PAPER • **OPEN ACCESS**

Deep learning with guided attention for early diagnosis of Alzheimer's disease

To cite this article: Gia Minh Hoang *et al* 2025 *Phys. Scr.* **100** 065020

Physica Scripta

PAPER

Deep learning with guided attention for early diagnosis of Alzheimer's disease

Gia Minh Hoang , Youngjoo Lee and Jae Gwan Kim

Department of Biomedical Science and Engineering, Gwangju Institute of Science and Technology, Republic of Korea

E-mail: jaekim@gist.ac.kr

Keywords: Alzheimer's disease, deep learning, magnetic resonance imaging, medical imaging, convolutional neutral networks

OPEN ACCESS

RECEIVED
24 February 2025

REVISED
10 April 2025

ACCEPTED FOR PUBLICATION
30 April 2025

PUBLISHED
23 May 2025

Abstract

Alzheimer's Disease (AD) is one of the most common forms of neurodegenerative disease that involves the accumulation of amyloid beta plaques and tau tangles. The early diagnosis of AD is crucial as it helps patients to start preventive interventions to slow the disease's progression. We created a Guided-Attention Feature Extraction Deep Learning Network (GADL) for the early diagnosis of Alzheimer's disease (AD). We applied a GADL for the prediction of mild cognitive impairment (MCI) progression to AD and classification between MCI and cognitively normal (CN). We trained the model with magnetic resonance imaging images in the Alzheimer's Disease Neuroimaging Initiative (ADNI) database by subject-level data splitting and verified its generalizability in the Australian Imaging Biomarkers and Lifestyle Flagship Study of Aging (AIBL) database. Our method outperformed other subject-level studies with an accuracy of 80.29% for the prediction of MCI progression to AD and 83.70% for CN versus MCI classification in the ADNI dataset. The accuracies of our models when they were applied to the AIBL dataset are recorded as 79.38% and 79.83%, respectively. These results prove the high performance of our models in terms of its generalizability. The evaluation results showed that the proposed approach has competitive performance in comparison with recent studies in terms of its performance and generalizability. These results suggest that deep learning with guided attention can be an effective early diagnosis technique and a prognostic tool for Alzheimer's disease.

1. Introduction

Alzheimer's disease (AD) is one of the most common neurodegenerative diseases. Currently, more than 6.7 million Americans aged ⩾65 years suffer from AD and this number could increase to more than 13.8 million by 2060 [1]. Mild cognitive impairment (MCI) occurs before dementia and, without timely treatment, MCI can slowly progress to AD within 3 years [2]. Although AD pathology is irreversible and incurable with current treatments, there are systematic treatments available that may help manage symptoms or delay disease progression, especially in the MCI stage [3]. Therefore, early diagnosis of AD is particularly important for early preventive interventions to alleviate disease progression. The clinical diagnosis of AD relies on various types of clinical information, including medical history, neurological assessments, and behavioral tests. However, these early diagnostic techniques are inaccurate, with 18.18% misdiagnoses leading to inappropriate medication [4].

Neuroimaging plays an important role in discovering the pathological signs for the diagnosis of AD, such as brain atrophy in the hippocampus or temporal lobe, abnormal gray matter volume, and amyloid depositions [5, 6]. Magnetic resonance imaging (MRI) can effectively provide information about brain structures with a high contrast of soft tissues and high spatial resolution, allowing important AD-related biomarker extraction for early diagnosis [7, 8]. However, manual diagnosis by doctors using MRI scans is time-consuming and depends on the expertise of the clinician, which could lead to an inappropriate diagnosis. Therefore, computer-aided approaches based on MRI are crucial for the rapid and accurate diagnosis of AD.

Deep learning has demonstrated remarkable success in computer-aided diagnosis, particularly in medical imaging classification. Owing to evident differences in brain atrophy between individuals who are cognitively

normal (CN) and those with AD, numerous convolutional neural network approaches based on structural MRI have exhibited excellent diagnostic capabilities for CN and AD [9–12]. These methods have shown that essential brain structures, such as the hippocampus and amygdala, may be of critical importance in AD diagnosis. Progressive cerebral atrophy initially manifests in the medial temporal lobe, followed by the hippocampus, amygdala, and para-hippocampus [13]. However, the diagnosis of AD from CN is too late for effective early interventions compared to early-stage diagnosis.

Recently, many studies have shifted their attention toward developing deep-learning-based approaches to address the classification between MCI and CN, as well as predicting the progression from MCI to AD [14, 15]. These efforts aimed to identify the signs of cognitive decline earlier, allowing for timely intervention and treatment strategies. In 2023, EL-Geneedy *et al* [14] proposed an MRI-based deep learning approach to classify patients who are CN and those with MCI and stratify them into different dementia stages. Hoang *et al* [15] demonstrated that an MRI-based vision transformer approach could classify the progression of MCI to AD with high performance. However, these approaches have not considered duplicate information when using multiple images of the same patient in both the training and test datasets. Ignoring this could lead to data leakage, which could significantly affect the reliability of the results.

Wen *et al* [16] highlighted the issue of overly optimistic outcomes arising from data leakage in several AD classification studies, emphasizing that when defining the training, validation, or test datasets, we must split them at the subject level and not at the image or slice level. There are two common strategies to split MRI scan data: (1) slice-level/image-level split, in which different slices or scans of the same subjects are contained both in training and test datasets (data leakages will occur), and (2) subject-level split, the scans from the same subjects can only be contained in training or test datasets. When models are trained with data leakage, their generalizability becomes uncertain. Although they may yield considerable results in specific databases, their performance can decrease dramatically when evaluated in other databases. Therefore, the evaluation of a model using an independent dataset is essential for clinical applications.

In this study, we propose a Guided-Attention Feature Extraction Deep Learning Network (GADL) that can predict the progression of MCI to AD and classify CN versus MCI more efficiently and accurately using brain atrophy information generated by AD versus CN classifier on the presumption that the same brain regions will also be important for MCI classification. Our model overcomes the data leakage problem by splitting the training, validation, and test datasets at the subject level, thereby avoiding overly optimistic and inappropriate results.

2. Material and method

2.1. Materials

The image data used in this study were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) [17] and the Australian Imaging Biomarkers and Lifestyle Flagship Study of Aging (AIBL) [18]. The ADNI and AIBL are the largest clinical medical imaging databases used for the development of early diagnostic methods and the identification of related biomarkers of AD.

The data were divided into five groups: CN, MCI, AD, progressive MCI (pMCI), and stable MCI (sMCI). We defined pMCI as patients with MCI who converted to AD after 3 years, whereas sMCI included patients with MCI who remained in the MCI state after 3 years. For the ADNI database, 1.5 T/3 T three-dimensional (3D) T1-weighted structural MRI scans of patients in the ADNI-1, ADNI-2, or ADNI-3 phases were obtained. We collected only patient data with complete non-imaging information, such as age, apolipoprotein E (APOE) type, Mini-Mental State Examination (MMSE) score, or neurological evaluation, including 405 patients with AD, 760 who were CN, 1078 with MCI, 270 with pMCI, and 273 with sMCI. For the AIBL database, we also obtained 1.5 T/3 T 3D T1-weighted MRI scans from patients with fully essential non-imaging information, including 243 patients who were CN, 66 with MCI, 16 with pMCI, and 25 with sMCI. Table 1 presents the key demographic details of the participants from the ADNI and AIBL datasets.

To remove noise from data acquisition and enhance differences in brain structure, we used a brain image pre-processing protocol involving image registration and bias field correction using the Functional Magnetic Resonance Imaging of the Brain Software Library (FSL; Oxford University Innovation, Oxford, UK), which can be downloaded from https://fsl.fmrib.ox. ac.uk/. First, we reoriented images to match the orientation of the standard template images (Montreal Neurological Institute 152) with the 'fslreorient2std' function of the FSL. Because the acquisition protocols of MRI images among the ADNI and AIBL are different, the FSL Linear Image Registration Tool algorithm was used to align each image to the Montreal Neurological Institute standard template. Next, a bias field correction was applied using the N4ITK method [19]. The detailed T1-weighted MRI scan pre-processing pipeline is shown in figure 1.

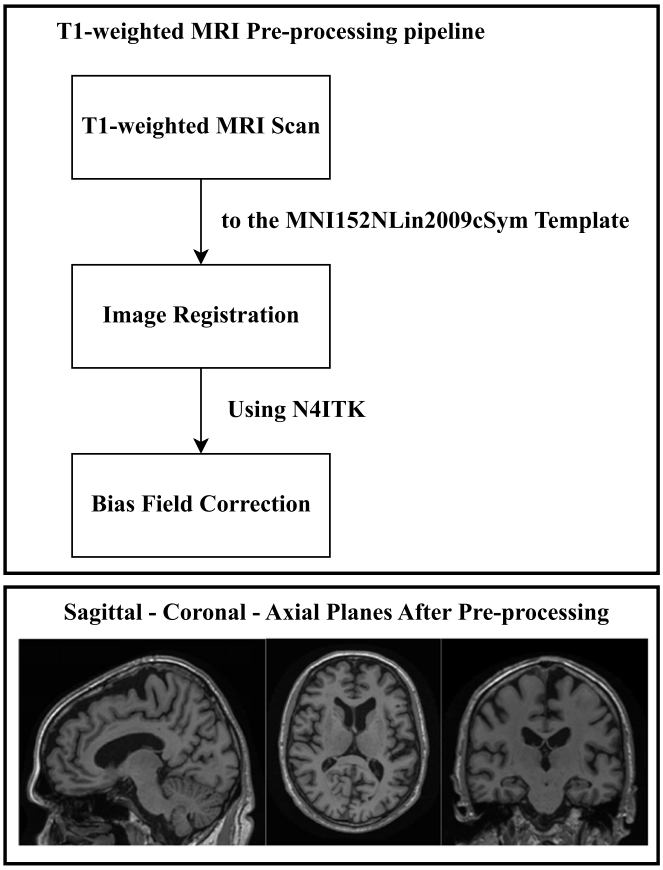

Figure 1. T1-weighted magnetic resonance imaging (MRI) pre-processing pipeline. The upper panel is a pre-processing protocol involving image registration and bias field correction. The lower panel is the sagittal, coronal, and axial plane of the MRI scan processed by our pipeline.

Table 1. Patient demographics of the Alzheimer's disease Neuroimaging Initiative (ADNI) and the Australian Imaging Biomarkers and Lifestyle Flagship Study of Aging (AIBL) database.

Dataset	Type	Gender	Age	MMSE
ADNI	CN	338 M/422 F	75.8 ± 6.9	27.05 ± 2.12
	MCI	628 M/450 F	74.3 ± 7.8	26.71 ± 2.61
	pMCI	156 M/114 F	74.2 ± 6.9	25.19 ± 2.11
	sMCI	166 M/107 F	74.8 ± 7.3	27.62 ± 2.00
	AD	227 M/178 F	75.7 ± 7.7	25.91 ± 1.26
AIBL	CN	96 M/147 F	74.9 ± 6.3	28.95 ± 2.33
	MCI	35 M/31 F	76.5 ± 7.0	26.65 ± 1.35
	pMCI	9 M/7 F	76.3 ± 5.4	24.98 ± 3.45
	sMCI	12 M/13 F	76.0 ± 6.5	27.13 ± 2.84

2.2. Overall architecture - GADL

In this study, we proposed a GADL for the early diagnosis of AD, including the prediction of the progression of MCI to AD and CN versus MCI classification. The overall architecture is illustrated in figure 2. Our approach contained two major components: a guided-attention feature extractor and a task classification model. Because MCI is an intermediate stage between CN and AD, brain atrophy that occurs in patients with AD also manifests in patients with MCI, albeit to a lesser extent. According to Calandrelli *et al* [20] there is evidence of cortical thickness and brain atrophy in regions similar to those affected by AD pathology, such as the middle temporal lobe, in patients with MCI who later progress to AD. Therefore, we guided our feature extractor by a pre-trained model using the CN versus AD task to learn the important brain structures that differ between CN and AD.

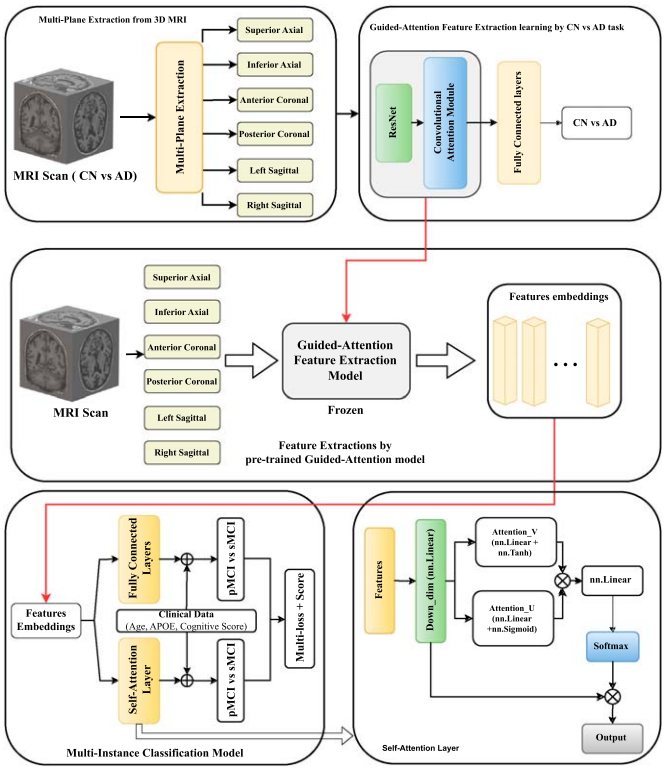

Figure 2. Guided-Attention Feature Extraction Deep Learning Network Architecture with two components: Guided-attention feature extractor and task classification model for specific task.

2.3. Guided-attention feature extractor component

In contrast to two-dimensional medical imaging, 3D MRI provides important pathological brain structural information in all three planes: axial, coronal, and sagittal. Directly applying the original 3D images with dimensions 193 × 229 × 193 requires a significant amount of computational time. Therefore, we extracted the original MRI scan into multiple sub-planes, such as left sagittal, right sagittal, and inferior axial, given that the original MRI scan, $I \in R^{B\times Sag\times Cor\times Ax}$, where $Sag \times Cor \times Ax$ is sagittal, coronal, and axial spatial resolution and the batch size is B, will be extracted to sub-plane $I_i' \in R^{B\times(w1\times H)\times(w2\times w)}$ with $w1 \times w2 = C$, where C, H, $andW$ are channels, height, and width of related planes ($Sag \times Cor \times Ax$ for sagittal-based planes, $Cor \times Sag \times Ax$ for coronal-based planes, and $Ax \times Sag \times Cor$ for axial-based planes).

In this study, we extracted 3D MRI scans into six sub-planes: left–right sagittal, superior–inferior axial, and anterior–posterior coronal, where

$$I_i' = Plane\ Extraction(I) \tag{1}$$

$$I' = [I_1', I_2', \ldots, I_6']. \tag{2}$$

Each sub-plane was trained individually with CN versus AD tasks to learn specific important brain domains in each plane using a convolutional neural network block D_{CNN}. The extracted features are $F_i \in R^{B\times 256}$, where

$$F_i = D_{CNN}(I_i') \tag{3}$$

$$F = [F_1, F_2, \ldots, F_6]. \tag{4}$$

The convolutional neural network block is a combination of the ResNet (Python Software Foundation, Wilmington, DE, USA) architecture and the convolutional attention module. The detailed architecture of the convolutional attention module is shown in figure 3. The dual-attention mechanism was first introduced by Woo *et al* [21] and consists of two components: channel and spatial attention. In this study, we applied dual attention to a convolutional attention module to extract spatial information from each sub-plane image. In the channel attention block, the maximum- and average-pooling layers were applied to aggregate the spatial information before using the shared multi-layer perceptron layer to compute the channel attention score by element-wise summation. In the spatial attention block, we aggregated channel information using two pooling operations (maximum and mean), followed by the generation of a spatial attention map by applying a convolution layer:

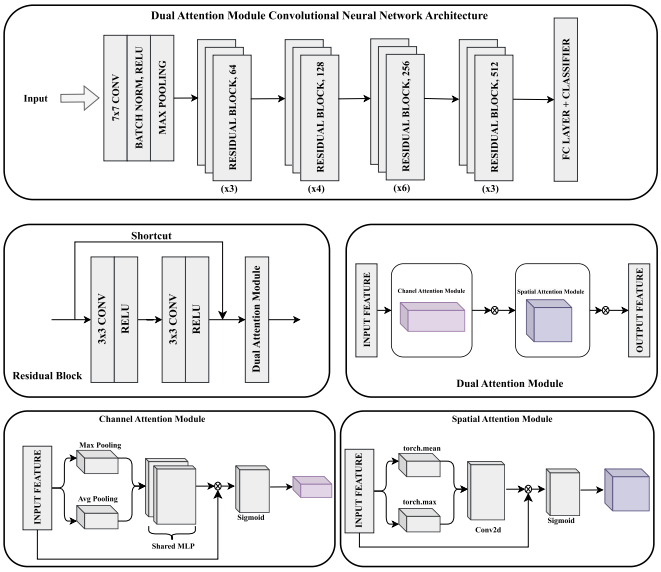

Figure 3. Convolutional block attention module with channel and spatial attention features extracted by the conventional neural network block, which then goes to the fully connected layer to classify cognitively normal (CN) and Alzheimer's Disease (AD) stages.

$$Y_i' = Fully\ Connected\ Layer(F_i) \tag{5}$$

$$Y' = torch.cat([Y_1', \quad Y_2', \ldots, Y_6']) \tag{6}$$

$$\mathscr{L}_{CN-AD} = CrossEntropyLoss(Y', Y) \tag{7}$$

where $Y' \in R^{B\times 6\times 256}$ is the output of the fully connected layer and Y is the target label for the CN versus AD classification task.

After training with the CN versus AD task, we froze the weight and used the convolutional neural network block as a pre-trained feature extractor in the task classification model to predict MCI progression and classify the CN versus MCI stages.

2.4. Classification model

In this component, we used the features extracted by the feature extractor to classify specific tasks, such as pMCI versus sMCI or CN versus MCI. The detailed architecture of the task classification model is shown in figure 2. The original MRI scan was extracted into six sub-planes using the same plane extraction as the previous component $I \in R^{B\times Sag\times Cor\times Ax} \rightarrow I'$: $[I_1', I_2', \ldots, I_6']$, $I_i' \in R^{B\times(w1\times H)\times(w2\times w)}$. Given that the convolutional neural network blocks are $E = [E_1, E_2, \ldots, E_6]$, in which E_i is the pre-trained feature extractor for each sub-plane, the features extracted by those models are $F = [F_1, F_2, \ldots, F_6]$, with $F_i \in R^{B\times 256}$:

$$F_i = E_i(I_i') \tag{8}$$

$$F = [F_1, F_2, \ldots, F_6] \rightarrow F' \in R^{B\times 6\times 256}. \tag{9}$$

After aggregating the features from $6 \times F_i \in R^{B\times 256}$ to $F' \in R^{B\times 6x256}$, we inputted those into dual-classifier architecture involving a fully connected layer and a self-attention classifier.

$$Y' = FullyConnected(F') \tag{10}$$

$$Y'' = SelfAttention(F'), \tag{11}$$

in which, Y' and $Y'' \in R^{B\times 2}$ are the classification score given by the fully connected layer and self-attention classifiers, respectively.

In the self-attention classifier, we aggregated feature embeddings using self-attention scores. After down sampling using the linear layer, we transformed each feature into two vectors, U and V:

$$F'' = nn.Linear(F') \tag{12}$$

$$\mathrm{U} = AttentionU(F'') \tag{13}$$

$$\mathrm{V} = AttentionV(F'') \tag{14}$$

$$W = nn.Linear(U \ * \ V) \tag{15}$$

$$Y'' = F'' \ * \ Softmax(W), \tag{16}$$

where *AttentionU* is the combination of *nn.Linear* and *nn.Sigmoid*; *AttentionV* is the combination of *nn.Linear* and *nn.Tanh*. After calculating the classifier scores using the dual classifier, the outputs are concatenated with normalized clinical data (APOE, MMSE, AGE, etc) to generate the final feature representation $(Y',\ Y'') \in R^{B\times 2} \oplus R^{B\times 5}$. This representation is subsequently fed into a classification module composed of a sequence of linear layers to perform disease classification.

$$Y'_{final} = Classifier\_1(Y') \tag{17}$$

$$Y''_{final} = Classifier\_2(Y''), \tag{18}$$

We estimated the loss of each classifier and overall loss:

$$\mathscr{L}_1 = CrossEntropyLoss(Y'_{final},\ Y) \tag{19}$$

$$\mathscr{L}_2 = CrossEntropyLoss(Y''_{final},\ Y) \tag{20}$$

$$\mathscr{L} = \alpha\mathscr{L}_1 + (1-\alpha)\mathscr{L}_2, \tag{21}$$

where Y'_{final} and $Y''_{final} \in R^{B\times 2}$ are classification logits and Y is the target label for the specific classification task, such as CN versus MCI or sMCI versus pMCI.

To address the issue of class imbalance in our dataset, we applied imbalance weighting directly within the loss function $\mathscr{L}_1$ and $\mathscr{L}_2$. Specifically, we used class weights inversely proportional to the frequency of each class, ensuring that minor classes had a higher impact on the loss calculation. This helped the model learn more effectively from all classes, rather than being biased toward the majority class. The weighted loss was computed as follows:

$$\mathscr{L}_{weighted} = \sum_{i=1}^{N} w_{y_i} \cdot \mathscr{L}(f(x_i),\ y_i) \tag{22}$$

where w_{y_i} is the weight assigned to the true class y_i of sample x_i and $\mathscr{L}$ is the base loss such as $\mathscr{L}_1$ and $\mathscr{L}_2$.

2.5. Experiment setting

In this study, we conducted experiments using one guided-attention task (CN versus AD) and two binary-classification tasks (sMCI versus pMCI and CN versus MCI). For the guided-attention task, the training and test sets for 80% and 20% of the total samples in the CN and AD datasets, respectively, all used subject-level splits. For other binary-classification tasks, we first split the total dataset into two parts: the test dataset and the training-validation dataset, with 20% and 80%, respectively, using subject-level splits. The training-validation dataset was then split into training and validation datasets using a five-fold cross-validation (CV) scheme. All the datasets were acquired from the ADNI database. The model was separately trained and evaluated for each classification task. For the CN versus MCI task, only cognitively normal (CN) and mild cognitive impairment (MCI) subjects were used. For the pMCI versus sMCI task, the model was trained and evaluated solely on MCI subjects, further divided into progressive MCI (pMCI) and stable MCI (sMCI). There is no sequential dependency between the two tasks; the output of the CN versus MCI classification does not feed into the pMCI versus sMCI task. Consequently, no misclassified samples from the first task are used in the second. This ensures that the performance of each classification task is not affected by the other.

To establish a baseline performance, we conducted an experiment using only the clinical data without incorporating any component of the proposed neural network. This experiment was performed using a logistic regression model, a widely used method for clinical prediction tasks [16, 20]. The results from this analysis reflect the performance of a traditional clinical approach and serve as a comparison point for evaluating the effectiveness of our proposed method. To evaluate the generalizability of our methods, we used CN, MCI, sMCI, and pMCI data from AIBL as an evaluation dataset. The GADL model was implemented using the PyTorch library [22]. The GADL was trained using an Adaptive Moment Estimation (Adam) as an optimizer for 100 epochs, with a learning rate of 1e-5, weight decay of 1e-4, and batch size of 8. Our experiments were performed on a machine with an Intel(R) Xeon(R) Gold 6258 R central processing unit @ 2.70 GHz with 256 GB random-access memory (Intel Corporation, Santa Clara, CA, USA). The training took approximately 20 h using a 4x NVIDIA GeForce RTX 3090 (Nvidia Corporation, Santa Clara, CA, USA).

The evaluation criteria used in this study were accuracy, sensitivity, specificity, and area under the curve as follows:

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{23}$$

$$Sen = \frac{TP}{TP + FN} \tag{24}$$

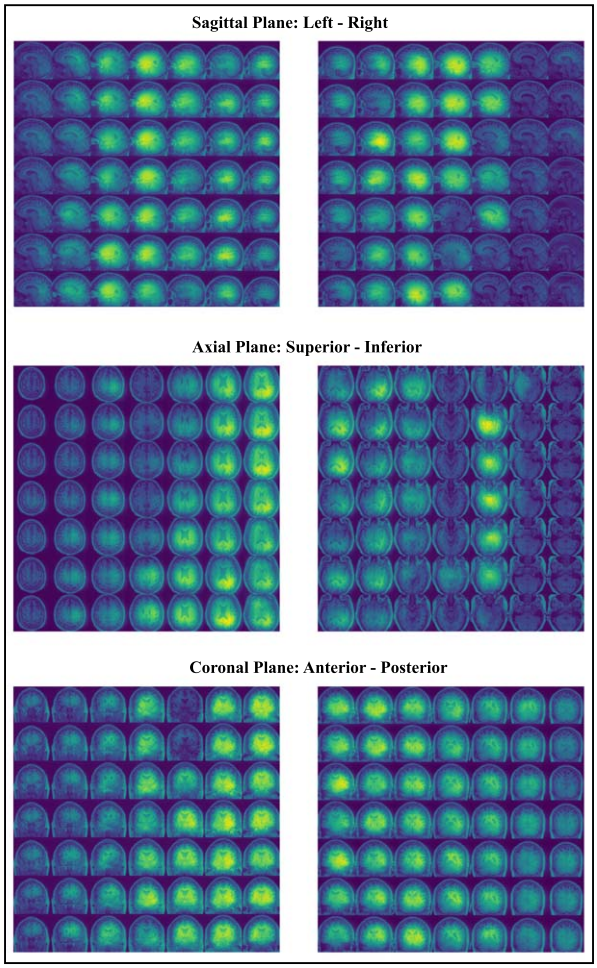

Figure 4. Visualization of important pathological locations of feature extraction by the guiding model. Upper panel: left-right sagittal planes. Middle panel: superior-inferior axial planes. Lower panel: anterior-posterior coronal planes.

$$Spec = \frac{TN}{TN + FP}, \tag{25}$$

where TP, TN, FP, and FN denote the true positive, true negative, false positive, and false negative values, respectively, and Acc, Sen, and Spec denote the accuracy, sensitivity, and specificity, respectively.

3. Results

3.1. Performance on guided-attention task with CN and AD classification

With the presumption that the same brain regions in CN versus AD will also be important for MCI classification, we guided our feature extractor by pre-training each sub-plane model using the CN versus AD task in the ADNI database. Our model achieved accuracies of 95.83%, 94.79%, 94.46%, 93.11%, 95.33%, and 94.74% using left-right sagittal, superior-inferior axial, and anterior-posterior coronal plane respectively. To ensure the reliability of the feature extractor model, we visualized the model activation heat map using explainable artificial intelligence. Figure 4 shows the highlighted regions measured using the attention score during model prediction. The upper panel of figure 4 shows the left and right sagittal planes, with attention scores highlighted in the medial temporal lobe, including the hippocampus and amygdala, which are among the first brain domains affected by AD [13]. In the middle panel, the guided-attention feature extraction model focuses on the gray matter and part of the temporal lobe in the superior and inferior axial planes. In the lower panel, the gray matter, hippocampus, and temporal lobe in the anterior and posterior coronal planes are highlighted using our model. These highlighted brain regions suggest that the features extracted by our model included essential brain structures for AD, such as the medial temporal lobe and hippocampus.

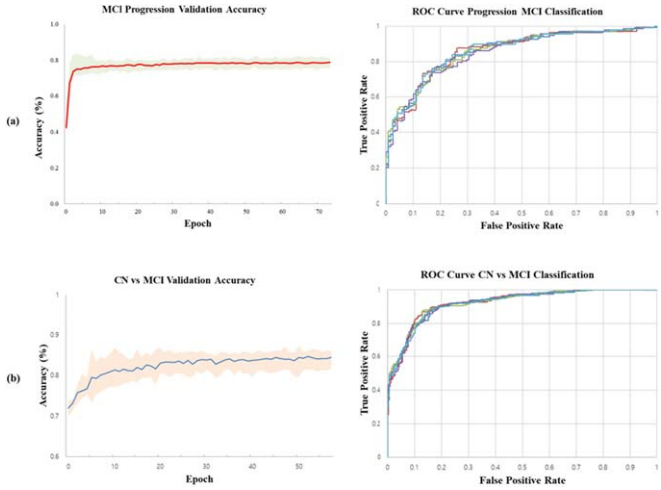

Figure 5. The five-fold cross-validation (CV) performance in the Alzheimer's disease Neuroimaging Initiative (ADNI) database for (a). the prediction of progression from mild cognitive impairment (MCI) to AD. (b). CN versus MCI classification tasks.

Table 2. Prediction performance of the progression of mild cognitive impairment (MCI) to Alzheimer's Disease in the ADNI dataset.

Study	Acc (%)	Sen (%)	Spec (%)	AUC (%)
Li *et al* [23]	71.1	—	—	76.19
Zhang *et al* [24]	73.68	65.91	79.05	72.1
Zhang *et al* [25]	76.88	77.25	76.5	78.45
Zhang *et al* [26]	73.2	53.26	82.96	68.06
Xin *et al* [27]	77.1	82.0	68.2	80.7
Guan *et al* [28]	73.54	69.46	76.55	75.7
Gao *et al* [29]	75.3	77.3	74.1	78.6
Hu *et al* [30]	77.2	79.97	71.59	81.53
Mulyadi *et al* [31]	67.85	—	—	75.67
Gao *et al* [32]	77.8	75.4	79.6	82.8
Clinical Data	71.09	77.11	65.94	74.97
GADL	80.29 ± 1.5	85.73 ± 2.1	74.43 ± 3.9	85.71 ± 0.4

3.2. Performance of MCI progression to AD prediction in the ADNI database

First, we predicted the progression of MCI to AD. Table 2 compares our classification performance with that of a previous study [23–32]. Using only clinical data, including age, APOE, MMSE, and neurological evaluation, the classification performance had 71.09% accuracy, 77.11% sensitivity, 65.94% specificity, and 74.94% AUC. When combining our GADL with clinical data, we achieved 80.29 ± 1.53% accuracy, 85.73 ± 2.13% sensitivity, 74.43 ± 3.9% specificity, and 85.71 ± 0.44% AUC. Our results outperformed those of current state-of-the-art studies using subject-level MRI scans for predicting the progression of MCI to AD [23–32]. The five-fold CV accuracy and receiver operating characteristic curves are shown in figure 5(a).

3.3. Performance of CN and MCI classification in the ADNI database

Second, we conducted a CN versus MCI experiment using the ADNI database. Table 3 shows the performance of the CN versus MCI classification in the ADNI database. Using only clinical data, the accuracy, sensitivity, specificity, and AUC were 78.2%, 72.28%, 83.07%, and 82.32%, respectively. When we applied the GADL to clinical data, we achieved 83.70 ± 1.47% accuracy, 85.97 ± 1.47% sensitivity, 79.93 ± 4.85% specificity, and 91.24 ± 0.11% AUC. Our approach consistently outperformed previous studies for all four indicators of classification performance [30, 31, 33–37]. The five-fold CV accuracy and receiver operating characteristic curves of the CN versus MCI classification tasks are shown in figure 5(b).

3.4. Generalization performance in the AIBL database

To verify the generalizability of our approach, we used the AIBL as an independent dataset to evaluate our model. The experimental performances for both pMCI versus sMCI and CN versus MCI are shown in table 4. The classification performance in pMCI versus sMCI task had 79.38 ± 1.70% accuracy, 76.66 ± 2.27% sensitivity, 87.30 ± 0.44% specificity, and 84.06 ± 1.07% AUC. For the CN versus MCI classification task, we

Table 3. Cognitively normal versus MCI classification performance in the ADNI dataset.

Study	Acc (%)	Sen (%)	Spec (%)	AUC (%)
Hu *et al* [30]	79.07	79.82	78.17	85.8
Mulyadi *et al* [31]	69.05	—	—	75.17
Hao *et al* [33]	76.1	85.53	62.79	75.82
Mora-Rubio *et al* [34]	66.41	65.3	65.91	—
Zhang *et al* [35]	71.26	74.61	67.96	71.47
Poloni and Ferrari [36]	75.58	72.94	77.81	83
Zhang *et al* [37]	73.77	89.13	50.41	73.14
Clinical Data	78.2	72.28	83.07	82.32
GADL	83.70 ± 1.4	85.97 ± 1.4	79.93 ± 4.8	91.24 ± 0.1

Table 4. Generalization performance of classification model in the AIBL datasets.

	pMCI versus sMCI	CN versus MCI
Accuracy (%)	79.38 ± 1.70	79.83 ± 7.22
Sensitivity (%)	76.66 ± 2.27	93.72 ± 3.02
Specificity (%)	87.30 ± 0.44	76.86 ± 9.39
AUC	84.06 ± 1.07	94.64 ± 0.18

Table 5. Performance of the progression of MCI to AD and CN versus MCI classification in the ADNI database by Slice-level data splitting.

	CN versus MCI	pMCI versus sMCI
Training Accuracy (%)	100 ± 0.00	99.16 ± 0.95
Validation Accuracy (%)	98.76 ± 0.57	98.30 ± 0.38
Test Accuracy (%)	71.63 ± 2.32	64.24 ± 1.21

achieved $79.83 \pm 7.22\%$ accuracy, $93.72 \pm 3.02\%$ sensitivity, $76.86 \pm 9.39\%$ specificity, and $94.64 \pm 0.18\%$ AUC. Compared to the performance in the ADNI database, our method showed no clear decline in performance for pMCI versus sMCI and CN versus MCI in most of the metrics.

3.5. Ablation study

To investigate the effect of our proposed approach including subject-level splitting and guided-attention mechanism, we conducted two ablation experiments. In the first experiments, we split data in training and validation datasets by slice level while the test dataset will be kept as an independent dataset. The results are shown in table 5. We observed that the model will return over-optimistic performance when training and validating in slice-level datasets. Therefore, when we evaluated them in independent test datasets, the performance dropped significantly. By training with subject-level, our approach could learn more informative features and overcome the limitations. In the second experiment, to validate the effectiveness of transferring the CN versus AD model to the pMCI versus sMCI classification task and other components, we conducted an ablation study comparing following training strategies: (a) using the pretrained CN versus AD model and fine-tuning it on pMCI versus sMCI data (Pretrained), (b) training a model from scratch directly on the pMCI versus sMCI classification task (From Scratch), (c) training the model without the dual attention module, and (d) training the model without the support of clinical data. As shown in table 6, the Pretrained model outperforms the model trained from scratch, demonstrating the benefit of leveraging knowledge learned from the CN versus AD classification task. Moreover, without dual-attention, and clinical data, the model performance decrease significantly ensuring the effectiveness of those components.

4. Discussion

Early-stage diagnosis and progression classification are crucial for AD treatment because it allows patients to start early intervention to delay disease progression or treat disease symptoms. However, traditional early diagnostic techniques using clinical data are inaccurate and lead to inappropriate medications, and computer-aided approaches with high accuracy and generalizability are lacking. In this study, we proposed a GADL for the

Table 6. Ablation study of the progression of MCI to AD in the ADNI database.

Pretrained	Dual attention	Clinical data	ACC	SEN	SPEC
✓	✓	✓	80.29 ± 1.5	85.73 ± 2.1	74.43 ± 3.9
✗	✓	✓	77.17 ± 2.4	80.00 ± 1.7	69.02 ± 2.8
✓	✗	✓	76.38 ± 1.6	79.28 ± 1.5	74.34 ± 2.0
✓	✓	✗	78.74 ± 2.6	85.71 ± 1.8	74.00 ± 3.3

early-stage diagnosis and progression classification of AD. We conducted two tasks: predicting the progression of MCI to AD and CN versus MCI classification, with considerable performance and high generalizability.

Our model contained two major components: a guided-attention feature extractor and a task classification model. In the guided attention feature extractor, we hypothesized that the convolution neural networks block can learn important brain atrophy regions to extract the necessary information by training with CN versus AD tasks. Brain atrophy occurs in patients with AD, especially in the grey matter and the temporal lobe, involving the hippocampus and amygdala [13], which can be easily recognized in MRI images. Because MCI is the intermediate stage between CN and AD, these brain atrophy regions also occur in MCI, but at a lower level.

However, without any guidance, the deep learning model cannot determine the brain regions that need to be focused on, which can lead to an overfitting of the model. By guiding the feature extractor with the results from the CN versus AD tasks, we can tell the model which brain structures or features are important as AD pathological biomarkers. To ensure the interpretability of our methods, we visualized the activation heat map of the model during prediction. As shown in figure 4, our model focused on the medial temporal lobe, particularly the hippocampus and amygdala, in agreement with other studies on the pathological regions of AD [13, 20, 38]. These results indicated that our feature extractor could extract important brain structures related to AD. A possible reason for this is that the dual-attention mechanism in the convolution attention module could effectively enhance and recognize the spatial differences in the brain regions between the AD and CN groups.

Using the pre-trained feature extractor, we extracted the necessary information and trained the classification model for specific tasks, such as sMCI versus pMCI or CN versus MCI classification. As shown in tables 2 and 3, the proposed method outperformed the other methods in terms of accuracy, sensitivity, specificity, and AUC for both classification tasks. These results supported our hypothesis that the brain atrophy features extracted with guiding tasks can be used to diagnose AD at an early stage as well as classify the progression of AD.

Data leakage is a major problem in medical imaging-based deep-learning studies, particularly in AD diagnosis. This can lead to overfitting in training or overoptimistic performance [16]. The main reason for data leakage is incorrect data splitting. In the ADNI and AIBL databases, each patient could have multiple time-point data. In addition, each image data point can be extracted into multi-slice two-dimensional images. Therefore, if we split the training, validation, and test datasets by slice or image levels, information that specifically represents the patient instead of the disease could occur in both the training and validation sets. The model learns that information, which leads to an overly optimistic performance. In this study, to overcome data leakage limitations, we split our dataset by subject level. We also checked for data leakage in all fold splitting by five-fold CV.

In computer-aided diagnosis, the generalizability of a model is an important criterion. Without a generalization evaluation, the model could yield considerable results in a specific database; however, the performance may decline dramatically when applied to other databases. This makes it difficult to apply the model to new data, which limits its clinical application. We evaluated our method using the AIBL database, an independent dataset with different national cohorts, and the ADNI database to verify its robustness and generalizability. As shown in table 4, the proposed approach could predict both classification tasks without an obvious decline in performance. These results suggest that the proposed GADL method has favorable generalizability and robustness across different datasets.

Although our GADL approach achieved favorable performance in both the ADNI and AIBL databases with both classification tasks (sMCI versus pMCI and CN versus MCI), there were still several limitations to this method. In other transfer learning tasks in computer vision, the feature extractor is typically trained using millions or hundreds of millions of images. However, the feature extractor in our method was trained only with limited data, which could affect the interpretability of the model regarding the important brain structures that need to be focused on. In the future, we plan to pre-train our feature extractor using more databases, in addition to the ADNI database, to overcome this limitation. Second, the spatial information extracted by the feature extractor can be further improved in the future with the development of computer vision techniques, especially vision transformers.

5. Conclusion

In conclusion, this paper proposed a novel method, the GADL, for the diagnosis of early stage and progressive AD, which includes two major components: (1) the guided-attention feature extractor extracts important brain atrophy information, guided by patients who are CN, with MCI AD tasks and (2) a task classification model to classify specific binary-classification tasks, such as patients with sMCI versus pMCI or patients who are CN versus patients with MCI based on the features extracted by the feature extractor. The proposed GADL approach was evaluated using two independent databases (the ADNI and AIBL) with subject-level data splitting to ensure model generalizability. Experimental results indicated that our GADL with brain information extracted by the feature extractor not only outperformed recent state-of-the-art subject-level studies in ADNI and AIBL datasets but also achieved high generalizability across different datasets.

Availability of data and material

This research uses data from the Alzheimer's Disease Neuroimaging Initiative (ADNI), which can be downloaded at https://adni.loni.usc.edu/ .

Abbreviations

ADNI: Alzheimer's Disease Neuroimaging Initiative; AIBL: Australian Imaging Biomarkers and Lifestyle Flagship Study of Aging; CV: cross-validation; FSL: Functional Magnetic Resonance Imaging of the Brain Software Library; GADL: Guided-Attention Feature Extraction Deep Learning Network.

Funding

This work was supported by National Research Foundation of Korea (NRF) grants funded by the Korean government (MSIT) [grant numbers NRF-2022R1A2C3009749, NRF-2022K1A3A1A20014975) and the Healthcare Artificial Intelligence Convergence Research & Development Program [grant number S1502-24-1004) through the National IT Industry Promotion Agency of Korea (NIPA).

Ethical statement

The authors declare that the work reported in this manuscript is original and has not been published elsewhere, nor is it currently under consideration for publication by any other journal. All authors have contributed significantly to the research and have approved the final version of the manuscript. The research was conducted in accordance with ethical standards, and no human participants or animals were involved, thus no ethical approval was required.

Data availability statement

The data that support the findings of this study are openly available at the following URL/DOI:https://adni.loni.usc.edu/.

ORCID iDs

Gia Minh Hoang https://orcid.org/0000-0002-8494-0096
Jae Gwan Kim https://orcid.org/0000-0002-1010-7712

References

[1] Anon 2023 2023 Alzheimer's disease facts and figures *Alzheimer's Dementia* **19** 1598–695
[2] Palmer K, Bäckman L, Winblad B and Fratiglioni L 2008 Mild cognitive impairment in the general population: occurrence and progression to alzheimer disease *The American Journal of Geriatric Psychiatry* **16** 603–11
[3] Robinson L, Tang E and Taylor J-P 2015 Dementia: timely diagnosis and early intervention *Brit. Med. J.* **350** h3029
[4] Gaugler J E, Ascher-Svanum H, Roth D L, Fafowora T, Siderowf A and Beach T G 2013 Characteristics of patients misdiagnosed with Alzheimer's disease and their medication use: an analysis of the NACC-UDS database *BMC Geriatr* **13** 137

[5] Chen X, Wang T, Lai H, Zhang X, Feng Q and Huang M 2022 Structure-constrained combination-based nonlinear association analysis between incomplete multimodal imaging and genetic data for biomarker detection of neurodegenerative diseases *Med. Image Anal.* **78** 102419
[6] Anon multi-band brain network analysis for functional neuroimaging biomarker identification | IEEE Journals & Magazine | IEEE Xplore
[7] Apostolova L G and Thompson P M 2008 Mapping progressive brain structural changes in early Alzheimer's disease and mild cognitive impairment *Neuropsychologia* **46** 1597–612
[8] Suk H-I, Lee S-W, Shen D and The Alzheimer's Disease Neuroimaging Initiative 2015 Latent feature representation with stacked auto-encoder for AD/MCI diagnosis *Brain Struct Funct* **220** 841–59
[9] Liu S, Masurkar A V, Rusinek H, Chen J, Zhang B, Zhu W, Fernandez-Granda C and Razavian N 2022 Generalizable deep learning model for early Alzheimer's disease detection from structural MRIs *Sci Rep.* **12** 17106
[10] Raza M, Awais M, Ellahi W, Aslam N, Nguyen H X and Le-Minh H 2019 Diagnosis and monitoring of alzheimer's patients using classical and deep learning techniques *Expert Syst. Appl.* **136** 353–64
[11] Feng X, Provenzano F A, Small S A and for the Alzheimer's Disease Neuroimaging Initiative 2022 A deep learning MRI approach outperforms other biomarkers of prodromal alzheimer's disease *Alzheimer's Research & Therapy* **14** 45
[12] Guan H, Liu Y, Yang E, Yap P-T, Shen D and Liu M 2021 Multi-site MRI harmonization via attention-guided deep domain adaptation for brain disorder identification *Med. Image Anal.* **71** 102076
[13] 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
[14] EL-Geneedy M, Moustafa H E-D, Khalifa F, Khater H and AbdElhalim E 2023 An MRI-based deep learning approach for accurate detection of Alzheimer's disease *Alexandria Engineering Journal* **63** 211–21
[15] Hoang G M, Kim U-H and Kim J G 2023 Vision transformers for the prediction of mild cognitive impairment to Alzheimer's disease progression using mid-sagittal sMRI *Front. Aging Neurosci.* **15**
[16] Wen J, Thibeau-Sutre E, Diaz-Melo M, Samper-González J, Routier A, Bottani S, Dormont D, Durrleman S, Burgos N and Colliot O 2020 Convolutional neural networks for classification of Alzheimer's disease: overview and reproducible evaluation *Med. Image Anal.* **63** 101694
[17] Anon ADNI | Alzheimer's Disease Neuroimaging Initiative
[18] Anon The AIBL study - aibl.org.au *AIBL*
[19] Tustison N J, Avants B B, Cook P A, Zheng Y, Egan A, Yushkevich P A and Gee J C 2010 N4ITK: Improved N3 Bias Correction *IEEE Trans. Med. Imaging* **29** 1310–20
[20] Calandrelli R, Panfili M, Onofrj V, Tran H E, Piludu F, Guglielmi V, Colosimo C and Pilato F 2022 Brain atrophy pattern in patients with mild cognitive impairment: MRI study *Transl Neurosci* **13** 335–48
[21] Woo S, Park J, Lee J-Y and Kweon I S 2018 CBAM: convolutional block attention module
[22] Paszke A *et al* 2019 PyTorch: An Imperative Style *High-Performance Deep Learning Library*
[23] Li Y, Yang B, Pan D, Zeng A, Wu L and Yang Y 2023 Early diagnosis of alzheimer's disease based on multimodal hypergraph attention network 2023*IEEE International Conference on Multimedia and Expo (ICME) 2023* (IEEE International Conference on Multimedia and Expo (ICME) 192–7
[24] Zhang Y, He X, Liu Y, Ong C Z L, Liu Y and Teng Q 2023 An end-to-end multimodal 3D CNN framework with multi-level features for the prediction of mild cognitive impairment *Knowl.-Based Syst.* **281** 111064
[25] Zhang Z, Gao L, Li P, Jin G and Wang J 2023 DAUF: A disease-related attentional UNet framework for progressive and stable mild cognitive impairment identification *Comput. Biol. Med.* **165** 107401
[26] Zhang Y, He X, Chan Y H, Teng Q and Rajapakse J C 2023 Multi-modal graph neural network for early diagnosis of Alzheimer's disease from sMRI and PET scans *Comput. Biol. Med.* **164** 107328
[27] Xin J, Wang A, Guo R, Liu W and Tang X 2023 CNN and swin-transformer based efficient model for Alzheimer's disease diagnosis with sMRI *Biomed. Signal Process. Control* **86** 105189
[28] Guan H, Yue L, Yap P-T, Xiao S, Bozoki A and Liu M 2023 Attention-guided autoencoder for automated progression prediction of subjective cognitive decline with structural MRI *IEEE Journal of Biomedical and Health Informatics* **27** 2980–9
[29] Gao X, Shi F, Shen D and Liu M 2023 Multimodal transformer network for incomplete image generation and diagnosis of Alzheimer's disease *Comput. Med. Imaging Graph.* **110** 102303
[30] Hu Z, Wang Z, Jin Y and Hou W 2023 VGG-TSwinformer: transformer-based deep learning model for early Alzheimer's disease prediction *Comput. Methods Programs Biomed.* **229** 107291
[31] Mulyadi A W, Jung W, Oh K, Yoon J S, Lee K H and Suk H-I 2023 Estimating explainable Alzheimer's disease likelihood map via clinically-guided prototype learning *NeuroImage* **273** 120073
[32] Gao X, Shi F, Shen D and Liu M 2022 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** 36–43
[33] Hao X, Li J, Ma M, Qin J, Zhang D and Liu F 2024 Hypergraph convolutional network for longitudinal data analysis in Alzheimer's disease *Comput. Biol. Med.* **168** 107765
[34] Mora-Rubio A, Bravo-Ortíz M A, Arredondo S Q, Torres J M S, Ruz G A and Tabares-Soto R 2023 Classification of Alzheimer's disease stages from magnetic resonance images using deep learning *PeerJ Comput. Sci.* **9** e1490
[35] Zhang J, He X, Liu Y, Cai Q, Chen H and Qing L 2023 Multi-modal cross-attention network for Alzheimer's disease diagnosis with multi-modality data *Comput. Biol. Med.* **162** 107050
[36] Poloni K M and Ferrari R J 2022 Automated detection, selection and classification of hippocampal landmark points for the diagnosis of Alzheimer's disease *Comput. Methods Programs Biomed.* **214** 106581
[37] Zhang J, He X, Qing L, Chen X, Liu Y and Chen H 2023 Multi-relation graph convolutional network for Alzheimer's disease diagnosis using structural MRI *Knowl.-Based Syst.* **270** 110546
[38] Anon MRI measures of entorhinal cortex vs hippocampus in preclinical AD | Neurology