A 3D Decoupling Alzheimer's Disease Prediction Network Based on Structural MRI

Shicheng Wei[1], Wencheng Yang[1,*], Eugene Wang[2,3], Song Wang[4], Yan Li[1,*]

[1]School of Mathematics and Computing, University of Southern Queensland, 487-535 West Street, Toowoomba, 4350, Queensland, Australia.
[2]Personalised Oncology Division, The Walter and Eliza Hall Institute of Medical Research, Parkville, VIC, Australia.
[3]Faculty of Medicine, Nursing and Health Sciences, Monash University, Melbourne, VIC, Australia.
[4]Dept of Engineering, La Trobe University, Bundoora, 3086, VIC, Australia.
[*]Corresponding Author.

Contributing authors: shicheng.wei@unisq.edu.au; wencheng.yang@unisq.edu.au; wang.e@wehi.edu.au; song.wang@latrobe.edu.au; yan.li@unisq.edu.au;

Abstract

Purpose: This paper aims to develop a three-dimensional (3D) Alzheimer's disease (AD) prediction method, thereby bettering current predictive methods, which struggle to fully harness the potential of structural magnetic resonance imaging (sMRI) data.

Methods: Traditional convolutional neural networks encounter pressing difficulties in accurately focusing on the AD lesion structure. To address this issue, a 3D decoupling, self-attention network for AD prediction is proposed. Firstly, a multi-scale decoupling block is designed to enhance the network's ability to extract fine-grained features by segregating convolutional channels. Subsequently, a self-attention block is constructed to extract and adaptively fuse features from three directions (sagittal, coronal and axial), so that more attention is geared towards brain lesion areas. Finally, a clustering loss function is introduced and combined with the cross-entropy loss to form a joint loss function for enhancing the network's ability to discriminate between different sample types.

Results: The accuracy of our model is 0.985 for the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset and 0.963 for the Australian Imaging, Biomarker & Lifestyle (AIBL) dataset, both of which are higher than the classification accuracy of similar tasks in this category. This demonstrates that our model can accurately distinguish between normal control (NC) and Alzheimer's Disease (AD), as well as between stable mild cognitive impairment (sMCI) and progressive mild cognitive impairment (pMCI).

Conclusion: The proposed AD prediction network exhibits competitive performance when compared with state-of-the-art methods. The proposed model successfully addresses the challenges of dealing with 3D sMRI image data and the limitations stemming from inadequate information in 2D sections, advancing the utility of predictive methods for AD diagnosis and treatment.

Keywords: Alzheimer's Disease, MRI, Convolutional Neural Network, CNN, Deep Learning

1 Introduction

Alzheimer's disease (AD) is a progressive irreversible neurodegenerative disorder [1]. Initial symptoms include short-term memory loss, and as the condition advances, individuals may experience language difficulties, confusion, and various behavioral issues. Over time, a patient's physical abilities decline, ultimately resulting in death. In contrast to a normal control (NC) population, patients with cognitive impairment can be broadly categorized into different stages: stable mild cognitive impairment (sMCI), progressive mild cognitive impairment (pMCI), and AD. Currently, there are no effective therapies or treatments to halt or reverse the progression of the disease. However, if AD is diagnosed at an early stage, doctors can intervene to slow down its progression. Nowadays, many technological methods are widely applied to assist in the clinical diagnosis of complex diseases using magnetic resonance imaging (fMRI), structural magnetic resonance imaging (sMRI), electroencephalogram (EEG) [2–4], polyethylene terephthalate (PET)

and so on. Studies [5–9] show that structural magnetic resonance imaging (sMRI) can be a powerful tool for clinicians to clarify the diagnosis for patients with suspected sMCI, pMCI, and AD, as sMRI data can provide vital information regarding structural changes in the brain that can be correlated with clinical findings. Various deep learning [10, 11] methods were employed for predicting AD using sMRI data [12–15]. The utilization of sMRI data is generally divided into two-dimensional (2D) and three-dimensional (3D) forms. Hoang et al. [16] proposed to employ a vision transformer to extract feature maps highlighting specific features in the data from three 2D sagittal slices, enabling the classification of mild cognitive impairment (MCI) and AD. Xing et al. [17] introduced a technique that combines the fusion attention and a residual network which extracts feature information from 2D sMRI data using the residual network, leading to effective classification outcomes.

However, contextual information along the depth dimension is lost when 3D images are sliced into 2D. To address this issue, researchers explored the utilization of entire 3D sMRI data. Zhang et al. [18] incorporated a self-attention mechanism and a residual learning method into a 3D convolutional neural network (CNN). This method makes use of both global and local information to prevent the loss of crucial contextual information. Bakkouri et al. [19] improved traditional 3D convolution using multiple scales. This approach allows for the extraction of brain atrophy features across various scales with convolution blocks of different sizes, ultimately enhancing the overall prediction effectiveness. Chen et al. [20] proposed a multi-view slicing attention mechanism integrated into a 3D CNN. 2D slicing is employed to remove redundant information, followed by the use of a 3D network for the feature extraction, thus mitigating the risk of overfitting. Liu et al. [21] leveraged 3D sMRI data to extract feature maps from three distinct angles. Using multi-scale convolutions to extract features, which are later combined in the channel dimension, the proposed network is able to acquire more information from a pathological perspective.

Most existing 3D-based methods either overlook the convolution of 3D data in different directions or struggle to encompass feature maps acquired from all three dimensions adequately. As a result, these methods are unable to exploit the intrinsic feature information embedded within the 3D sMRI data. To address these issues, this study, serving as a substantial extension of our previous work [22], develops a backbone network that excels at extracting focal information from 3D sMRI scans across axial, sagittal, and coronal directions. Specifically, we design a multi-scale decoupling (MSD) block to isolate information from different groups so that a complete fusion of local information is achieved. Moreover, we introduce a self-attention (SA) block to extract feature maps from sagittal, coronal, and axial planes. Furthermore, we integrate the clustering loss with the cross-entropy loss to form a joint loss function, giving rise to enhanced prediction performance.

The key contributions of this work are summarised as follows.

1. Augmentation of information extraction capacity: The designed MSD block augments the ability of the proposed method to extract detailed information through a comprehensive integration of local features.
2. Improved image analysis and AD diagnosis: By simultaneously taking into account the features from all three directions (i.e., sagittal, coronal and axial), the SA block effectively directs attention towards critical atrophic lesions, thereby improving the overall ability of the model to analyze structural sMRI images.
3. Enhancement of identification outcomes: The joint loss function strengthens grouping and clustering within the same category, thereby enhancing the identification of specific brain regions or patterns that are highly indicative of AD.
4. Formal analysis: Beyond examining the overall performance of our model, we delve into its performance on individual subject basis. This analysis, focusing on variations in model efficacy across different subjects, is aimed at uncovering characteristics specific to Alzheimer's disease and understanding their influence on the model's predictions. Consequently, our approach offers fresh insights that could be instrumental for future studies in related fields.

This paper is organized as follows. The Introduction section is followed by the Methodology section, which includes an outline of the datasets used in the experiments and a detailed description of the proposed method. Section 3 presents the experiment setup and results. Section 4 discusses and analyzes the proposed method and provides visual representations of the feature maps extracted by our model. The concluding remarks are made in Section 5.

2 Methodology

2.1 Datasets

Large datasets from Australian Imaging, Biomarker & Lifestyle (AIBL) and Alzheimer's Disease Neuroimaging Initiative (ADNI), namely ADNI-1 and ADNI-2 [23], which are publicly available, are used in our experiments. For the ADNI dataset, data collection for ADNI-1 started in 2004 and concluded in 2009, while the data collection of ADNI-2 commenced in 2011 and continued until 2016. ADNI-2 is considered a continuation or extension of ADNI-1. Given that databases ADNI-1 and ADNI-2 have issues such as duplication, missing data, and poor data quality, a selection process is implemented to exclude such records. As a result, the selected dataset comprises 170 AD, 156 pMCI, 202 sMCI, and 206 NC records from ADNI-1, plus 102 AD, 101 pMCI, 329 sMCI, and 147 NC records from ADNI-2. The AIBL dataset contains

Table 1: Demographic information of subjects included in this study.

Dataset	Category	Gender (Male/Female)	Patient Age (Mean±SD[1])	Education[2] (Mean±SD)	CDR[3] (Mean±SD)	MMSE[4] (Mean±SD)
ADNI-1	AD	88/82	75.37±7.48	14.61±3.18	0.74±0.24	23.22±2.03
	pMCI	94/62	74.57±7.11	15.74±2.90	0.5±0.0	26.53±1.70
	sMCI	131/71	74.55±7.59	15.54±3.11	0.49±0.03	27.37±1.76
	NC	103/103	75.85±5.10	15.92±2.86	0.0±0.0	29.14±0.98
ADNI-2	AD	58/44	74.44±7.89	15.99±2.51	0.77±0.27	22.99±2.16
	pMCI	55/46	72.54±6.96	15.99±2.58	0.50±0.04	27.55±1.78
	sMCI	174/155	71.11±7.51	16.16±2.67	0.49±0.02	28.18±1.64
	NC	72/75	73.72±6.39	16.68±2.42	0.0±0.0	29.06±1.21
AIBL	AD	30/44	73.35±7.93	-	0.93±0.55	20.18±5.44
	pMCI	7/4	74.90±5.97	-	0.50±0.00	26.27±1.60
	sMCI	33/36	75.36±7.54	-	0.47±0.13	27.04±2.13
	NC	30/55	75.52±6.63	-	0.029±0.117	28.71±1.35

1. SD refers to standard deviation. 2. Number of years of education. 3. CDR is a scale used to assess the severity of cognitive impairment in individuals diagnosed with AD or other forms of dementia. 4.MMSE is calculated based on the subject's performance on a series of questions and tasks that evaluate cognitive ability.

211 AD, 133 MCI, and 768 NC patients. However, due to image quality and completeness requirements, only 74 AD, 80 MCI, and 85 NC records are selected.

Table 1 summarizes relevant clinical data regarding the participants of this study, including their gender, age, education level, clinical dementia rating (CDR), and Mini-Mental State Examination (MMSE) scores. Notably, there are sizable differences in MMSE scores observed among the AD, MCI, and NC groups. For instance, in the AD group of ADNI-1, there are 88 male and 82 female participants. Their average age is 75.37 years old, with a standard deviation of 7.48. On average, they have completed 14.61 years of education, with a standard deviation of 3.18. The mean CDR for this group is 0.74, with a standard deviation of 0.24. Additionally, the mean MMSE score is 23.22, with a standard deviation of 2.03.

2.2 Data Preprocessing

The sMRI images from the selected datasets in this study undergo initial processing with the Statistical Parametric Mapping (SPM) tool in MATLAB. SPM is a comprehensive tool used for the preprocessing of sMRI images. The process begins with the conversion of images to the NIFTI format, followed by reorientation and cropping to standardize orientation and removal of non-brain elements. Segmentation is then performed to classify voxels into different tissue types, incorporating bias field correction. The images are spatially normalized to a standard template, typically the MNI space, for comparability across subjects. SPM's GUI facilitates the selection and parameterization of these steps, and it also supports batch processing via MATLAB scripting. Once preprocessing is completed, the images are down-sampled to be used as inputs to the proposed CNN.

Before the training phase, a preconditioning process is applied, which includes addressing cranial anatomy, correcting for signal strength variations, and performing spatial registration. Data from the AD, MCI, and NC categories used in the experiments are evenly distributed to ensure unbiased training results. This allocation guarantees an equal representation of each category in the training phase.

2.3 The Proposed Method

We propose a 3D decoupling, self-attention network, which is composed of three main components: the MSD Block, SA Block and Classification Block, as shown in Fig. 1. Specifically, the input to the MSD block is preprocessed through downsampling and data augmentation. The downsampling process involves two convolutional layers of 32 and 64 channels, respectively, to reduce the image size and computational load. The data augmentation process transforms an original image into three distinct dimensional views: axial, coronal, and sagittal, for the convolution operation. Subsequently, sMRI data from the axial, sagittal, and coronal directions are used as the input to the MSD block. This block is responsible for extracting essential information and discarding extraneous data irrelevant to AD. After combining feature maps from all three directions, the resultant feature map is fed into the SA block. The SA block integrates the feature information derived from the three-dimensional views (axial, coronal, and sagittal), identifying areas

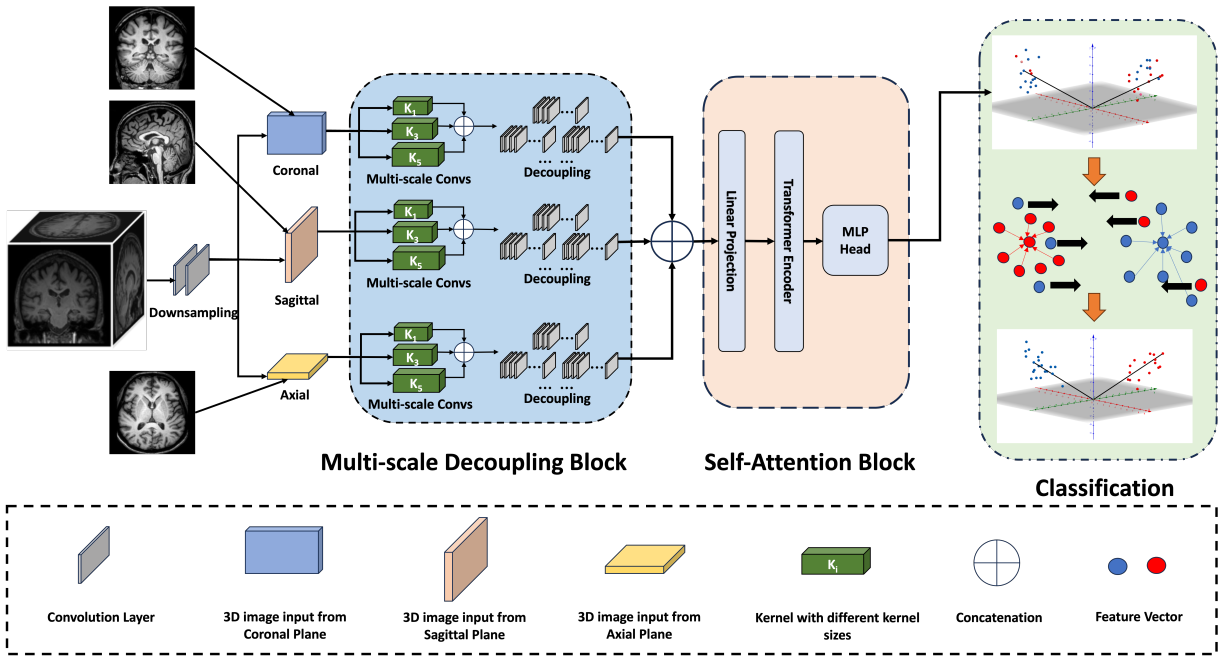


Fig. 1: Framework of the proposed network.

of interest relevant to AD. Finally, after operations like flattening and linear projection, the feature map is classified into one of three categories: NC, MCI, or AD. In the Classification part, the proposed clustering loss function is integrated into the overall loss to enhance discrimination between different classes, while minimizing variations within the same class, thus improving prediction accuracy. Details regarding each of the three main components in the proposed methodology are delineated below.

2.3.1 The MSD Block

It is widely known that increasing the depth and width of CNNs can boost their performance. As a classical model, the Visual Geometry Group(VGG) network [24] employed a strategy of stacking blocks of the same shape to increase the network depth, and was later adopted by deep CNN models. However, as network size grows, so does the number of parameters, causing the issue of overfitting when training data are insufficient. Therefore, how to extract feature maps from 3D sMRI images without a huge number of parameters is of significance in this research.

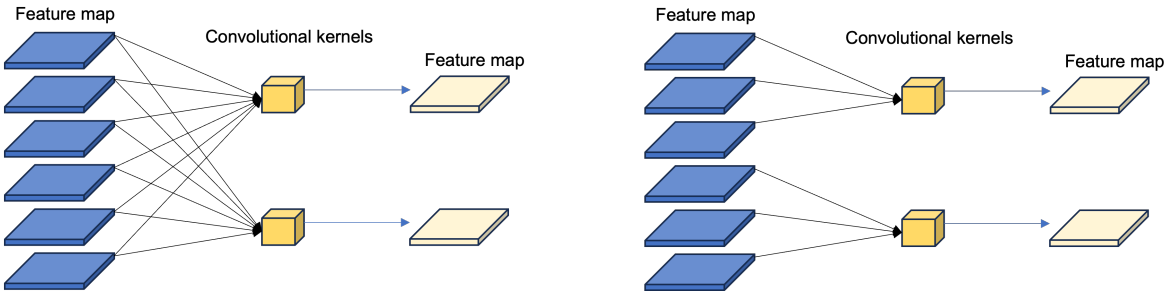
Since AD-related biomarkers are likely to manifest differently across scales, through multi-scale convolutions, the designed network is able to understand these biomarkers more accurately. Moreover, as pathological changes in AD may present differently at different scales, multi-scale convolutions can capture these changes more comprehensively, aiding the detection and accurate classification of the disease. Furthermore, conventional single-scale convolution might miss critical information due to the fixed size of the convolutional kernel. Multi-scale convolutions alleviate these issues by capturing important features from multiple scales, thus preventing information loss.

Han et al. [25] introduced a multi-scale CNN for AD prediction, utilizing distinct kernel sizes across layers to extract features. However, this approach only uses two average pooling and max pooling layers to extract multi-scale features, and it likely results in the loss of crucial information pertaining to AD within the original sMRI data. Information loss is inherent in each convolution layer, an inevitable aspect of the training process. To mitigate the information loss, we conduct multi-scale convolutions whose kernels are of different sizes based on 3D sMRI data from all three directions (i.e., axial, sagittal, and coronal). In this way, comprehensive features are captured at various scales. The obtained feature maps are combined after the multi-scale convolutions. The combined feature map is represented by $\mathbf{F}_{ms}$, expressed as

$$\begin{aligned}\mathbf{F}_{ms} = \mathbf{F}_a * \mathbf{K}_1(x_1, y_1, z_1) \oplus \mathbf{F}_s * \mathbf{K}_3(x_2, y_2, z_2) \\ \oplus \mathbf{F_c} * \mathbf{K}_5(x_3, y_3, z_3)\end{aligned} \tag{1}$$

where $\mathbf{F}_a$, $\mathbf{F}_s$, and $\mathbf{F}_c$ denote the feature maps obtained in the axial, sagittal and coronal planes, respectively; $\mathbf{K}_1$, $\mathbf{K}_3$, and $\mathbf{K}_5$ denote the convolution kernels of sizes $1\times1\times1$, $3\times3\times3$, and $5\times5\times5$, respectively. Symbol $\oplus$ represents the concatenation operation; and (x_i, y_i, z_i) are spatial coordinates, for $i \in \{1, 2, 3\}$.

While the multi-scale convolution benefits feature extraction, given the wealth of information contained in 3D sMRI data [26], a simple concatenation operation cannot fully establish the relationship between features, limiting both local and global feature learning. To address this issue, we propose the idea of "decoupling". That is, we introduce multi-channel-based group convolution [27] to decouple channels in different groups, which significantly reduces the number of parameters and computational load. Typically, when applying two convolution kernels to a feature map, each kernel is used on every channel of the feature map, as shown in (a) of Fig. 2. By contrast, when channels of a feature map are grouped, each kernel only



(a) Conventional Convolution (b) Group Convolution

Fig. 2: Comparison between conventional convolution and group convolution.

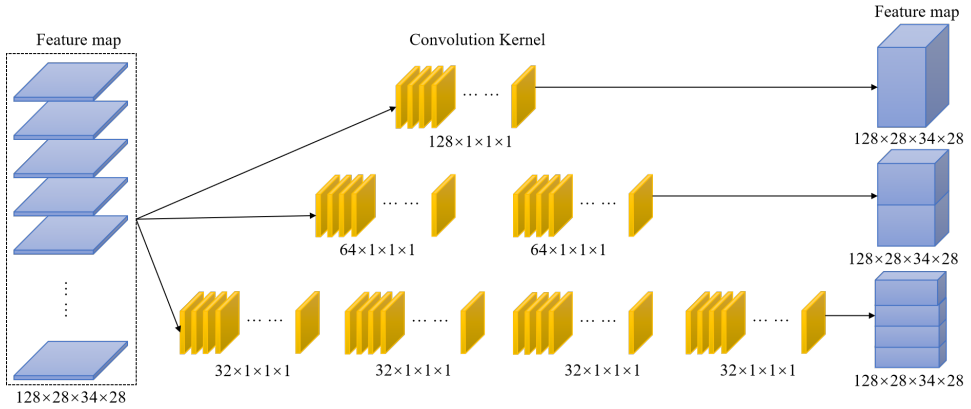


Fig. 3: The proposed decoupling process.

needs to be applied to the corresponding channel of the feature map, as shown in (b) of Fig. 2. Motivated by this, we apply the group convolution separately to the feature maps obtained after the concatenation.

The proposed decoupling process is depicted in Fig. 3. In the decoupling phase, there are sets of group convolution kernels with each set having a different number of groups. In each set, there are a channels with $a = C/G$, where C is the total number of channels and G the total number of groups in the set. During the 3D convolution, kernels in group g are only convolved with the feature maps in channel $[(g-1)a, g\times a)$, where $g = 1, 2, \ldots, G$. Let $\mathbf{O}(x, y, z, c)$ represent the value at position (x, y, z, c) in the output tensor and we get

$$\mathbf{O}(x,y,z,c) = \sum_{i=1}^{H_f}\sum_{j=1}^{W_f}\sum_{k=1}^{D_f}\sum_{m=(g-1)a}^{g\times a} \mathbf{F}_{ms}(x+i-\frac{H_f}{2}, y+j-\frac{W_f}{2}, z+k-\frac{D_f}{2}, c+m-\frac{a}{2}) \cdot \mathbf{F}_g(i,j,k,m) \quad (2)$$

where H_f, W_f, D_f denote the height, width, and depth of the convolutional filter, respectively. $\mathbf{F}_{ms}(x+i-\frac{H_f}{2}, y+j-\frac{W_f}{2}, z+k-\frac{D_f}{2}, c+m-\frac{a}{2})$ represents the feature map $\mathbf{F}_{ms}$ obtained in (1) at position $(x+i-\frac{H_f}{2}, y+j-\frac{W_f}{2}, z+k-\frac{D_f}{2}, c+m-\frac{a}{2})$ in the input tensor upon shifting the filter along the spatial dimensions and across groups; and $\mathbf{F}_g(i,j,k,m)$ represents the filter weight at position (i,j,k,m) for group g.

As illustrated in Fig. 3, the first set consists of one group with 128 channels, each of size $1 \times 1 \times 1$. The second set comprises two groups, each having 64 channels of size $1 \times 1 \times 1$. Finally, the third set contains four groups, each with 32 channels of size $1 \times 1 \times 1$. This, in turn, addresses the issue of overfitting that can arise from increasing the network's depth. The feature map is then convolved with these 3 sets of group convolutions.

By learning features independently within each group, the group convolutions in the decoupling process can enhance the robustness of our model. This independent feature learning enables the proposed network to focus on diverse aspects of the input data, potentially improving generalization performance over different subjects or under different imaging conditions. This approach also improves the collaboration among local information within the feature maps, allowing for more precise identification of intricate lesion areas while simultaneously extracting comprehensive and information-rich features.

2.3.2 The SA Block

Two attention mechanisms are incorporated in the proposed network. They are spatial attention and channel attention, enabling the model to reduce computational complexity while excluding redundant image information. Specifically, after the convolutions on 3D sMRI data are performed, the spatial attention mechanism is introduced to allocate reasonable weights to the obtained multiple feature maps. Additionally, the inclusion of channel attention following multi-scale

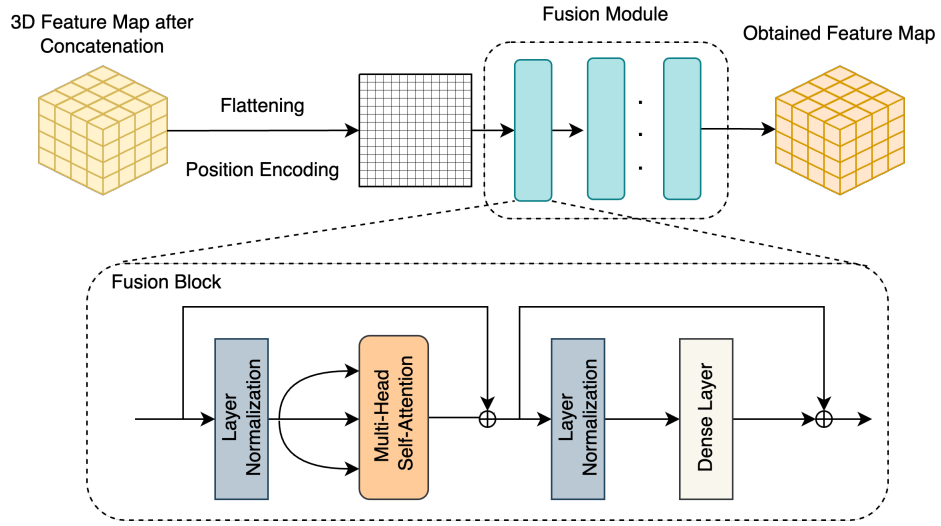


Fig. 4: Structure of the Self-Attention Block.

convolution ensures that the model can capture information from structural changes in brain regions but also differentiate the weight changes across various channels. After the concatenation of the three 3D feature maps along the channel dimension, they are fed into the SA [28] block to facilitate a rational fusion of the combined feature map.

Once features are extracted from 3D sMRI images in each of the three directions (sagittal, coronal and axial) from the MSD block, it is important to integrate these features across the three directions. Due to variations in observation angles and relative positions of the features from different directions, concatenating feature maps directly would not integrate the features well. This would, in return, hinder the model's ability to capture the accurate information about the features' position and structure. To harness the complementary nature of features from different directions, we incorporate a self-attention strategy in the design of the SA block. The design of this block is geared towards promoting synergistic relationships among features, ensuring that they mutually enhance one another.

Fig. 4 shows the structure of the SA block. Feature flattening and spatial position encoding are applied to the 3D feature map obtained after the concatenation so as to convert it into a one-dimensional representation, while preserving the original spatial information. This not only makes the proposed model more robust but also streamlines the input of features into a fusion module, responsible for merging the features from different directions and positions, achieved by the steps outlined below.

First, a number of volumetric patches are obtained from the combined feature representations of 3D sMRI images. Next, each patch is flattened into a vector (e.g., $\mathbf{x}_i$, where subscript i denotes the ith vector). Then, the flattened patches are linearly embedded to allow the network to learn the most suitable encoding matrices, such that $\mathbf{x}_i$ is projected into Query ($\mathbf{Q}$), Key ($\mathbf{K}$), and Value ($\mathbf{V}$) as follows.

$$\mathbf{Q} = \mathbf{x}_i \mathbf{W}_q, \quad \mathbf{K} = \mathbf{x}_i \mathbf{W}_k, \quad \mathbf{V} = \mathbf{x}_i \mathbf{W}_v \tag{3}$$

where $\mathbf{W}_q, \mathbf{W}_k, \mathbf{W}_v$ are the learnable encoding matrices. After the linear projection in (3), a layer normalization (LN) is applied to $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$, respectively. The LN operation helps stabilize and normalize feature vectors.

Because the attention scores influence the attention assigned to each element in the input sequence, they are calculated by assessing the similarities between $\mathbf{Q}$ and $\mathbf{K}$. This is done through the multi-head self-attention mechanism, described by

$$Attention(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = SoftMax\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}}\right)\mathbf{V} \tag{4}$$

where d is the encoding dimension of $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$; and $\mathbf{K}^T$ is the transpose of $\mathbf{K}$.

The outputs $Attention(\mathbf{Q}, \mathbf{K}, \mathbf{V})$ from self-attention layers go through a normalization layer, followed by a linear transformation applied in the dense connection layer. Finally, the outputs of multiple self-attention layers are combined with $\mathbf{x}_i$ using residual connections, thus alleviating performance degradation caused by excessive stacking of the fusion blocks [29]. Adaptive fusion of features is achieved by stacking multiple fusion blocks in different directions and positions.

With the multi-head self-attention mechanism expressed by (4), the SA block can prevent the proposed model from over-focusing on its position during encoding and attention score calculation. The operation in (4) differs from single-head self-attention in that it simultaneously projects $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ into multiple encoding spaces for individual self-attention calculations. The resulting outputs are projected back to the original dimensions, enhancing the model's representation capacity.

High-level details and contextual understanding provided by the SA block can reduce false positives and false negatives in AD prediction. This minimizes the risk of providing unnecessary treatment to normal subjects, or missing early intervention opportunities for AD patients. Also, by processing individual sMRI data with attention to unique brain features, the SA block can assist in developing personalized profiles of AD progression, which is particularly useful in tailoring treatment plans and monitoring the efficacy of interventions.

2.3.3 The Classification Block

A popular strategy in image classification is to minimize the cross-entropy loss, a process that effectively maximizes the logarithm of the probability associated with the target labels. However, only applying the softmax function [30] tends to over-emphasize the loss associated with the accurate label whilst struggling to consider losses of alternative label positions, thus decreasing the priority of reducing the probability of predicting incorrect labels. Consequently, issues such as overfitting and poor generalization occur.

To overcome the limitations of depending on Softmax alone, we introduce a clustering loss function [31], expressed as

$$L_{SC} = \frac{1}{2}\sum_{i=1}^{m} ||\mathbf{S}_i - \mathbf{SC}_{y_i}||^2 \tag{5}$$

where m is the total amount of training data in one batch; $\mathbf{S}_i$ is the score vector of the ith sample; and $\mathbf{SC}_{y_i}$ is the score center of Class y_i. The dimension of $\mathbf{S}_i$ and $\mathbf{SC}_{y_i}$ is equal to the number of categories. By minimizing the distance between the sample score and score center of their respective category, the classification stage facilitates the automatic learning of score centers for each category within the image data.

The clustering loss function (5) effectively controls the growth of predicted scores. In the context of tag position scores, it is usually desirable for the score to accurately represent the importance of the tag. By introducing the clustering loss function, the network is pushed to attain the label prediction outcome in a more steady manner. In other words, if the tag score grows rapidly, the clustering loss function (5) would intervene, regulating the network to allow a gradual score increase [32]. This enables the network to adopt a more cautious approach to its predictions, ultimately increasing the network's generalizability.

In the training process, $\mathbf{SC}_{y_i}$ is initialized as an all-zero vector, and constantly updated during training. The updates of $\mathbf{SC}_{y_i}$ are depicted below:

$$\Delta\mathbf{SC}_{y_i} = \frac{\sum_{i=1}^{m} \delta(y_i = j)\cdot(\mathbf{SC}_{y_i} - \mathbf{S}_i)}{1 + \sum_{i=1}^{m} \delta(y_i = j)} \tag{6}$$

where y_i denotes the label of the ith point; j is the class number; and $\delta(\cdot)$ is an indicator function that equals 1 if the ith data is assigned to the jth cluster (i.e., when $y_i = j$), and 0 otherwise.

In the proposed method, we combine the cross-entropy loss L_{CE} determined by the Softmax function and the clustering loss L_{SC} to train the network. The overall loss function is given by

$$L_{Total} = L_{CE} + \lambda \cdot L_{SC} \tag{7}$$

L_{SC} is expressed in (5); and L_{CE} is written as

$$L_{CE} = -\sum_{i=1}^{m} \frac{e^{S_{y_i}}}{\sum_{j=1}^{m} e^{S_{y_i}}} \tag{8}$$

where S_{y_i} is the score of class y_i in the ith sample; m is the total number of classes over which Softmax distributes probabilities; and λ is the coefficient adjusting the contribution of L_{SC}.

The overall loss L_{Total} is calculated in each iteration, and the weight parameter is updated using the gradient descent method. The updated weight parameter W^+ is obtained by

$$W^+ = W - \eta\frac{\partial L_{Total}}{\partial W} \tag{9}$$

where η is the learning rate adjusted by the early stop strategy; and W is the weight parameter from the previous iteration. This iterative training process imposes a constraint on the scores of each category, guiding them to remain close to their respective score centers. As a result, it curbs overfitting by preventing scores for the correct tag positions from growing infinitely.

3 Results

3.1 The Experimental Settings

The proposed model is implemented based on Python 3.9 and Pytorch-GPU (3090) and run with the PyTorch framework. All experiments are carried out under the Windows 10 operating system. Other hardware configurations include CPU (Intel i7-12700K@3.6GHz), 64GB memory, and 1TB hard disk.

To achieve optimal performance with minimal computation load, the MSD block is designed to repeat a number of times across four stages. In the first stage, rather than the multi-scale convolution, two 3D convolutions are conducted, and the MSD block repeats four times with 64 input channels and 128 output channels. In the second stage, the MSD block runs four times with 128 input channels and 256 output channels. During this stage, the multi-scale convolution runs with kernels of sizes $1\times1\times1$, $3\times3\times3$ and $5\times5\times5$. However, during the second half of Stage 2, the kernel sizes are reduced to $1\times1\times1$ and $3\times3\times3$. The process in the third stage is the same as in Stage 2 except that the repetition of the MSD block is six times with 256 input channels and 512 output channels. In the fourth stage, the process mirrors that of Stage 1, with 512 input channels and 1024 output channels. The decoupling operation remains consistent across all four stages.

To guarantee a smooth operation of the proposed network, we employ the data augmentation method [33] to address overfitting, and incorporate deep learning approaches like the addition of batch normalization, and utilization of the early stop technique [34] to continually optimize the model's hyperparameters. Images from different planes are randomly rotated in small increments, and images are randomly masked as part of these strategies. It is worth noting that with the early stop technique, our model can automatically stop running if the loss is not reduced for five consecutive iterations, thereby saving time in the search for hyperparameters close to the trough of the gradient descent curve.

In the proposed network, the adjustment of hyperparameters (e.g., batch size, number of epochs, learning rate, and weight decay) is contingent upon the early stop strategy. Should the overall loss fail to decrease within the margin of 0.01, the execution would stop. Upon cessation, the hyperparameters undergo further refinement in a predefined range using the control variable method [35]. This approach expedites the discovery of optimal parameters and facilitates the convergence of the loss function to its minimum value. As part of the specified range, each hyperparamter adjustment must not change its original value by more than one order of magnitude.

3.2 The Performance Evaluation Metrics

To evaluate the performance of the proposed AD prediction model, we adopt the performance measures of accuracy (ACC), sensitivity (SEN), specificity (SPE), and the area under the ROC curve (AUC) [45]. ACC refers to the proportion of all test results that are correctly identified. SEN represents the ability of a test to correctly identify individuals who have a given disease or condition (true positives). SPE is defined as the ability of a test to correctly identify individuals who do not have the disease (true negatives). AUC is a single measurement that summarizes the performance of a test across all possible classification thresholds. The calculation formulas [46] for ACC, SEN and SPE are:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{10}$$

$$SEN = \frac{TP}{TP + FN} \tag{11}$$

$$SPE = \frac{TN}{TN + FP} \tag{12}$$

where TP, TN, FP, and FN stand for true positives, true negatives, false positives, and false negatives, respectively.

3.3 Existing Related Methods

We compare the proposed model with other related state-of-the-art methods and provide a brief description of each method below.

Slice+SVM: The system proposed in [36] is a combination of a support vector machine (SVM) trained with various texture descriptors derived from MRI slice data and a SVM trained with markers constructed from MRI voxels. The two sets of SVMs are tuned according to a weighted-sum rule to yield a final decision.

Slice+CNN: The network in [17] uses a simple CNN with several convolution layers, pooling layers, and a fully connected layer to extract features from 2D sMRI images. The pooling operation is for reconstructing feature maps to save computational resources.

Slice+WholeBrain+CNN: The model in [20] is characterized by two primary branches comprising a 2D convolutional network and a 3D convolutional network. Features from both networks are subsequently combined through the fully connected layer, and classification is conducted through a Softmax function.

WholeBrain CNN+Self ATT: Zhang et al. [18] proposed a 3D convolution followed by a self-attention block with a residual block to capture both local and global features from sMRI images.

WholeBrain CNN+Global ATT: In [21], 3D images are initially subjected to a 3D convolution from three distinct angles, generating a feature map formed by the concatenation of three individual feature maps. This allows for the observation of a broader range of information. The feature map is subsequently directed to a multi-scale convolution block, facilitating the extraction of features that span from local to global areas.

ATT: The network in [15] employs an attention-based framework to extract multi-level discriminative information from sMRI data, supporting the diagnosis of AD.

Self ATT: Gao et al. [37] proposed a task-induced pyramid and an attention generative adversarial network for the imputation and classification of multi-modal brain images. This helps generate better image details by focusing on relevant features across an entire image.

Spatial Disease ATT: The attention network in [38] utilizes an average pooling layer for attention extraction and employs weakly supervised discriminative localization to aid classification.

Spatial+Channel ATT: The network in [39] has two phases in its attention mechanism. In the initial attention phase, global representations are meticulously aggregated through second-order attention pooling. Following this, the second attention stage judiciously disperses these global features to each individual spatial location. This precise distribution ensures that every point within the feature map is infused with tailored global information. Multiple pooling layers are then added to obtain a comprehensive global image representation.

MIL: The model in [40] employs a dual-attention multi-instance deep learning network for diagnosing patients with early AD and MCI. The multiple instance learning (MIL) technique aids the model in balancing

Table 2: Performance comparison on the ADNI dataset.

Reference	Method	AD vs. NC				pMCI vs. sMCI			
		ACC	AUC	SEN	SPE	ACC	AUC	SEN	SPE
Nanni et al. [36]	Slice+SVM	0.876	0.903	0.841	-	0.671	0.865	0.345	-
Xing et al. [17]	Slice+CNN	0.953	-	0.889	0.974	-	-	-	-
Chen et al. [20]	Slice+ WholeBrain CNN	0.911	0.950	0.914	0.888	0.801	0.789	0.520	0.856
Zhang et al. [18]	WholeBrain CNN+ Self ATT	0.910	-	0.910	0.920	0.820	-	0.810	0.810
Liu et al. [21]	WholeBrain CNN+ Global ATT	0.977	**0.977**	**0.968**	**0.985**	0.883	**0.892**	**0.840**	**0.944**
Gao et al. [37]	Self ATT	0.920	0.956	0.891	0.940	0.753	0.786	0.773	0.741
Lian et al. [38]	Spatial Disease ATT	0.919	0.965	0.887	0.945	0.827	0.793	0.579	0.866
Guan et al. [39]	Spatial+ Channel ATT	0.872	0.927	0.890	0.856	0.793	0.776	0.546	0.841
Our model	WholeBrain CNN+ Global ATT	**0.985**	0.946	0.962	0.983	**0.894**	0.887	0.796	0.880

Note: Results of [18, 21, 36, 38] in this table are cited from [21], while other results are cited from corresponding papers.
ATT refers to the attention mechanism.
The best performance is highlighted in bold, while the second best performance is underlined.

Table 3: Performance Comparison on the AIBL dataset.

Reference	Method	AD vs. NC			
		ACC	AUC	SEN	SPE
Guan et al. [15]	ATT	0.903	0.953	0.873	0.908
Lian et al. [38]	Spatial Disease ATT	0.898	**0.974**	0.873	0.908
Zhu et al. [40]	MIL	0.911	0.950	0.914	0.888
Liu et al. [21]	WholeBrain CNN+Global ATT	0.949	0.951	0.929	0.972
Our model	WholeBrain CNN+Global ATT	**0.963**	0.955	**0.933**	**0.980**

Note: **The best performance** is highlighted in bold, while the second best performance is underlined.

Table 4: Performance Comparison on the OASIS dataset.

Reference	Method	AD vs. NC			
		ACC	AUC	SEN	SPE
Salami et al. [41]	WholeBrain CNN	0.877	-	-	-
Islam et al. [42]	WholeBrain CNN	0.932	-	-	-
Saratxaga et al. [43]	WholeBrain CNN	0.92	-	-	-
He et al. [44]	WholeBrain CNN+Global ATT	0.92	-	-	-
Our model	WholeBrain CNN+Global ATT	**0.945**	0.970	0.920	0.960

Note: **The best performance** is highlighted in bold, while the second best performance is underlined.

the relative contribution of features, so that the most relevant lesion area in the brain can be identified.

3.4 Experiment Results

The proposed method is implemented and tested using the datasets of ADNI, AIBL and OASIS, and is compared with the state-of-the-art methods. The experimental results on datasets ADNI, AIBL and OASIS

are shown in Table 2, Table 3 and Table 4, respectively. The methods under comparison in these tables are described in the Section 3.3. The proposed method shows a superior prediction performance and reflects the robustness of the proposed model.

In the classification experiments for AD and NC, it is evident that our model achieves the highest prediction accuracy among all the evaluated methods with an ACC of 0.985 on the ADNI dataset.

Unlike the noticeable differences in images between the AD and NC groups, the differentiation of images between the pMCI and sMCI groups is less eminent, which makes it hard to distinguish. The proposed model achieves an accuracy of 0.894, which is the best among all the methods under comparison. This is because 1) our model can extract features at local and global levels and performs decoupling operations among different channels; and 2) the attention mechanisms (both spatial attention and channel attention) allow the proposed model to effectively localize the regions of brain physiological changes.

It is worth pointing out that results obtained through deep learning algorithms exhibit higher accuracy than those through alternative algorithms. This reflects the limitations of alternative machine learning algorithms (e.g., SVM, Long Short-Term Memory (LSTM), and similar approaches) with only shallow semantic information obtained from sMRI images. In contrast, deep learning algorithms excel in their ability to accurately identify and classify brain lesion areas. Confusion matrices presented in Fig. 5 and Fig. 6 illustrate the comparative robustness of our model in classifying AD vs. NC, and sMCI vs. pMCI, respectively. The different numbers of actual positive and negative labels in these figures reflect the class imbalance in our dataset. Specifically, for classification task AD vs. NC, our dataset contains a total of 625 samples, among which 272 samples belong to the positive class and 353 samples belong to the negative class. For classification task sMCI vs. pMCI, our dataset contains a total of 788 samples, among which 257 samples belong to the positive class and 531 samples belong to the negative class. This imbalance is inherent to the dataset and is representative of real-world scenarios where certain classes may be underrepresented. The results clearly indicate that the proposed method outperforms other methods under comparison.

3.5 Ablation Study

In order to thoroughly assess the efficacy of the MSD block, SA block, and regularized loss function of the proposed method, we conduct ablation studies on different combinations of the network components. The outcomes of these experiments are presented in Fig. 7. The 'Baseline' model in Fig. 7 means that the proposed network does not take the MSD block and SA block into consideration, and relies solely on the cross-entropy loss function for the classification task.

From Fig. 7 we can make the following observations:

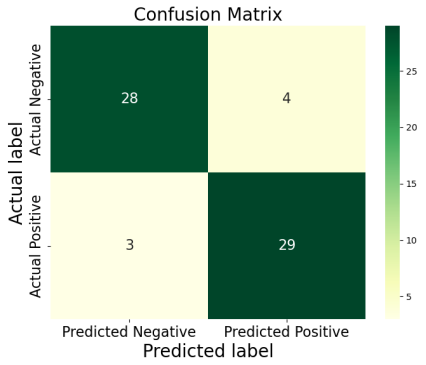
1. Adding the MSD block to the baseline model, namely 'Baseline+MSD', surpasses the baseline model in both classification accuracy and AUC. Conversely, the single-channel convolution tends to be susceptible to overfitting concerns due to its constrained feature extraction capacity and inadequate feature representation. Having the MSD block not only reduces parameters but also strengthens feature quality, promoting the amalgamation of global and local information and empowering the model to capture fine-grained features efficiently.
2. The inclusion of the SA block, that is, 'Baseline+MSD+SA', enables a synergistic fusion of features from all three directions, allowing the model to concurrently consider features from all the aspects, resulting in heightened classification accuracy.
3. The whole model gives the best classification accuracy and AUC, which is attributable to the joint loss function (i.e., combining the clustering loss with the cross-entropy loss). This outcome validates the effectiveness of the joint loss function in learning class centroids, minimizing the distance between data samples within the same class while increasing the separation between distinct classes. It is important to note that relying solely on the cross-entropy loss function often makes the network over-prioritize the prediction label, thus potentially impeding its generalizability. By introducing the clustering loss function, the method exhibits a smoother scoring progression, refraining from making absolute predictions and improving its ability to generalize. The joint loss function plays a pivotal role in boosting performance, ensuring more precise and generalizable predictions.

4 Discussion

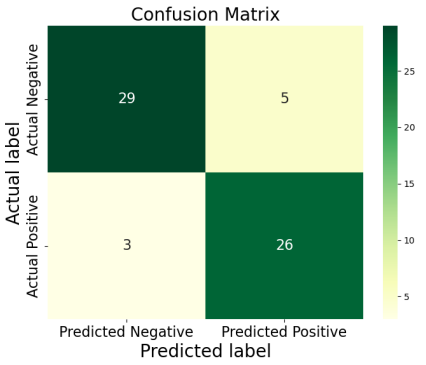
Several methods [17, 47, 48] were designed to process 2D image data, where feature information was extracted from various slices by the CNN, and achieved a classification accuracy exceeding 80 percent. However, these methods, primarily relying on 2D CNNs and single feature selection, exhibit notable drawbacks:

1. Discontinuous brain slices can influence the assessment of afflicted brain regions.
2. Disease prediction, driven by feature selection, may be biased towards distinguishing brain structures while overlooking natural individual variation, which can lead to misclassification.

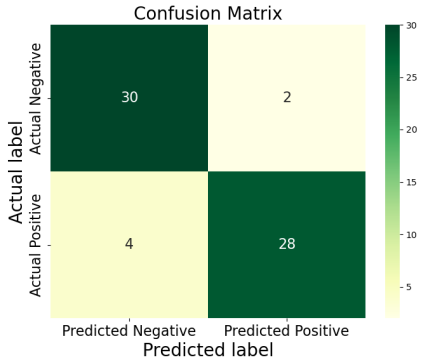
In contrast to 2D CNNs, 3D CNNs utilizing whole-brain images offer the potential for higher classification accuracy as they can comprehensively consider individual variation. Nevertheless, due to a myriad of information in 3D MRI images compared to 2D images, it is more challenging to fully and accurately identify brain atrophy features. As a remedy, multi-scale convolution is proposed in this paper, so that small- and large-scale changes in the brain structure can be captured simultaneously.



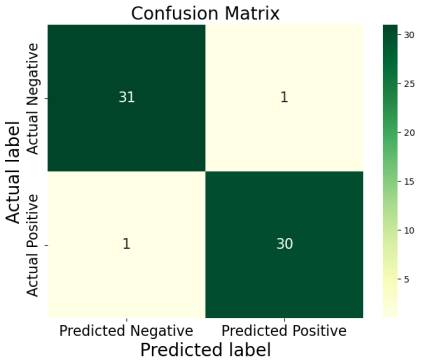
(a) Confusion Matrix of Chen's Model [20]



(b) Confusion Matrix of Guan's Model [39]

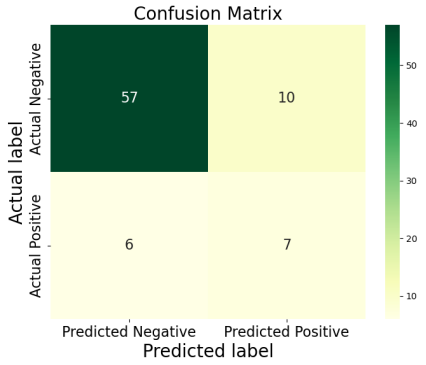


(c) Confusion Matrix of Lian's Model [38]

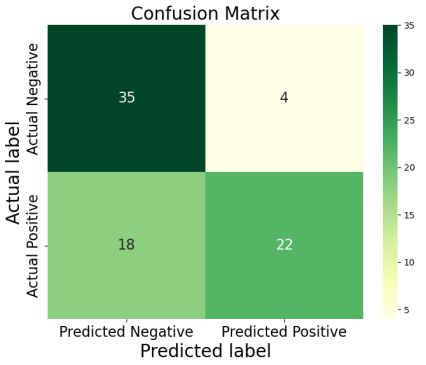


(d) Confusion Matrix of Our Model

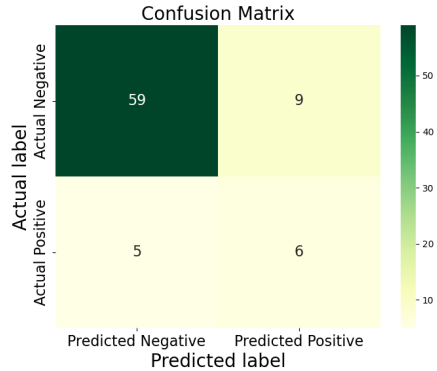
Fig. 5: Confusion matrices of different models on the classification task of AD vs. NC over dataset ADNI.



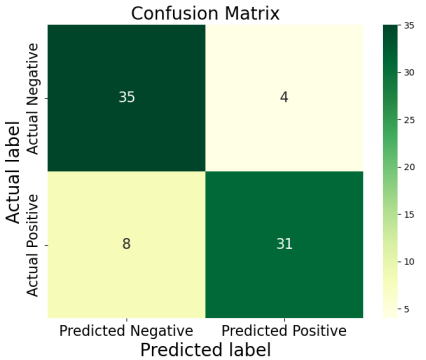
(a) Confusion Matrix of Chen's Model [20]



(b) Confusion Matrix of Guan's Model [39]



(c) Confusion Matrix of Lian's Model [38]



(d) Confusion Matrix of Ours Model

Fig. 6: Confusion matrices of different models on the classification task of sMCI vs. pMCI over dataset ADNI.

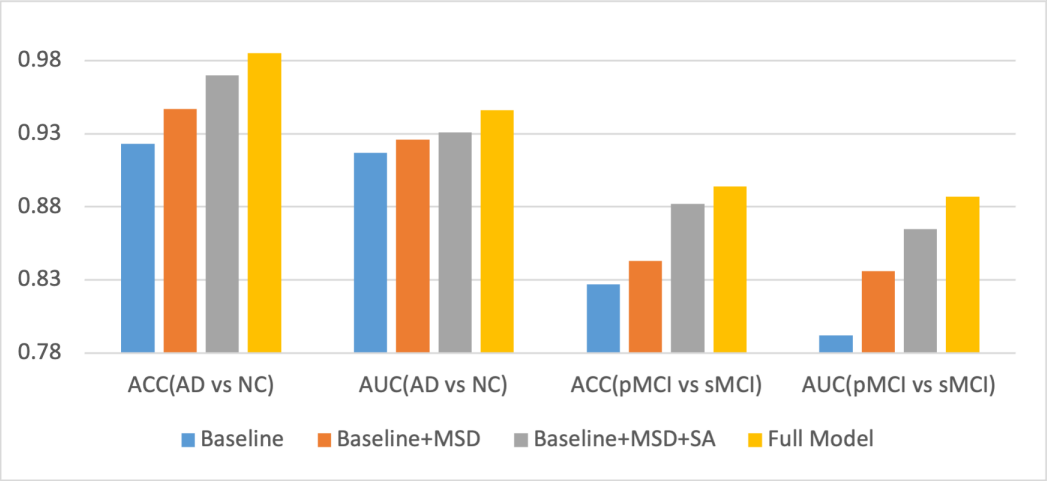
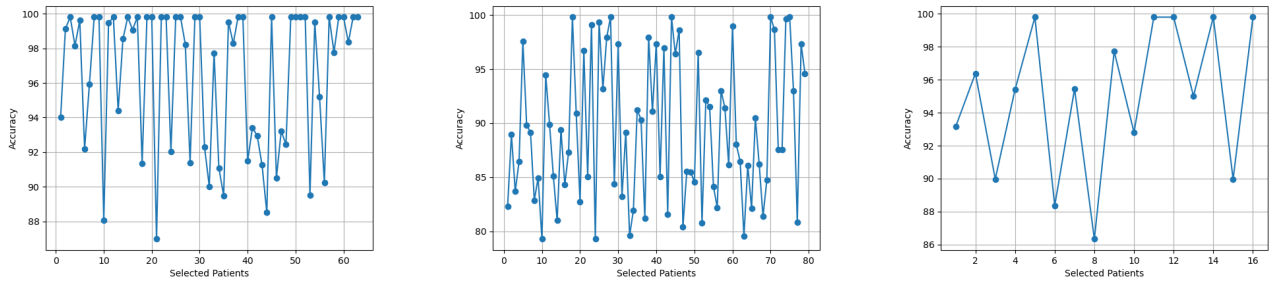
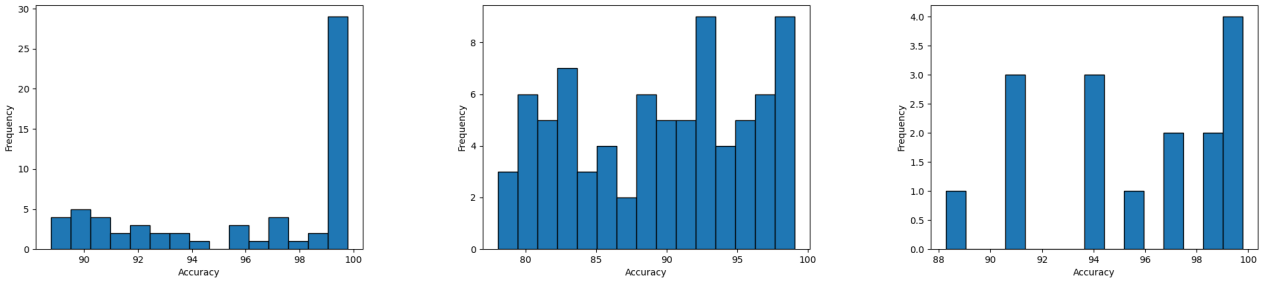


Fig. 7: Results of ablation studies. 'Baseline' means that only the cross-entropy loss function is in action for the classification task while MSD Block, SA Block and the clustering loss function are all deactivated.



(a) AD vs. NC on ADNI (b) pMCI vs. sMCI on ADNI (c) AD vs. NC on AIBL

Fig. 8: Line plots obtained from both ADNI and AIBL datasets.



(a) AD vs. NC on ADNI (b) pMCI vs. sMCI on ADNI (c) AD vs. NC on AIBL

Fig. 9: Histograms obtained from both ADNI and AIBL datasets.

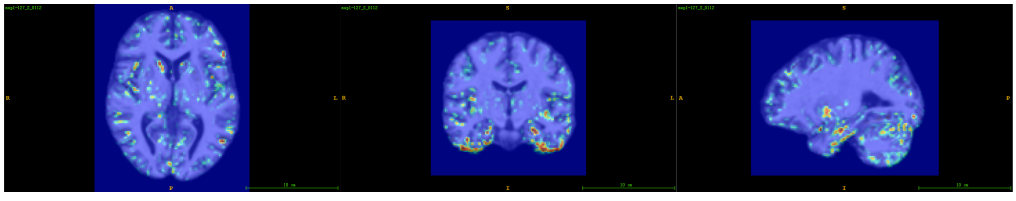


Fig. 10: An AD patient's locations of pathology identified by the proposed model.

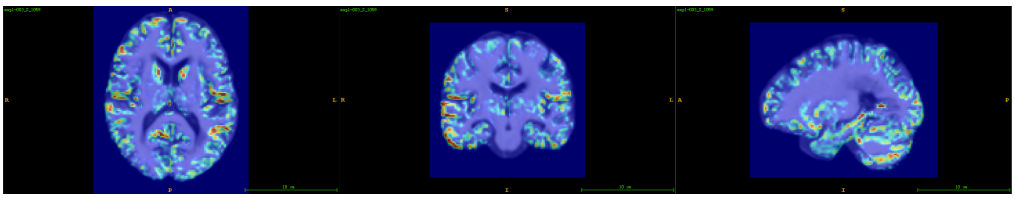


Fig. 11: An MCI patient's locations of pathology identified by the proposed model.

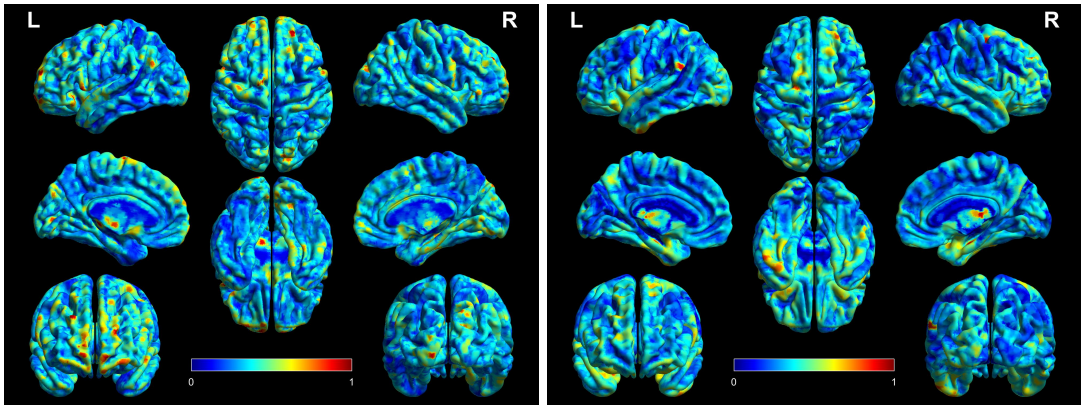


Fig. 12: 3D Heatmaps of the brain with extracted features generated by the proposed model.

However, when performing 3D convolutions, there is a potential for information interaction between different regions, leading to image misclassification. Some brain structural features may be unrelated to AD. Yet they become involved in the convolution operations, resulting in less weights being assigned to affected brain areas during weight calculations, impacting classification accuracy. In essence, 3D CNN models using multi-scale convolutions may struggle to capture structural changes across the entire brain.

To handle the issues associated with structural variations in the brain, the input 3D sMRI images may be rotated, resulting in a transformation from one image to three images. The proposed mode is then partitioned into three branches, with the addition of the MSD convolution operation to mitigate interactions among different image regions. Furthermore, the SA mechanism is incorporated to obtain the fused image. Each step of this process aims to selectively retain more lesion-related features while eliminating a substantial volume of redundant information within the 3D image. Nevertheless, parameters generated by a 3D CNN with multiple branches are far more than those generated by a 2D CNN. To tackle this issue, we resort to the proposed group convolution with $1 \times 1 \times 1$ convolution kernels to realize dimensionality reduction, which saves the computational cost.

In this study, we have also compiled data on the performance of our model for each individual patient. Fig. 8 illustrates the line plots that depict the model's performance for each patient in both AD vs. NC and pMCI vs. sMCI tasks. Similarly, Fig. 9 presents the histograms that intuitively express the proportion of our model's classification accuracy among patients. From these figures, it is evident that while our model demonstrates strong overall performance in both tasks, there are certain patients for whom the diagnosis is not accurately rendered. This inconsistency may be attributable to individual variability in the human brain structure and function. In some cases, patients with severe conditions may exhibit only minor neurological changes, posing challenges for accurate classification. Future research should therefore aim to address these challenges by focusing on individual differences, enhancing the model's ability to accurately diagnose patients with subtle neurological variations.

To assess whether the proposed method can extract feature maps effectively, we carry out a series of visual experiments. As depicted in Fig. 10 and Fig. 11, Fig. 10 illustrates the pathological areas detected by the model during the AD classification task, whereas Fig. 11 highlights the pathological regions identified by the model in the MCI conversion prediction task. Therefore, the AD and MCI patients' pathological locations identified by the proposed method agree with AD-related research literature [49, 50]. Notably, the 3D heatmap [51] (Fig. 12) of the brain indicates that the focus area is in close proximity to the hippocampus, substantiating established research that AD is intrinsically linked to the hippocampus, which is responsible for memory storage.

5 Conclusion

To overcome the limitations of feature information in 2D sections and the challenges in dealing with 3D sMRI images, we have proposed a new AD prediction method that consists of three main components – the MSD block, SA block and Classification block. The MSD block helps utilize global and local information and capture detailed features, while reducing the computational load through the decoupling process.

The SA block can fuse features from the three directions (i.e., axial, sagittal, and coronal), allowing more attention to be directed towards brain lesion areas. Moreover, the joint loss function in the classifier, which integrates the clustering loss with the cross-entropy loss, greatly improves generalizability and prediction performance of the proposed method. The experiment results validate the proposed design, as our model achieves excellent classification accuracy of 0.985 on the ADNI dataset and 0.963 on the AIBL dataset when compared with other related methods. Furthermore, the proposed algorithm effectively discriminates AD patients from the NC group and pMCI patients from the sMCI group, manifesting its potential for clinical applications. Additionally, the method proposed in this paper could be applied to other areas, such as EEG classification tasks [52, 53].

Declarations

- 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.
- Consent for publication
  All authors have checked the manuscript and have agreed to the submission.
- Authors' contributions
  All authors contributed to the study conception, design, and manuscript writing. All authors read and approved the final manuscript.

References

[1] Philip Scheltens, Bart De Strooper, Miia Kivipelto, Henne Holstege, Gael Chételat, Charlotte E Teunissen, Jeffrey Cummings, and Wiesje M van der Flier. Alzheimer's disease. *The Lancet*, 397(10284):1577–1590, 2021.

[2] Tai Nguyen-Ky, Peng Wen, and Yan Li. Consciousness and depth of anesthesia assessment based on bayesian analysis of eeg signals. *IEEE Transactions on Biomedical Engineering*, 60(6):1488–1498, 2013.

[3] Thomas Schmierer, Tianning Li, and Yan Li. A novel empirical wavelet sodp and spectral entropy based index for assessing the depth of anaesthesia. *Health Information Science and Systems*, 10(1):10, 2022.

[4] Siuly Siuly, Yan Li, Peng Wen, and Omer Faruk Alcin. Schizogooglenet: The googlenet-based deep feature extraction design for automatic detection of schizophrenia. *Computational Intelligence and Neuroscience*, 2022(1):1992596, 2022.

[5] Jessica Izzo, Ole A. Andreassen, Lars T. Westlye, and Dennis van der Meer. The association between hippocampal subfield volumes in mild cognitive impairment and conversion to alzheimer's disease. *Brain Research*, 1728:146591, 2020.

[6] Yan Li, Peng Wen, David Powers, and C Richard Clark. Lsb neural network based segmentation of mr brain images. In *IEEE SMC'99 Conference Proceedings. 1999 IEEE International Conference on Systems, Man, and Cybernetics (Cat. No. 99CH37028)*, volume 6, pages 822–825. IEEE, 1999.

[7] Md R. Bashar, Yan Li, and Peng Wen. Study of EEGs from Somatosensory Cortex and Alzheimer's Disease Sources, September 2011.

[8] Yan Li and Zheru Chi. Mr brain image segmentation based on self-organizing map network. *International Journal of Information Technology*, 11, 2005.

[9] Adam M Brickman, Laura B Zahodne, Vanessa A Guzman, Atul Narkhede, Irene B Meier, Erica Y Griffith, Frank A Provenzano, Nicole Schupf, Jennifer J Manly, Yaakov Stern, et al. Reconsidering harbingers of dementia: progression of parietal lobe white matter hyperintensities predicts alzheimer's disease incidence. *Neurobiology of aging*, 36(1):27–32, 2015.

[10] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[11] Rubina Sarki, Khandakar Ahmed, Hua Wang, Yanchun Zhang, and Kate Wang. Convolutional neural network for multi-class classification of diabetic eye disease. *EAI Endorsed Transactions on Scalable Information Systems*, 9(4), 2021.

[12] Dinesh Pandey, Hua Wang, Xiaoxia Yin, Kate Wang, Yanchun Zhang, and Jing Shen. Automatic breast lesion segmentation in phase preserved dce-mris. *Health Information Science and Systems*, 10(1):9, 2022.

[13] Shui-Hua Wang, Qinghua Zhou, Ming Yang, and Yu-Dong Zhang. ADVIAN: Alzheimer's Disease VGG-Inspired Attention Network Based on Convolutional Block Attention Module and Multiple Way Data Augmentation. *Frontiers in Aging Neuroscience*, 13, 2021.

[14] Zhehao Zhang, Linlin Gao, Guang Jin, Lijun Guo, Yudong Yao, Li Dong, Jinming Han, and the Alzheimer's Disease NeuroImaging Initiative. THAN: task-driven hierarchical attention network for the diagnosis of mild cognitive impairment and Alzheimer's disease. *Quantitative Imaging*

in Medicine and Surgery, 11(7):3338354–3333354, July 2021. Publisher: AME Publishing Company.

[15] Hao Guan, Yunbi Liu, Erkun Yang, Pew-Thian Yap, Dinggang Shen, and Mingxia Liu. Multi-site MRI harmonization via attention-guided deep domain adaptation for brain disorder identification. *Medical Image Analysis*, 71:102076, July 2021.

[16] Gia Minh Hoang, Ue-Hwan Kim, and Jae Gwan Kim. Vision transformers for the prediction of mild cognitive impairment to alzheimer's disease progression using mid-sagittal smri. *Frontiers in Aging Neuroscience*, 15, 2023.

[17] Ying Xing, Yu Guan, Bin Yang, and Jingze Liu. Classification of smri images for alzheimer's disease by using neural networks. In Shiqi Yu, Zhaoxiang Zhang, Pong C. Yuen, Junwei Han, Tieniu Tan, Yike Guo, Jianhuang Lai, and Jianguo Zhang, editors, *Pattern Recognition and Computer Vision*, pages 54–66, Cham, 2022. Springer Nature Switzerland.

[18] Xin Zhang, Liangxiu Han, Wenyong Zhu, Liang Sun, and Daoqiang Zhang. An explainable 3d residual self-attention deep neural network for joint atrophy localization and alzheimer's disease diagnosis using structural mri. *IEEE Journal of Biomedical and Health Informatics*, 26(11):5289–5297, 2022.

[19] Ibtissam Bakkouri, Karim Afdel, Jenny Benois-Pineau, and Gwenaelle Catheline. Recognition of alzheimer's disease on smri based on 3d multi-scale cnn features and a gated recurrent fusion unit. In *2019 International Conference on Content-Based Multimedia Indexing (CBMI)*, pages 1–6, 2019.

[20] Lin Chen, Hezhe Qiao, and Fan Zhu. Alzheimer's disease diagnosis with brain structural mri using multiview-slice attention and 3d convolution neural network. *Frontiers in Aging Neuroscience*, 14, 2022.

[21] Fei Liu, Huabin Wang, Shiuan-Ni Liang, Zhe Jin, Shicheng Wei, and Xuejun Li. Mps-ffa: A multiplane and multiscale feature fusion attention network for alzheimer's disease prediction with structural mri. *Computers in Biology and Medicine*, 157:106790, 2023.

[22] Shicheng Wei, Yan Li, and Wencheng Yang. An Adaptive Feature Fusion Network for Alzheimer's Disease Prediction. In *Health Information Science*, Lecture Notes in Computer Science, pages 271–282, Singapore, 2023. Springer Nature.

[23] R. C. Petersen, P. S. Aisen, L. A. Beckett, M. C. Donohue, A. C. Gamst, D. J. Harvey, Jr C. R. Jack, W. J. Jagust, L. M. Shaw, A. W. Toga, J. Q. Trojanowski, and M. W. Weiner. Alzheimer's disease neuroimaging initiative (adni). *Neurology*, 74(3):201–209, 2010.

[24] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2015.

[25] Ruizhi Han, Zhulin Liu, and CL Philip Chen. Multi-scale 3d convolution feature-based broad learning system for alzheimer's disease diagnosis via mri images. *Applied Soft Computing*, 120:108660, 2022.

[26] Kun Hu, Yijue Wang, Kewei Chen, Likun Hou, and Xiaoqun Zhang. Multi-scale features extraction from baseline structure mri for mci patient classification and ad early diagnosis. *Neurocomputing*, 175:132–145, 2016.

[27] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

[28] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[29] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778. IEEE, 2016.

[30] Kunal Banerjee, Vishak Prasad C, Rishi Raj Gupta, Karthik Vyas, Anushree H, and Biswajit Mishra. Exploring alternatives to softmax function, 2020.

[31] Elie Aljalbout, Vladimir Golkov, Yawar Siddiqui, Maximilian Strobel, and Daniel Cremers. Clustering with deep learning: Taxonomy and new methods. *arXiv preprint arXiv:1801.07648*, 2018.

[32] Dong Liang, Lanfen Lin, Hongjie Hu, Qiaowei Zhang, Qingqing Chen, Yutaro lwamoto, Xianhua Han, and Yen-Wei Chen. Combining convolutional and recurrent neural networks for classification of focal liver lesions in multi-phase ct images. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, pages 666–675, Cham, 2018. Springer International Publishing.

[33] Phillip Chlap, Hang Min, Nym Vandenberg, Jason Dowling, Lois Holloway, and Annette Haworth.

A review of medical image data augmentation techniques for deep learning applications. *Journal of Medical Imaging and Radiation Oncology*, 65(5):545–563, 2021.

[34] Lutz Prechelt. Early stopping-but when? In *Neural Networks: Tricks of the trade*, pages 55–69. Springer, 2002.

[35] Jeremy B Bernerth and Herman Aguinis. A critical review and best-practice recommendations for control variable usage. *Personnel psychology*, 69(1):229–283, 2016.

[36] Loris Nanni, Sheryl Brahnam, Christian Salvatore, and Isabella Castiglioni. Texture descriptors and voxels for the early diagnosis of alzheimer's disease. *Artificial Intelligence in Medicine*, 97:19–26, 2019.

[37] Xingyu Gao, Feng Shi, Dinggang Shen, and Manhua Liu. Task-induced pyramid and attention gan for multimodal brain image imputation and classification in alzheimer's disease. *IEEE journal of biomedical and health informatics*, 26(1):36–43, 2021.

[38] Chunfeng Lian, Mingxia Liu, Yongsheng Pan, and Dinggang Shen. Attention-guided hybrid network for dementia diagnosis with structural mr images. *IEEE transactions on cybernetics*, 52(4):1992–2003, 2020.

[39] Hao Guan, Chaoyue Wang, Jian Cheng, Jing Jing, and Tao Liu. A parallel attention-augmented bilinear network for early magnetic resonance imaging-based diagnosis of alzheimer's disease. *Human Brain Mapping*, 43(2):760–772, 2022.

[40] Wenyong Zhu, Liang Sun, Jiashuang Huang, Liangxiu Han, and Daoqiang Zhang. Dual attention multi-instance deep learning for alzheimer's disease diagnosis with structural mri. *IEEE Transactions on Medical Imaging*, 40(9):2354–2366, 2021.

[41] Farzaneh Salami, Ali Bozorgi-Amiri, Ghulam Mubashar Hassan, Reza Tavakkoli-Moghaddam, and Amitava Datta. Designing a clinical decision support system for alzheimer's diagnosis on oasis-3 data set. *Biomedical Signal Processing and Control*, 74:103527, 2022.

[42] Jyoti Islam and Yanqing Zhang. Brain mri analysis for alzheimer's disease diagnosis using an ensemble system of deep convolutional neural networks. *Brain informatics*, 5:1–14, 2018.

[43] Cristina L Saratxaga, Iratxe Moya, Artzai Picón, Marina Acosta, Aitor Moreno-Fernandez-de Leceta, Estibaliz Garrote, and Arantza Bereciartua-Perez. Mri deep learning-based solution for alzheimer's disease prediction. *Journal of personalized medicine*, 11(9):902, 2021.

[44] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[45] Tom Fawcett. An introduction to roc analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006.

[46] Bowen Zheng, Ang Gao, Xiaona Huang, Yuhan Li, Dong Liang, and Xiaojing Long. A modified 3d efficientnet for the classification of alzheimer's disease using structural magnetic resonance images. *IET Image Processing*, 17(1):77–87, 2023.

[47] Atif Mehmood, Muazzam Maqsood, Muzaffar Bashir, and Yang Shuyuan. A deep siamese convolution neural network for multi-class classification of alzheimer disease. *Brain Sciences*, 10(2), 2020.

[48] Ahsan Bin Tufail, Yong-Kui Ma, and Qiu-Na Zhang. Binary classification of alzheimer's disease using smri imaging modality and deep learning. *Journal of digital imaging*, 33:1073–1090, 2020.

[49] Christiane Möller, Hugo Vrenken, Lize Jiskoot, Adriaan Versteeg, Frederik Barkhof, Philip Scheltens, and Wiesje M van der Flier. Different patterns of gray matter atrophy in early-and late-onset alzheimer's disease. *Neurobiology of aging*, 34(8):2014–2022, 2013.

[50] Jing Yang, PingLei Pan, Wei Song, Rui Huang, JianPeng Li, Ke Chen, QiYong Gong, JianGuo Zhong, HaiChun Shi, and HuiFang Shang. Voxelwise meta-analysis of gray matter anomalies in alzheimer's disease and mild cognitive impairment using anatomic likelihood estimation. *Journal of the neurological sciences*, 316(1-2):21–29, 2012.

[51] Mingrui Xia, Jinhui Wang, and Yong He. Brainnet viewer: a network visualization tool for human brain connectomics. *PloS one*, 8(7):e68910, 2013.

[52] Muhammad Tariq Sadiq, Siuly Siuly, Ahmad Almogren, Yan Li, and Paul Wen. Efficient novel network and index for alcoholism detection from eegs. *Health Information Science and Systems*, 11(1):27, 2023.

[53] Y Li, P Wen, et al. Analysis and classification of eeg signals using a hybrid clustering technique. In *IEEE/ICME International Conference on Complex Medical Engineering*, pages 34–39. IEEE, 2010.