the decoder is composed of 8 branches accordingly in parallel. This allows the decoder to adapt to different organizational segmentation targets, while the encoder can learn universal features. **Figure 1**, Step1 depicts the Med3D structure. Med3D pretrained models can be used for classification, detection, and segmentation. We used the parameters of the models pretrained by the 3DSeg-8 dataset for the initialization of our network. Transfer learning strategy effects were evaluated in various

FIGURE 2 | Comparison of MRI data before and after preprocessing. **(A,B)** Show the differences before and after preprocessing from the sagittal, axial, and coronal view of the brain, respectively.

ResNet networks in Med3D, including, ResNet-18, ResNet-34, ResNet-50, ResNet-101, ResNet-152 (He et al., 2016).

Transfer Learning Using Contrastive Learning

We extracted general 3D medical image features by Med3D (Chen et al., 2019). However, while Med3D includes MRI and CT of brains, lungs, chests, and other organs, whereas our MCI data set only include brain MRI data, there are still domain transfer concerns between the dataset of Med3D and MRI of sMCI and pMCI (Chen et al., 2019), while our MCI data set is only composed of brain MRI data. Contrastive learning, a special unsupervised learning method with a supervisory function, was introduced in this study to further increase the correlation between the target and source domains. The labels of contrastive learning are generated by the data itself rather than by manual labeling (Liu et al., 2020). It uses unlabeled data to train models and learn embeddings of the data by maximizing the consistency between different augmented views of the same sample and minimizing the consistency between different samples through a contrast loss function (Tian et al., 2020).

Currently, there are various representative contrastive learning methods, such as MoCo (He et al., 2020), SimCLR (Chen T. et al., 2020), and PIRL (Misra and van der Maaten, 2020). Because of the sample size and computational resources constraints, we chose MoCo as our pretraining model in our study. Unlike the end-to-end gradient update of SimCLR, MoCo introduces a dynamic queueing dictionary, which is updated by adding new training batches to the queue while removing the oldest ones from the queue according to the first-in-first-out principle and keeping the length of the queueing dictionary unchanged. This approach allows MoCo to obtain good training results with small batch size.

Given and preprocessed sample x, contrastive learning obtains two augmented samples x_q and x_k by data augmentation of sample x. x_q and x_k are referred to as query data and key data, respectively. q and k are the latent representation of x_q and x_k using a query encoder $q=f_q(x_q;\theta_q)$ and key encoder $k=f_k(x_k;\theta_k)$ with weight θ_q and θ_k, respectively. If the query and the key belong to the same sample, it is marked as a positive pair $k+$. Otherwise, it is a negative sample pair $k-$. The auxiliary task in contrastive learning is: given a pair (x_q, x_k), determining whether it is a
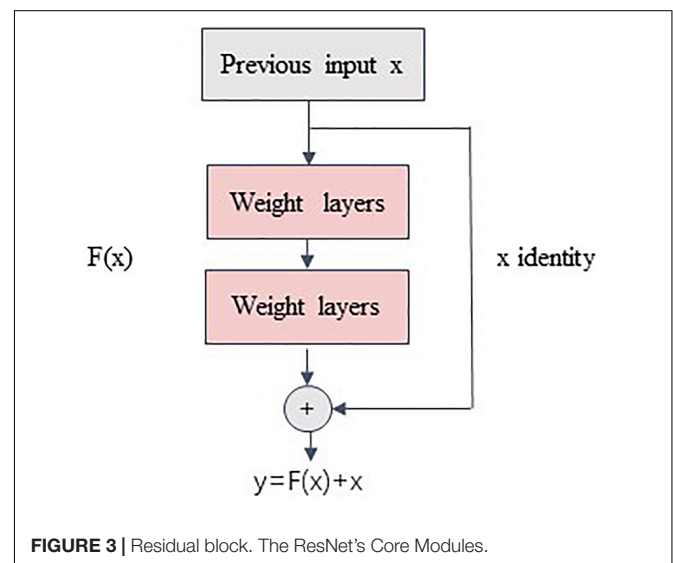

FIGURE 3 | Residual block. The ResNet's Core Modules.

positive or negative sample pair and making the positive samples closer together and the negative samples further apart.

Give a query q. MoCo applies a queue storing a set of keys k from different samples, including one k^+ and several k^-. The contrastive loss can be defined as:

$$\mathcal{L}_{q,k+,\{k^-\}} = -\log \frac{\exp(q \cdot k^+/\tau)}{\exp(q \cdot k^+/\tau) + \sum_{k^-} \exp(q \cdot k^-/\tau)} \quad (1)$$

Here, τ is the temperature parameter. The model updates the encoder weights by minimizing the contrastive loss.

In MoCo, the key encoder is neither updated by backpropagation nor copied from the query encoder. Still, a running average of the key encoder is used, which is known as a momentum encoder. The updating of θ_q and θ_k can be formulated as:

$$\theta_q \leftarrow \theta_q - \alpha \frac{\partial \mathcal{L}}{\partial \theta_q}$$
$$\theta_k \leftarrow m\theta_k + (1 - m)\theta_q \quad (2)$$

Here, m $\in$ [0, 1) is the momentum coefficient. As in Eq. (2), θ_k is updated more smoothly than θ_q which is updated by backpropagation.

It was shown that data augmentation methods such as Gaussian blur and color distortion for natural images might not be applicable in the medical image. For example, Gaussian blur on MRI can potentially change the label of the data. Therefore, we improved the data augmentation method in MoCo by using random rotation ($\pm$ 10°), horizontal flip, and random scaling ($\pm$ 10%) on 3D MRI. In detail, we rotated images at any angle between –10° and +10° along the three axes. Random scaling was also applied to randomly scale the image by $\pm$ 10%. If the image size is larger than the size of the original image, the same volume of the original image is extracted by cropping the center region of the image. If the volume is less than its original size, it is filled with 0 in the reduced region. **Figure 4** shows the schematic diagram of the three data augmentation methods.

We performed MoCo on the full unlabeled MRI data using the Med3D pretrained ResNet as the encoder. In addition, as it is shown that adding a projection head helps to learn feature representation better (Chen T. et al., 2020; Chen X. et al., 2020), we added a projection head, as shown in **Figure 1**. We used two MLP with 128-D hidden layers and a ReLU activation function as the projection head as:

$$z_q = g(q) = W^2\sigma(W^1q)$$
$$z_k = g(k) = W^2\sigma(W^1k) \quad (3)$$

where W^1 and W^2 represent the hidden and output layer weights, and σ is the ReLU activation function.

Classification

The last step of our model is classification, where the labeled data were divided into training, validation, and test sets to fine-tune the pretrained encoders. We added a linear classifier to the frozen backbone model to complete the classification of sMCI and pMCI as proposed by Chen X. et al. (2020).

Model Evaluation

We first used different ResNet, including ResNet-10, ResNet-34, ResNet-50, ResNet-101 as our two-stage model backbone network, and selected the ResNet with the best performance (ResNet-50, see in results) as our backbone in the following comparative experiment.

Evaluation of Transferring Learning Strategies

We first conducted the following comparative experiments with different transfer learning strategies. All the models in different transfer learning strategies used the best ResNet selected by the first experiment.

λ Med3D, pretrained ResNet using Med3D and fine-tuned the model to complete the classification of sMCI and pMCI.

λ MoCo, random initialization of weights, followed with the modified MoCo in the method to train the ResNet without using sMCI/pMCI labels. Then fine-tune the model using labeled data and do the classification.

λ Only ResNet, random initialization of weights, and training ResNet from scratch.

λ Med3D+MoCo, our two-stage model.

Comparison With Previous Studies

To comprehensively understand the performance of our method, we reviewed the state-of-the-art literature, which utilized deep learning to predict the conversion from MCI to Alzheimer's using MRI. We selected studies that achieved criteria for a fair comparison, including (1) only used MRI. (2) published in the last 3 years. (3) the data were from ADNI.

We selected five evaluation metrics to evaluate our model accurately. (1) Accuracy (Acc) is used to measure the proportion of correctly classified samples. (2) Sensitivity (Sens), also known as the true positive rate, is the proportion of predicted positive results that are true positives. (3) Specificity (Spec) is the proportion of correctly identified negatives. (4) F1-score (F1) is the reconciled average of sensitivity and retrieval. (5) Area Under ROC Curve (AUC) represents how the false-positive rate increases with the true-positive rate and increases the area under the characteristic curve. The aforementioned evaluation metrics were calculated based on True Positive (TP), False Positive (FP), False Negative (FN), and True Negative (TN), and these four indicators form a confusion matrix.
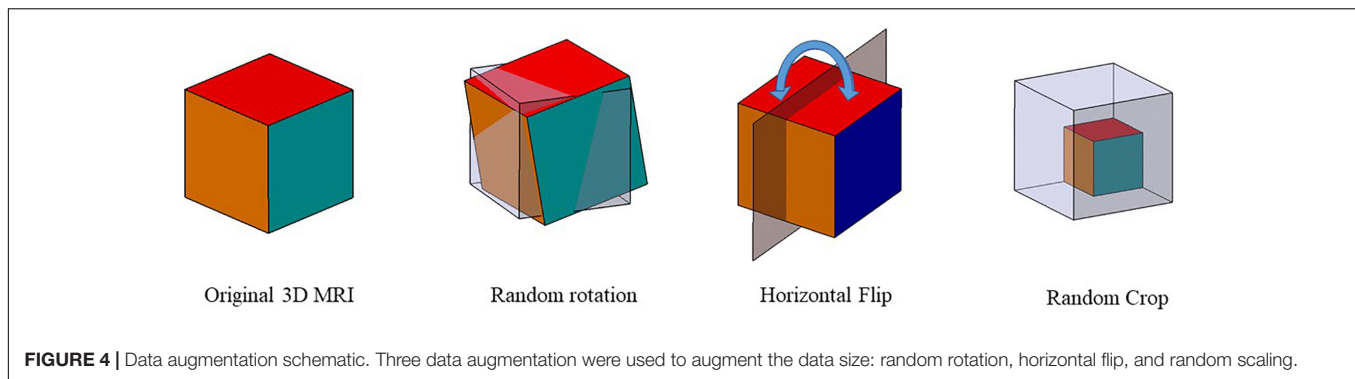
In our study, sMCI and pMCI were referred to as positive and negative examples, respectively. We can calculate accuracy, sensitivity, specificity, and F1 as follows:

$$Acc = \frac{TP + TN}{TP + TN + FN + FP} \quad (4)$$

$$SPE = \frac{TN}{TN + FP} \quad (5)$$

$$SEN = \frac{TP}{TP + FN} \quad (6)$$

$$\mathbb{F}1 = 2 \times \frac{TP}{(2 \times TP + FP + FN)} \quad (7)$$

FIGURE 4 | Data augmentation schematic. Three data augmentation were used to augment the data size: random rotation, horizontal flip, and random scaling.

Finally, we used the non-parametric bootstrap to construct each evaluation metrics' CIs. A total of 10,000 bootstrap replicates were extracted from the test set, and the performance metrics of the model on each bootstrap replicate were calculated. This generated a distribution for each estimate and reported 95% bootstrap percentile intervals (2.5 and 97.5 percentile) (Efron and Tibshirani, 1993).

Experimental Settings

The model was implemented in PyTorch. We used Stochastic Gradient Descent (SGD) with a weight decay of 0.0001 and momentum of 0.99 as our optimizer. A minibatch size of 16 and a cosine annealing learning rate with an initialized value of 0.01 were used in training. Other hyperparameters are the same as default values in MoCo (He et al., 2020). All unlabeled data were used to train MoCo. We trained the classifier with optimized cross-entropy loss and a learning rate of 0.001 in 100 epochs. The dataset was randomly split into training and testing data with a ratio of 8:2. Optimal hyperparameters were selected using fivefold cross-validation on the training set, and the optimal model was chosen for model evaluation on the test set (**Figure 5**). All experiments were run on NVIDIA GTX 2080.

RESULTS

Results of Different ResNet Models Using a Two-Stage Model

In this part, we investigated the classification performance of different Med3D pretrained CNN backbones on the test set, including ResNet-18, ResNet-34, ResNet-50, and ResNet-101. As highlighted in **Table 2** and shown in **Figure 6A**, ResNet-50 achieved the best performance with an accuracy of 0.819 and an AUC of 0.835, indicating complex models with more parameters may not always work best. ResNet-50, the model with the best performance, was used in our following experiments. The confusion maps of different ResNet are shown in **Figure 6B**.

Results of Different Transfer Learning Strategies Using ResNet-50

Table 3 and **Figure 7** show the results of the comparison of accuracy, sensitivity, specificity, F1, and ROC for different

transfer learning strategies (based on ResNet-50) mentioned in section "Experimental Settings," where the optimal results are highlighted. As **Table 3** and **Figure 7A** indicate, our method achieves better results compared to other methods in terms of accuracy (0.819), sensitivity (0.786), specificity (0.850), and F1 score (0.807). Similarly, **Figure 7A** shows ROC curves of different transfer learning strategies, where our method has the best AUC of 0.835 compared with other methods. The confusion maps of different transfer learning strategies are shown in **Figure 7B**.

Results of the Relevant Brain Region

In this study, the 3D Grad-CAM method was used to identify brain regions associated with sMCI or pMCI and improve the interpretability of the model. After weight back-propagation of trained models, we obtained average relevance heatmaps of each class in the test dataset. For comparison, we highlighted temporal superior, temporal middle, hippocampus, thalamus, precuneus, cingulate in the automated anatomical labeling (AAL2)[2] in **Figure 8** first row. **Figure 8** second and third rows show each class's last convolutional layer's attention heatmap. As shown in **Figure 8**, the hippocampus, temporal superior, temporal middle, thalamus, and cingulate are relevant for both sMCI and pMCI. But precuneus is recognized as a unique feature of pMCI.

Comparisons With Previous Studies

We further used four evaluation metrics to compare our results to previous state-of-the-art deep learning studies on sMCI/pMCI classification published in the last 3 years, including accuracy, sensitivity, specificity, and AUC. **Table 4** summarizes the results in relation to the literature, and the best results are indicated by the bold text. In the classification tasks of sMCI and pMCI using deep learning, our method achieves better or comparable classification results in terms of accuracy, specificity, and AUC.

DISCUSSION

This study proposed a two-stage method that combined contrastive learning and transfer learning for predicting conversion from MCI to AD based on MRI. Pretrained models from sizeable medical image datasets were used to initialize the
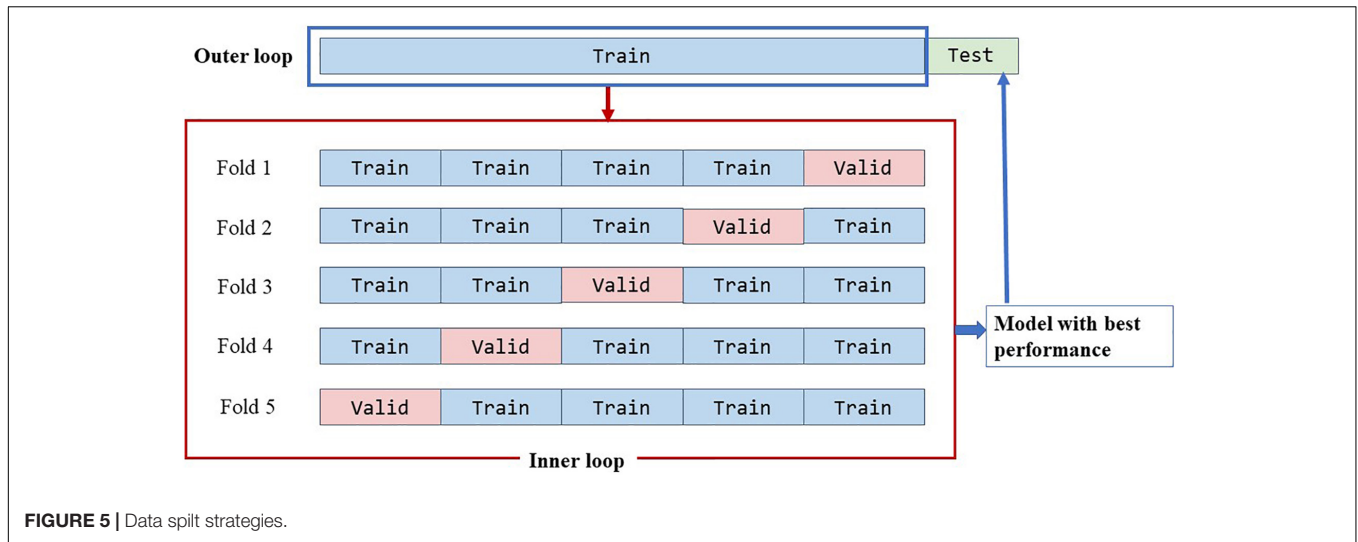
---

[2]http://www.gin.cnrs.fr/en/tools/aal-aal2/

FIGURE 5 | Data spilt strategies.

TABLE 2 | Performance of different ResNet using two-stage model.

Encoder	Acc (95% CI)	Sens (95% CI)	Spec (95% CI)	F1 (95% CI)
ResNet-18	0.707 (0.674, 0.735)	0.679 (0.637, 0.720)	0.733 (0.698, 0.772)	0.691 (0.651, 0.723)
ResNet-34	0.716 (0.690, 0.741)	0.679 (0.644, 0.715)	0.750 (0.716, 0.790)	0.697 (0.670, 0.726)
ResNet-50	**0.819 (0.798, 0.841)**	**0.786 (0.754, 0.821)**	**0.850 (0.815, 0.877)**	**0.807 (0.783, 0.834)**
ResNet-101	0.759 (0.730, 0.783)	0.767 (0.731, 0.808)	0.750 (0.716, 0.785)	0.754 (0.724, 0.779)

The bold numbers denote the maximum value of each column.
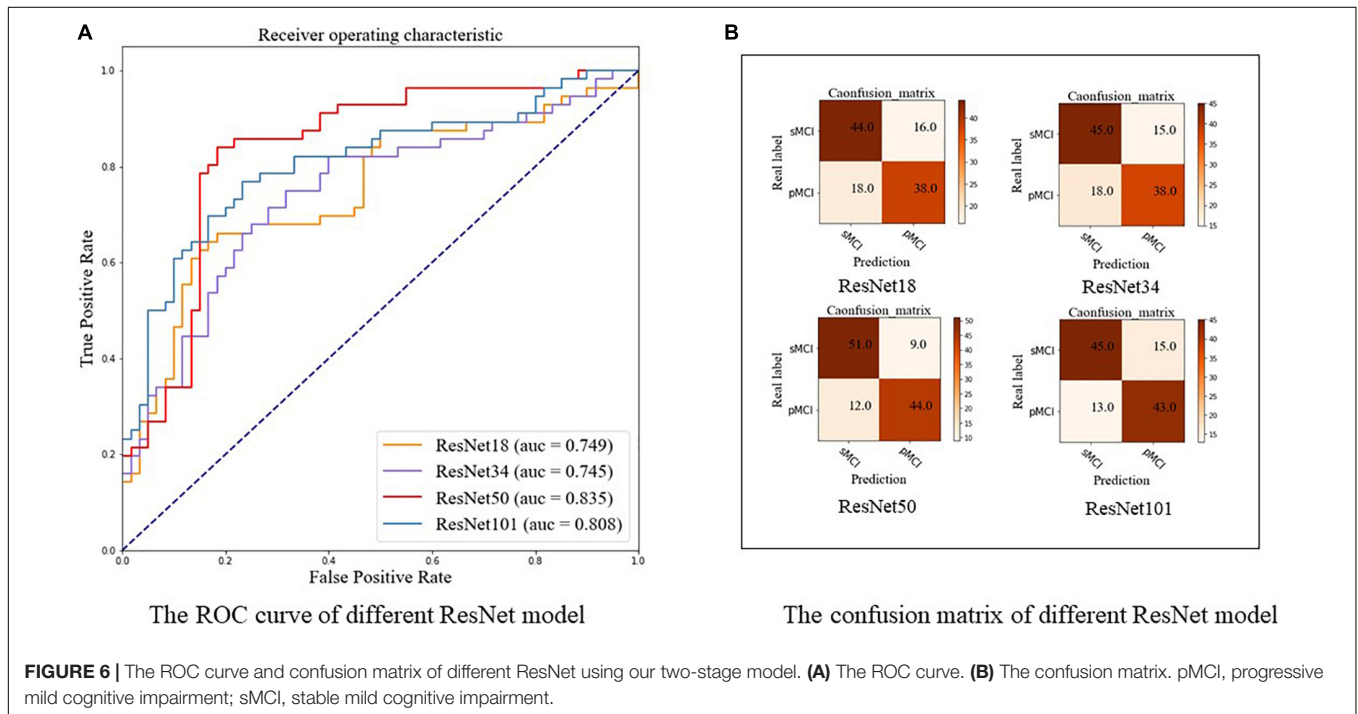Acc, Accuracy; CI, Confidence Interval; F1, F1-score; Sens, Sensitivity; Spec, Specificity.

FIGURE 6 | The ROC curve and confusion matrix of different ResNet using our two-stage model. **(A)** The ROC curve. **(B)** The confusion matrix. pMCI, progressive mild cognitive impairment; sMCI, stable mild cognitive impairment.

model parameters and obtain general imaging features. Training on unlabeled target datasets using contrastive learning was used to get target imaging features. At last, the network was fine-tuned using the labeled target dataset to complete the classification.

In addition, 3D Grad-CAM was used to highlight brain regions potentially associated with pMCI/sMCI classification. We demonstrated the validity of our model through multiple evaluation experiments. The experiments showed that our model

TABLE 3 | Performance of different transfer learning strategies using ResNet-50.

	Acc (95% CI)	Sens (95% CI)	Spec (95% CI)	F1 (95% CI)
Med3D	0.655 (0.630,0.682)	0.661 (0.623,0.702)	0.650 (0.609,0.685)	0.649 (0.619, 0.6768)
MoCo	0.733 (0.704, 0.756)	0.786 (0.747, 0.820)	0.683 (0.651, 0.725)	0.740 (0.712, 0.762)
Only ResNet-50	0.716 (0.694, 0.744)	0.750 (0.712, 0.789)	0.683 (0.656, 0.722)	0.718 (0.694, 0.744)
Med3D+MoCo	**0.819 (0.798, 0.841)**	**0.786 (0.754, 0.821)**	**0.850 (0.815, 0.877)**	**0.807 (0.783, 0.834)**

The bold numbers denote the maximum value of each column.
Acc, Accuracy; CI, Confidence Interval; F1, F1-score; MoCo, Momentum Contrast; Sens,Sensitivity; Spec, Specificity.
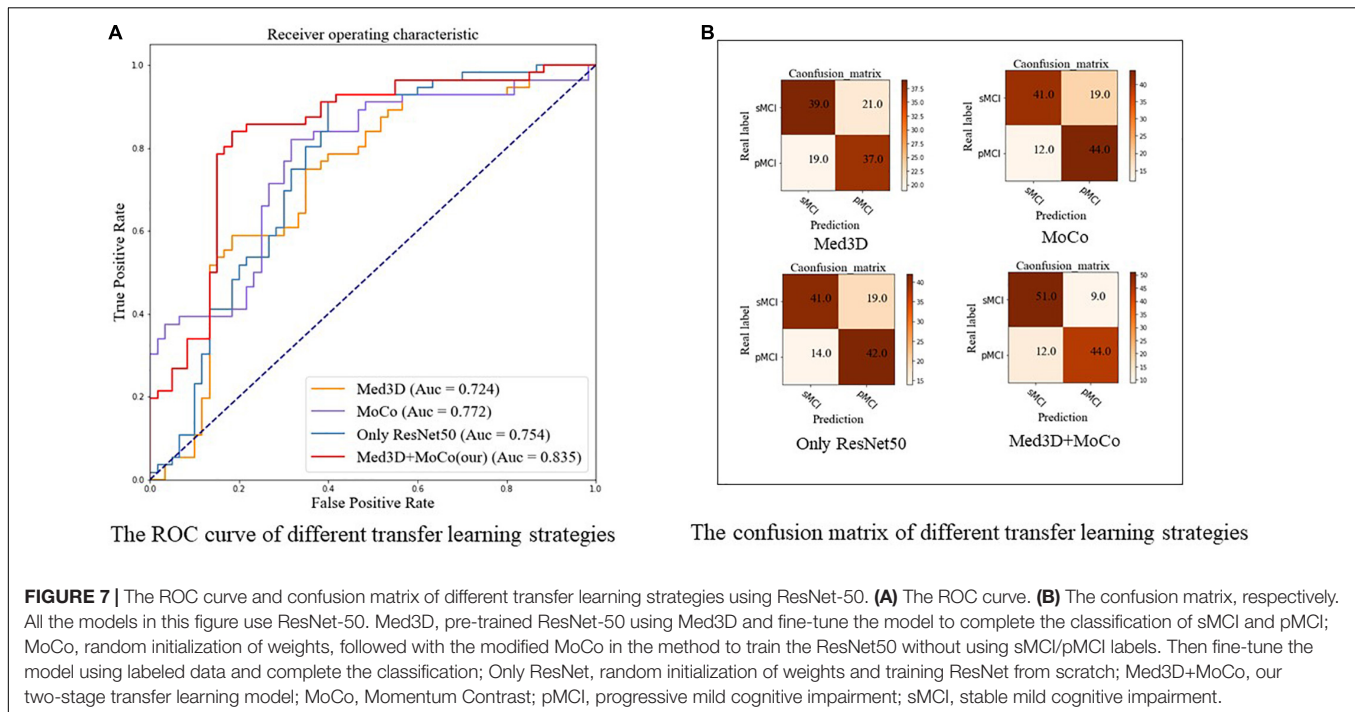
FIGURE 7 | The ROC curve and confusion matrix of different transfer learning strategies using ResNet-50. **(A)** The ROC curve. **(B)** The confusion matrix, respectively. All the models in this figure use ResNet-50. Med3D, pre-trained ResNet-50 using Med3D and fine-tune the model to complete the classification of sMCI and pMCI; MoCo, random initialization of weights, followed with the modified MoCo in the method to train the ResNet50 without using sMCI/pMCI labels. Then fine-tune the model using labeled data and complete the classification; Only ResNet, random initialization of weights and training ResNet from scratch; Med3D+MoCo, our two-stage transfer learning model; MoCo, Momentum Contrast; pMCI, progressive mild cognitive impairment; sMCI, stable mild cognitive impairment.

outperformed both transfer learning and contrastive learning individually and achieved better or comparable results than previous state-of-the-art studies.

Several factors improve the performance of our classification model. The first contribution is the proposal of a two-stage model. **Table 3** shows that the classification accuracy of ResNet-50 trained from scratch on sMCI and pMCI is 6.03% higher than that of ResNet-50 pre-trained in Med3D, implying that direct transfer learning for two data sets that are not highly correlated does not always achieve good results, and may even result in a negative transfer. The performance of transfer learning is affected by the various factors such as the size of pretrained samples, the relevance of the source and target domains. Thus, not all the transfer learning can improve the model's performance (Huh et al., 2016; Zhuang X. et al., 2019; Alzubaidi et al., 2020, 2021; Mustafa et al., 2021). For example, Alzubaidi et al. (2021) found that the model trained from scratch performed better than those pretrained by ImageNet using three different medical imaging datasets. This observation inspired us to develop a two-stage model. Our two-stage model is sample efficient when compared with existing transfer learning-based models for sMCI and pMCI classification (Oh et al., 2019; Gao et al., 2020; Bae et al., 2021).

In each of these studies, additional AD and NC samples were collected for training. But our model does not require additional data collection, makes full use of each sample, and produces better or equivalent results. For example, Bae et al. (2021) developed a transfer learning model based on 3D ResNet29. In the source task, the model is pretrained using MRI scans of 2,084 normal samples and 1,406 AD samples. Then they used pMCI and sMCI samples to fine-tune the model to accomplish the target task of classifying pMCI and sMCI. In comparison to our results, they got the same accuracy but lower AUC.

To the best of our knowledge, we are the first to use the MoCo pretrained model for sMCI and pMCI classification. Compared with the models, only pretrained by Med3D and ResNet-50 trained from scratch, our method improved accuracy by 16.4 and 10.3%, respectively, further demonstrating the importance and efficiency of including contrastive learning into our method. Pretraining by contrastive learning allows the model to have a feature representation with better generalization at the same domain of the target task (Sun et al., 2019). Recent studies have shown that fine-tuning on a well-trained contrastive learning model can achieve comparable or even better results than fully supervised learning (Wu et al., 2018; Zhuang C. et al., 2019),
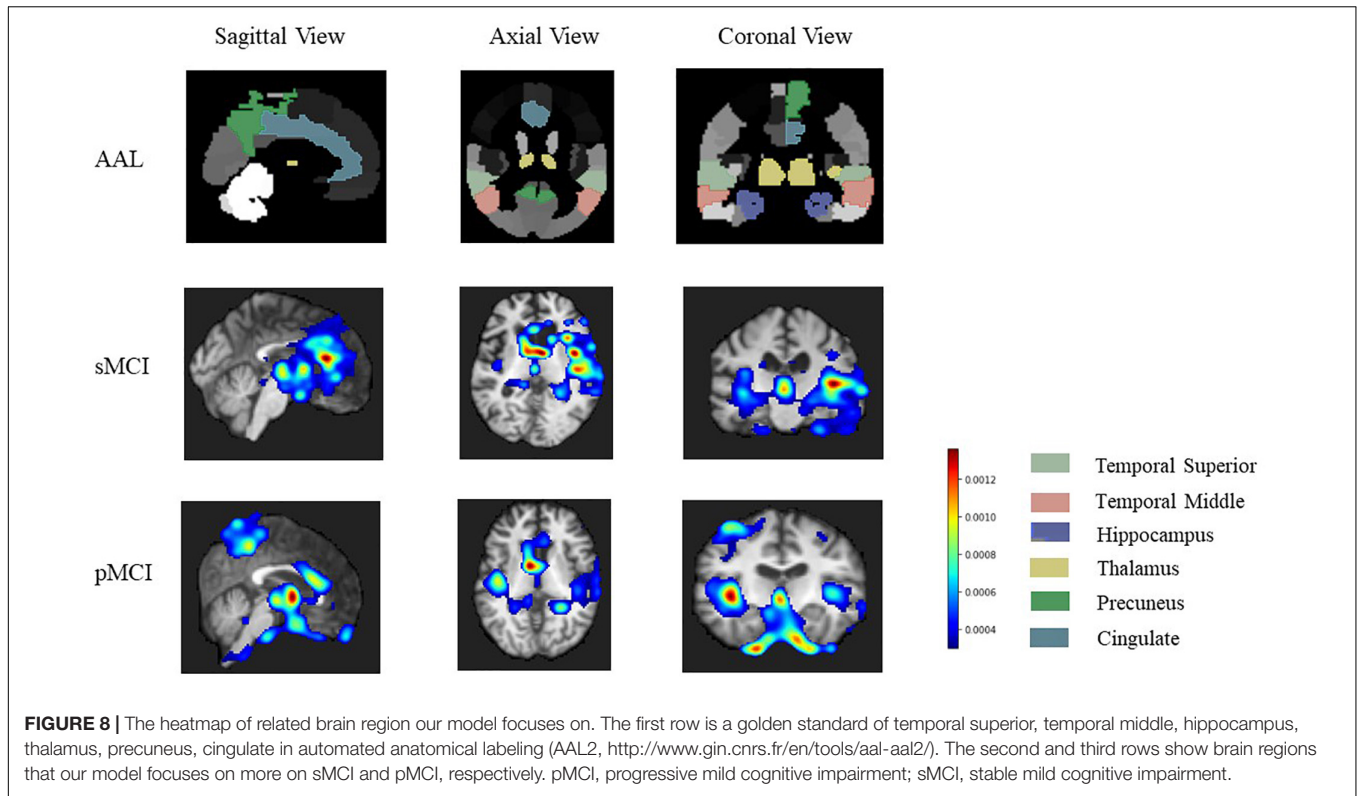
FIGURE 8 | The heatmap of related brain region our model focuses on. The first row is a golden standard of temporal superior, temporal middle, hippocampus, thalamus, precuneus, cingulate in automated anatomical labeling (AAL2, http://www.gin.cnrs.fr/en/tools/aal-aal2/). The second and third rows show brain regions that our model focuses on more on sMCI and pMCI, respectively. pMCI, progressive mild cognitive impairment; sMCI, stable mild cognitive impairment.

TABLE 4 | A summarized comparison of state-of-the-art research on MRI using deep learning in sMCI and pMCI classification.

Research	Conversion time	sMCI/pMCI number	Network	Acc	Sens	Spec	AUC
Liu M. et al. (2018)	36 Months	100/164	Landmark detection and 3D CNN	0.77	0.42	0.82	0.78
Liu J. et al. (2018)	18 Months	160/120	Whole brain hierarchical network	0.72	0.75	0.71	0.72
Lin et al. (2018)	36 Months	139/169	2D CNN	0.79	**0.86**	0.68	0.83
Shmulev and Belyaev (2018)	60 Months	532/327	3DResNet/VoxCNN	0.62	0.75	0.54	0.70
Shi et al. (2018)	18 Months	56/43	2D Deep polynomial network	0.79	0.68	0.87	0.80
Basaia et al. (2019)	36 months	253/510	3D CNN	0.75	0.75	0.75	NA
Li et al. (2019)	36 months	95/126	Self-weighting grading biomarker	0.69	0.82	0.51	0.70
Oh et al. (2019)	36 months	101/106	3D CNN + Transfer learning from CN/AD	0.73	0.77	0.71	0.79
Spasov et al. (2019)	36 months	228/181	3D CNN	0.72	0.63	0.81	0.79
Abrol et al. (2020)	24 Months	409/217	2D Multiscale Deep Neural Network	0.75	0.73	0.76	0.71
Gao et al. (2020)	36 months	129/168	3D CNN + Transfer learning for AD age prediction	0.81	0.76	0.77	0.76
Pan et al. (2020)	18 Months	173/105	CNN and ensemble learning	0.62	NA	NA	0.59
Wen et al. (2020a)	36 Months	298/295	3D CNN	0.74	0.80	0.68	NA
Bae et al. (2021)	36 Months	222/228	3D ResNet29 + Transfer learning from CN /AD	**0.82**	0.72	NA	0.83
Guan et al. (2021)	36 Months	401/197	3D CNN	0.79	0.55	0.84	0.78
Zhang J. et al. (2021)	18 Months	251/162	3D DenseNet + Attention	0.79	0.75	0.82	**0.86**
Our	36 months	297/280	3D ResNet and transfer learning from self	**0.82**	0.79	**0.85**	0.84

The bold numbers denote the maximum value of each column.

Acc, Accuracy; AD, Alzheimer's disease; AUC, Area Under Curve; F1, F1-score; NC, Normal control; pMCI, progressive mild cognitive impairment; Sens, Sensitivity; Spec, Specificity; sMCI, stable mild cognitive impairment.

which is consistent with our findings. In addition, one of the critical factors limiting the performance of contrastive learning is the slow convergence rate (Chen T. et al., 2020; Chen X. et al., 2020; Tian et al., 2020). As shown in **Table 3**, compared with MoCo trained from scratch, our method improved accuracy by 6.89%, which indicates that transfer learning can accelerate

the convergence of the MoCo model and improve the model performance. MoCo and transfer learning can reinforce and complement one other.

In addition, our model uses complete 3D MRI as model input. Unlike models using 2D slices, the 3D model makes full use of the spatial information of the brain to improve the accuracy of the

model (Wen et al., 2020b). Furthermore, some previous studies used feature engineering or cherry-picked regions of interest as input features (Gerardin et al., 2009; Ahmed et al., 2014; Basaia et al., 2019), which ignored the contributions of other features in the model, resulting in information loss in some cases. For example, Basaia et al. (2019) chose brain gray matter to train the model, neglecting the role of cerebrospinal fluid or white matter in early diagnosis of AD (Jack et al., 2010; Weiler et al., 2015). Our model differs from the previous studies by using an end-to-end model to learn from all possible features in medical images, which improves model performance.

In **Figure 8**, the hippocampus, temporal, and thalamus are highlighted in both sMCI and pMCI. Hippocampus and amygdala in the middle temporal lobe have been considered as crucial brain regions for the diagnosis of early AD (Visser et al., 2002; Braak and van Braak, 2004b; Burton et al., 2009; Costafreda et al., 2011). The hippocampus is essential for memory formation, and the recent studies have found that the hippocampus atrophy of pMCI is more pronounced than sMCI (Devanand et al., 2007; Risacher et al., 2009; Costafreda et al., 2011). Similarly, the amygdala, which is primarily responsible for emotion and expression, is intimately linked to emotional changes of AD, such as anxiety and irritability (Unger et al., 1991; Poulin et al., 2011). Thalamic damage is associated with decreased body movement and coordination, attention, and awareness in AD (Braak and van Braak, 2004a; de Jong et al., 2008; Cho et al., 2014; Aggleton et al., 2016). In addition to the hippocampus and temporal, which have been widely studied in AD, our heatmap also reveals that pMCI is also closely related to the precuneus with high-level memory and cognitive functions, which is in line with the previous studies (Whitwell et al., 2008; Bailly et al., 2015; Perez et al., 2015; Colangeli et al., 2016; Kato et al., 2016; Zhang H. et al., 2021). The results, as mentioned earlier, further indicate that some structural brain region abnormalities play an important role in predicting early AD. In summary, our discovery of important brain regions is supported by abundant literature, which helps construct a more comprehensive brain biomarker atlas to predict MCI progression.

CONCLUSION

In conclusion, our two-stage model increases both the accuracy of early AD detection as well as the transparency of the model. Notably, a comprehensive comparison of different 3D ResNet networks provides references for related research. Furthermore, the combination of transfer learning and contrastive learning solves the negative transfer problem and alleviates the model overfitting problem due to a lack of medical data. Notably, it also substantially improves the diagnostic performance of this tricky classification problem in neuroscience. Our model only uses low-invasive, low-cost, and widely available MRI data, which significantly expands the application scenarios of the model.

However, this study also has some limitations that merit additional exploration. First of all, we will explore more options for the model's various modules, such as different data augmentation methods and pretrained models on model effectiveness. When a larger dataset becomes available, we will also continue to validate our model. At final, it is worth noting that a direct comparison of different methods using the same evaluation metrics is straightforward but may not be the optimal solution. Factors such as sample size, dataset split strategy, sMCI, and pMCI definitions, and test data selection can have an impact on model outcomes. A more statistically robust comparison should be proposed in our future studies. Despite these limitations, our model provides a new solution to avoid overfitting because of the insufficient medical data and allows early identification of AD.

AUTHOR'S NOTE

The numerical calculations in this article have been done on the supercomputing system in the Supercomputing Center of Wuhan University.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

PL and LL designed the study and wrote the article. PL implemented the algorithm and preprocessed the data. LH, NZ, HL, and TT provided critical suggestions. All authors contributed to the article and approved the submitted version.

FUNDING

ACKNOWLEDGMENTS

REFERENCES

Abrol, A., Bhattarai, M., Fedorov, A., Du, Y., Plis, S., and Calhoun, V. (2020). Deep residual learning for neuroimaging: an application to predict progression to Alzheimer's disease. *J. Neurosci. Methods* 339:108701. doi: 10.1016/j.jneumeth.2020.108701

Abrol, A., Fu, Z., Salman, M. S., Silva, R. F., Du, Y., Plis, S., et al. (2021). Deep learning encodes robust discriminative neuroimaging representations to outperform standard machine learning. *Nat. Commun.* 12:353. doi: 10.1038/s41467-020-20655-6

Afzal, S., Maqsood, M., Nazir, F., Khan, U., Aadil, F., Awan, K. M., et al. (2019). A Data Augmentation-Based Framework to Handle Class Imbalance Problem for Alzheimer's Stage Detection. *IEEE Access* 7, 115528–115539. doi: 10.1109/access.2019.2932786

Aggleton, J. P., Pralus, A., Nelson, A. J. D., and Hornberger, M. (2016). Thalamic pathology and memory loss in early Alzheimer's disease: moving the focus from the medial temporal lobe to Papez circuit. *Brain* 139, 1877–1890. doi: 10.1093/brain/aww083

Ahmed, O., Ben, Benois-Pineau, J., Allard, M., Amar, C., Ben, et al. (2014). Classification of Alzheimer_s disease subjects from MRI using hippocampal visual features. *Multimed. Tools Appl.* 74, 1249–1266. doi: 10.1007/s11042-014-2123-y

Alzubaidi, L., Duan, Y., Al-dujaili, A., Ibraheem, I. K., Alkenani, A. H., Santamaría, J., et al. (2021). Deepening into the suitability of using pre-trained models of ImageNet against a lightweight convolutional neural network in medical imaging: an experimental study. *PeerJ Comput. Sci.* 7:e715. doi: 10.7717/peerj-cs.715

Alzubaidi, L., Fadhel, M. A., Al-Shamma, O., Zhang, J., Santamaría, J., Duan, Y., et al. (2020). Towards a Better Understanding of Transfer Learning for Medical Imaging: a Case Study. *Appl. Sci.* 10:4523. doi: 10.1016/j.morpho.2019.09.001

Association, A. (2019). 2019 Alzheimer's disease facts and figures. *Alzheimer's & Dement.* 15, 321–387.

Bae, J., Stocks, J., Heywood, A., Jung, Y., Jenkins, L. M., Hill, V. B., et al. (2021). Transfer learning for predicting conversion from mild cognitive impairment to dementia of Alzheimer's type based on a three-dimensional convolutional neural network. *Neurobiol. Aging* 99, 53–64. doi: 10.1016/j.neurobiolaging.2020.12.005

Bailly, M., Destrieux, C., Hommet, C., Mondon, K., Cottier, J. P., Beaufils, É, et al. (2015). Precuneus and Cingulate Cortex Atrophy and Hypometabolism in Patients with Alzheimer's Disease and Mild Cognitive Impairment: MRI and 18F-FDG PET Quantitative Analysis Using FreeSurfer. *Biomed Res. Int* 2015:583931. doi: 10.1155/2015/583931

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* 21:101645. doi: 10.1016/j.nicl.2018.101645

Bernal, J., Kushibar, K., Asfaw, D. S., Valverde, S., Oliver, A., Marti, R., et al. (2019). Deep convolutional neural networks for brain image analysis on magnetic resonance imaging: a review. *Artif. Intell. Med.* 95, 64–81. doi: 10.1016/j.artmed.2018.08.008

Braak, H., and van Braak, E. (2004b). Neuropathological stageing of Alzheimer-related changes. *Acta Neuropathol.* 82, 239–259. doi: 10.1007/BF00308809

Braak, H., and van Braak, E. (2004a). Alzheimer's disease affects limbic nuclei of the thalamus. *Acta Neuropathol.* 81, 261–268. doi: 10.1007/BF00305867

Burton, E. J., Barber, R., Mukaetova-Ladinska, E. B., Robson, J. I., Perry, R. H., Jaros, E., et al. (2009). Medial temporal lobe atrophy on MRI differentiates Alzheimer's disease from dementia with Lewy bodies and vascular cognitive impairment: a prospective study with pathological verification of diagnosis. *Brain* 132, 195–203. doi: 10.1093/brain/awn298

Carreira, J., and Zisserman, A. (2017). *Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. 2017 IEEE Conf. Comput. Vis. Pattern Recognit.* Piscataway: IEEE, 4724–4733.

Chen, S., Ma, K., and Zheng, Y. (2019). Med3D: transfer learning for 3D medical image analysis. *arXiv* [Preprint]. arXiv:1904.00625

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. E. (2020). *A Simple Framework for Contrastive Learning of Visual Representations. ArXiv abs/2002.0.*

Chen, X., Fan, H., Girshick, R. B., and He, K. (2020). Improved baselines with momentum contrastive learning. *arXiv* [Preprint]. arXiv:2003.04297

Cho, H., Kim, J., Kim, C., Ye, B. S., Kim, H. J., Yoon, C. W., et al. (2014). Shape changes of the basal ganglia and thalamus in Alzheimer's disease: a three-year longitudinal study. *J. Alzheimers. Dis.* 40, 285–295. doi: 10.3233/JAD-132072

Choi, H., and Jin, K. (2018). Predicting cognitive decline with deep learning of brain metabolism and amyloid imaging. *Behav. Brain Res.* 344, 103–109. doi: 10.1016/j.bbr.2018.02.017

CodaLab. (2021). *Competition*. Available online at: https://competitions.codalab.org/competitions/17094#results (Accessed October 1, 2021).

Colangeli, S., Boccia, M., Verde, P. A., La, Guariglia, P., Bianchini, F., et al. (2016). Cognitive Reserve in Healthy Aging and Alzheimer's Disease. *Am. J. Alzheimer's Dis. Other Dementias* 31, 443–449.

Costafreda, S. G., Dinov, I. D., Tu, Z., Shi, Y., Liu, C.-Y., Kloszewska, I., et al. (2011). Automated hippocampal shape analysis predicts the onset of dementia in mild cognitive impairment. *Neuroimage* 56, 212–219. doi: 10.1016/j.neuroimage.2011.01.050

de Jong, L. W., van der Hiele, K., Veer, I. M., Houwing, J. J., Westendorp, R. G. J., Bollen, E. L. E. M., et al. (2008). Strongly reduced volumes of putamen and thalamus in Alzheimer's disease: an MRI study. *Brain* 131, 3277–3285. doi: 10.1093/brain/awn278

Devanand, D. P., Pradhaban, G., Liu, X., Khandji, A. G., de Santi, S., Segal, S., et al. (2007). Hippocampal and entorhinal atrophy in mild cognitive impairment. *Neurology* 68, 828–836. doi: 10.1212/01.wnl.0000256697.20968.d7

Efron, B., and Tibshirani, R. (1993). *An Introduction to the Bootstrap*. New York: Chapman and Hall/CRC.

Gao, F., Yoon, H., Xu, Y., Goradia, D., Luo, J., Wu, T., et al. (2020). AD-NET: age-adjust neural network for improved MCI to AD conversion prediction. *Neuroimage* 27:102290. doi: 10.1016/j.nicl.2020.102290

Gerardin, E., Chetelat, G., Chupin, M., Cuingnet, R., Desgranges, B., Kim, H. S., et al. (2009). Multidimensional classification of hippocampal shape features discriminates Alzheimer's disease and mild cognitive impairment from normal aging. *Neuroimage* 47, 1476–1486. doi: 10.1016/j.neuroimage.2009.05.036

Guan, H., Wang, C., Cheng, J., Jing, J., and Liu, T. (2021). A parallel attention_augmented bilinear network for early magnetic resonance imaging-based diagnosis of Alzheimer's disease. *Hum. Brain Mapp.* 43, 760–772. doi: 10.1002/hbm.25685

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. B. (2020). "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, 9726–9735.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition," in *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Piscataway, NJ: IEEE), 770–778.

Heun, R., Mazanek, M., Atzor, K.-R., Tintera, J., Gawehn, J., Burkart, M., et al. (1997). Amygdala-Hippocampal Atrophy and Memory Performance in Dementia of Alzheimer Type. *Dement. Geriatr. Cogn. Disord.* 8, 329–336. doi: 10.1159/000106651

Huh, M., Agrawal, P., and Efros, A. A. (2016). What makes ImageNet good for transfer learning? *arXiv* [Preprint]. arXiv:1608.08614

Irvin, J. A., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., et al. (2019). *CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison*. California: AAAI Press.

Jack, C. R., Bernstein, M. A., Fox, N. C., Thompson, P., Alexander, G., Harvey, D., et al. (2008). The Alzheimer's Disease Neuroimaging Initiative (ADNI): MRI methods. *J. Magn. Reson. IMAGING* 27, 685–691. doi: 10.1002/jmri.21049

Jack, C. R., Knopman, D. S., Jagust, W. J., Shaw, L. M., Aisen, P. S., Weiner, M. W., et al. (2010). Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade. *Lancet Neurol.* 9, 119–128. doi: 10.1016/S1474-4422(09)70299-6

Johnson, K. A., Fox, N. C., Sperling, R. A., and Klunk, W. E. (2012). Brain imaging in Alzheimer disease. *Cold Spring Harb. Perspect. Med.* 2:a006213.

Kato, T., Inui, Y., Nakamura, A., and Ito, K. (2016). Brain fluorodeoxyglucose (FDG) PET in dementia. *Ageing Res. Rev.* 30, 73–84. doi: 10.1016/j.arr.2016.02.003

Klöppel, S., Stonnington, C. M., Chu, C., Draganski, B., Scahill, R. I., Rohrer, J. D., et al. (2008). Automatic classification of MR scans in Alzheimer's disease. *Brain* 131), 681–689. doi: 10.1093/brain/awm319

Li, Y., Fang, Y., Zhang, H., and Hu, B. (2019). Self-Weighting Grading Biomarker Based on Graph-Guided Information Propagation for the Prediction of Mild Cognitive Impairment Conversion. *IEEE Access* 7, 116632–116642.

Lin, W. M., Tong, T., Gao, Q. Q., Guo, D., Du, X. F., Yang, Y. G., et al. (2018). Convolutional Neural Networks-Based MRI Image Analysis for the Alzheimer's Disease Prediction From Mild Cognitive Impairment. *Front. Neurosci* 12:777. doi: 10.3389/fnins.2018.00777

Liu, J., Li, M., Lan, W., Wu, F.-X., Pan, Y., and Wang, J. (2018). Classification of Alzheimer's Disease Using Whole Brain Hierarchical Network. *IEEE/ACM Trans. Comput. Biol. Bioinforma.* 15, 624–632. doi: 10.1109/TCBB.2016.2635144

Liu, M., Zhang, J., Adeli, E., and Shen, D. (2018). Landmark-based deep multi-instance learning for brain disease diagnosis. *Med. Image Anal* 43, 157–168. doi: 10.1016/j.media.2017.10.005

Liu, X., Zhang, F., Hou, Z., Wang, Z., Mian, L., Zhang, J., et al. (2020). Self-supervised learning: generative or contrastive. *arXiv* [Preprint]. arXiv:2006.08218

Liu, Y., Li, Z., Ge, Q., Lin, N., and Xiong, M. (2019). Deep Feature Selection and Causal Analysis of Alzheimer's Disease. *Front. Neurosci.* 13:1198. doi: 10.3389/fnins.2019.01198

Livingston, G., Sommerlad, A., Orgeta, V., Costafreda, S. G., Huntley, J., Ames, D., et al. (2017). Dementia prevention, intervention, and care. *Lancet* 390, 2673–2734.

Medical Segmentation Decathlon. (2021). *Medical Segmenation Decathlon.* Available online at: http://medicaldecathlon.com/index.html (Accessed October 1, 2021)

Menze, B. H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J. S., et al. (2015). The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). *IEEE Trans. Med. Imaging* 34, 1993–2024.

Misra, I., and van der Maaten, L. (2020). "Self-supervised learning of pretext-invariant representations," in *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, 6706–6716. doi: 10.1109/CVPR42600.2020.00674

Mousavian, M., Chen, J., and Greening, S. G. (2019). "Depression detection using feature extraction and deep learning from sMRI images," in *Proceedings of the 2019 18th IEEE International Conference On Machine Learning and Applications (ICMLA)*, Boca Raton, FL, 1731–1736.

Mustafa, B., Loh, A., von Freyberg, J., MacWilliams, P., Wilson, M., McKinney, S. M., et al. (2021). Supervised transfer learning at scale for medical imaging. *arXiv* [Preprint]. arXiv:2101.05913

Naz, S., Ashraf, A., and Zaib, A. (2021). Transfer learning using freeze features for Alzheimer neurological disorder detection using ADNI dataset. *Multimed. Syst.* 28, 85–94.

Oh, K., Chung, Y.-C., Kim, K. W., Kim, W.-S., and Oh, I.-S. (2019). Classification and Visualization of Alzheimer's Disease using Volumetric Convolutional Neural Network and Transfer Learning. *Sci. Rep.* 9, 18150.

Pan, D., Zeng, A., Jia, L., Huang, Y., Frizzell, T., and Song, X. (2020). Early Detection of Alzheimer's Disease Using Magnetic Resonance Imaging: a Novel Approach Combining Convolutional Neural Networks and Ensemble Learning. *Front. Neurosci.* 14:259. doi: 10.3389/fnins.2020.00259

Perez, S. E., He, B., Nadeem, M., Wuu, J., Scheff, S. W., Abrahamson, E. E., et al. (2015). Resilience of Precuneus Neurotrophic Signaling Pathways Despite Amyloid Pathology in Prodromal Alzheimer's Disease. *Biol. Psychiatr.* 77, 693–703. doi: 10.1016/j.biopsych.2013.12.016

Petersen, R. C. (2000). Mild cognitive impairment: transition between aging and Alzheimer's disease. *Neurologia* 15, 93–101.

Poulin, S., Dautoff, R., Morris, J. C., Barrett, L. F., and Dickerson, B. C. (2011). Amygdala atrophy is prominent in early Alzheimer's disease and relates to symptom severity. *Psychiatry Res. Neuroimaging* 194, 7–13. doi: 10.1016/j.pscychresns.2011.06.014

Raghu, M., Zhang, C., Kleinberg, J., and Bengio, S. (2019). "Transfusion: understanding transfer learning for medical imaging," in *Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*, Vancouver, BC.

Risacher, S. L., Saykin, A. J., West, J. D., Shen, L., Firpi, H. A., and McDonald, B. C. (2009). Baseline MRI Predictors of Conversion from MCI to Probable AD in the ADNI Cohort. *Curr. Alzheimer Res.* 6, 347–361. doi: 10.2174/156720509788929273

Roberson, E. D., and Mucke, L. (2006). 100 Years and Counting: prospects for Defeating Alzheimer's Disease. *Science (80-.)*. 314, 781–784. doi: 10.1126/science.1132813

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., et al. (2015). ImageNet Large Scale Visual Recognition Challenge. *Int. J. Comput. Vis.* 115, 211–252. doi: 10.1007/s11263-015-0816-y

Shen, D. G., Wu, G. R., and Suk, H. I. (2017). "Deep Learning in Medical Image Analysis,". 19, 221–248. doi: 10.1146/annurev-bioeng-071516-044442

Shi, J., Zheng, X., Li, Y., Zhang, Q., and Ying, S. H. (2018). Multimodal Neuroimaging Feature Learning With Multimodal Stacked Deep Polynomial Networks for Diagnosis of Alzheimer's Disease. *IEEE J. Biomed. Heal. INFORMATICS* 22, 173–183. doi: 10.1109/JBHI.2017.2655720

Shmulev, Y., and Belyaev, M. (2018). "Predicting conversion of mild cognitive impairments to Alzheimer's disease and exploring impact of neuroimaging," in *Graphs in Biomedical Image Analysis and Integrating Medical Imaging and Non-Imaging Modalities. GRAIL 2018, Beyond MIC 2018. Lecture Notes in Computer Science*, Vol. 11044, eds D. Stoyanov, et al. (Cham: Springer). doi: 10.1007/978-3-030-00689-1_9

Spasov, S. E., Passamonti, L., Duggento, A., Liu, P., and Toschi, N. (2019). A parameter-efficient deep learning approach to predict conversion from mild cognitive impairment to Alzheimer's disease. *Neuroimage* 189, 276–287. doi: 10.1016/j.neuroimage.2019.01.031

Sun, Y., Tzeng, E., Darrell, T., and Efros, A. A. (2019). Unsupervised domain adaptation through self-supervision. *arXiv* [Preprint]. arXiv:1909.11825

Tajbakhsh, N., Shin, J. Y., Gurudu, S., Hurst, R. T., Kendall, C. B., Gotway, M., et al. (2016). Convolutional Neural Networks for Medical Image Analysis: full Training or Fine Tuning? *IEEE Trans. Med. Imaging* 35, 1299–1312. doi: 10.1109/TMI.2016.2535302

Tian, Y., Sun, C., Poole, B., Krishnan, D., Schmid, C., and Isola, P. (2020). What makes for good views for contrastive learning. *arXiv* [Preprint]. arXiv:2005.10243

Tobon-Gomez, C., Geers, A. J., Peters, J., Weese, J., Pinto, K., Karim, R., et al. (2015). Benchmark for Algorithms Segmenting the Left Atrium From 3D CT and MRI Datasets. *IEEE Trans. Med. Imaging* 34, 1460–1473. doi: 10.1109/TMI.2015.2398818

Tran, D., Bourdev, L. D., Fergus, R., Torresani, L., and Paluri, M. (2014). *C3D: Generic Features for Video Analysis. ArXiv abs/1412.0*.

Tustison, N., Avants, B., Cook, P., Zheng, Y., Egan, A., Yushkevich, P., et al. (2010). N4ITK: improved N3 Bias Correction. *IEEE Trans. Med. Imaging* 29, 1310–1320. doi: 10.1109/TMI.2010.2046908

Unger, J. W., Lapham, L. W., McNeill, T. H., Eskin, T. A., and Hamill, R. W. (1991). The amygdala in Alzheimer's disease: neuropathology and Alz 50 Immunoreactivity. *Neurobiol. Aging* 12, 389–399. doi: 10.1016/0197-4580(91)90063-p

Visser, P. J., Verhey, F. R. J., Hofman, P. A. M., Scheltens, P., and Jolles, J. (2002). Medial temporal lobe atrophy predicts Alzheimer's disease in patients with minor cognitive impairment. *J. Neurol. Neurosurg. Psychiatr.* 72, 491–497. doi: 10.1136/jnnp.72.4.491

Weiler, M., Agosta, F., Canu, E., Copetti, M., Magnani, G., Marcone, A., et al. (2015). Following the Spreading of Brain Structural Changes in Alzheimer's Disease: a Longitudinal. *Multimodal MRI Study. J. Alzheimers. Dis.* 47, 995–1007. doi: 10.3233/JAD-150196

Wen, J., Thibeau-Sutre, E., Samper-González, J., Routier, A., Bottani, S., Durrleman, S., et al. (2020b). Convolutional Neural Networks for Classification of Alzheimer's Disease: overview and Reproducible Evaluation. *Med. Image Anal.* 63:101694. doi: 10.1016/j.media.2020.101694

Wen, J., Thibeau-Sutre, E., Diaz-Melo, M., Samper-González, J., Routier, A., Bottani, S., et al. (2020a). Convolutional neural networks for classification of Alzheimer's disease: overview and reproducible evaluation. *Med. Image Anal.* 63:101694. doi: 10.1016/J.MEDIA.2020.101694

Whitwell, J. L., Shiung, M. M., Przybelski, S. A., Weigand, S. D., Knopman, D. S., Boeve, B. F., et al. (2008). MRI patterns of atrophy associated with progression to AD in amnestic mild cognitive impairment. *Neurology* 70, 512–520. doi: 10.1212/01.wnl.0000280575.77437.a2

Wu, Z., Xiong, Y., Yu, S. X., and Lin, D. (2018). "Unsupervised feature learning via non-parametric instance discrimination," in *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, 3733–3742. doi: 10.1109/CVPR.2018.00393

Yang, X., Bian, C., Yu, L., Ni, D., and Heng, P. (2017). "Hybrid Loss Guided Convolutional Networks for Whole Heart Parsing," in *Statistical Atlases and Computational Models of the Heart. ACDC and MMWHS Challenges*, eds M. Pop, et al. (Cham: Springer).

Zeng, G., and Zheng, G. (2018). Deep Learning-Based Automatic Segmentation of the Proximal Femur from MR Images. *Adv. Exp. Med. Biol.* 1093, 73–79. doi: 10.1007/978-981-13-1396-7_6

Zhang, H., Wang, Y., Lyu, D., Li, Y., Li, W., Wang, Q., et al. (2021). Cerebral blood flow in mild cognitive impairment and Alzheimer's disease: a systematic review and meta-analysis. *Ageing Res. Rev* 71:101450. doi: 10.1016/j.arr.2021.101450

Zhang, J., Zheng, B., Gao, A., Feng, X., Liang, D.-P., and Long, X. (2021). A 3D densely connected convolution neural network with connection-wise attention mechanism for Alzheimer's disease classification. *Magn. Reson. Imaging* 78, 119–126. doi: 10.1016/j.mri.2021.02.001

Zhuang, C., Zhai, A., and Yamins, D. (2019). "Local aggregation for unsupervised learning of visual embeddings," in *Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, 6001–6011. doi: 10.1109/ICCV.2019.00610

Zhuang, X., Li, Y., Hu, Y., Ma, K., Yang, Y., and Zheng, Y. (2019). "Self-supervised Feature Learning for 3D Medical Images by Playing a Rubik's Cube," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019. MICCAI 2019*, eds D. Shen, et al. (Cham: Springer). doi: 10.1016/j.media.2020.101746

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.