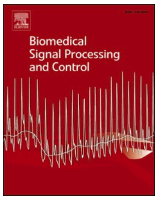
Biomedical Signal Processing and Control 88 (2024) 105669

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

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



A hierarchical attention-based multimodal fusion framework for predicting the progression of Alzheimer’s disease

Peixin Lu [c,d,*], Lianting Hu [a,b], Alexis Mitelpunkt [h,i], Surbhi Bhatnagar [d], Long Lu [c,e,f,g,*], Huiying Liang [a,b,*]

[a] *Medical Big Data Center, Guangdong Provincial People’s Hospital (Guangdong Academy of Medical Science), Guangzhou, China*
[b] *Guangdong Cardiovascular Institute, Guangdong Provincial People’s Hospital (Guangdong Academy of Medical Science), Guangzhou, China*
[c] *School of Information and Management, Wuhan University, Wuhan, China*
[d] *Division of Biomedical Informatics, Cincinnati Children’s Hospital Medical Center, Cincinnati, OH, USA*
[e] *Institute of Pediatrics, Guangzhou Women and Children’s Medical Center, Guangzhou Medical University, Guangzhou, Guangdong, China*
[f] *The Center for Healthcare Big Data Research, Big Data Institute, Wuhan University, Wuhan, China*
[g] *School of Public Health, Wuhan University, Wuhan, China*
[h] *Pediatric Rehabilitation, Department of Rehabilitation, Dana-Dwek Children’s Hospital, Tel Aviv Medical Center, Tel Aviv, Israel*
[i] *Sackler Faculty of Medicine, Tel Aviv University, Tel Aviv, Israel*

ARTICLE INFO

Keywords:
Early detection
Alzheimer’s disease
Multimodal fusion
Hierarchical attention

ABSTRACT

Early detection and treatment can slow the progression of Alzheimer’s Disease (AD), one of the most common neurodegenerative diseases. Recent studies have demonstrated the value of multimodal fusion in early AD detection. However, most approaches to this have failed to consider data modality domains, their relationships, and variations in their relative importance. To address these challenges, we propose a Hierarchical Attention-Based Multimodal Fusion framework (HAMF) that utilizes imaging, genetic and clinical data for early AD detection. In the HAMF model, attention mechanisms are utilized to learn the appropriate weights for each modality and to understand the interaction between modalities through hierarchical attention. HAMF performs better than state-of-the-art methods, achieving an accuracy of 87.2% and an AUC of 0.913, which are superior to unimodal models. By comparing the results of different unimodal and multimodal models, we find that multimodal fusion can improve model performance more than unimodal models and clinical data is the most important modality. Our ablation experiment confirmed the effectiveness of HAMF. Finally, we used SHapley Additive exPlanations (SHAP) to improve the model’s interpretability. We provide the model as a guide for future research in the field, and as a framework for generating actional advice and decision support system for clinical practitioners.

1. Introduction

Alzheimer’s Disease (AD) is one of the most common and severe neurodegenerative diseases [12]. By 2050, AD and other types of dementia are expected to affect at least 131 million people, according to the Alzheimer’s Association [3]. The disease’s primary cause is the accumulation of abnormal protein deposits, primarily beta-amyloid plaques and tau tangles, in the brain. These protein aggregates disrupt neural communication and lead to the characteristic cognitive decline associated with AD [4]. Mild cognitive impairment (MCI) is a condition that may or may not progress to Alzheimer’s disease (AD) or other types of dementia. While MCI is considered an early stage of AD, it is important to note that not all MCI cases will develop into AD. In fact, studies have shown that only 10–12 % of patients with MCI progress to AD each year [5]. MCI can be classified as either progressive (pMCI) or stable (sMCI) depending on its conversion to AD in a specific period. Although AD cannot be cured effectively, some studies have shown that early detection and intervention can slow its progression [6,7]. Therefore, early detection of AD such as predicting MCI conversion to AD has become an increasing focus for ongoing research.

As the deep learning (DL) models show substantial performances in a wide array of clinical decision support systems, several studies have

[*] Corresponding authors at: School of Information and Management, Wuhan University, Wuhan, China (P. Lu and L. Lu), Medical Big Data Center, Guangdong Provincial People’s Hospital (Guangdong Academy of Medical Science), Guangzhou, China (H. Liang).
E-mail addresses: lupx@whu.edu.cn (P. Lu), bioinfo@gmail.com (L. Lu), lianghuiying@hotmail.com (H. Liang).

https://doi.org/10.1016/j.bspc.2023.105669
Received 4 June 2023; Received in revised form 14 September 2023; Accepted 29 October 2023
Available online 7 November 2023

focused on developing DL-based models for predicting the conversion of MCI to AD using various unimodal medical data, such as Magnetic Resonance Imaging (MRI), Positron Emission Tomography (PET), clinical data as well as genetic information [9–11]. Nevertheless, a unimodal approach does not suffice for clinical decision-making, which frequently requires people to consider information from multiple sources. As a practical example, when diagnosing complex disorders, clinicians not only consult clinical records and medical histories but also perform pathological or medical imaging tests to determine a diagnosis more accurately.

Multimodal information can provide a more comprehensive view and can improve the accuracy of classification and prediction. For example, MRI, cerebrospinal fluid markers, and Single Nucleotide Polymorphisms (SNPs) can be used to obtain information about brain morphology, cerebrospinal pathology, and genetic information related to AD. Multimodal fusion is currently being successfully used in a wide range of domains, such as sentiment analysis, object recognition, and image segmentation [12,13]. Several models for diagnosing AD or predicting MCI to AD conversion based on multimodal fusion have been developed to increase the accuracy and effectiveness of diagnosis and prediction [14–22]. Despite some advancements and successes, existing studies have neglected the importance of cross-modal interaction and the varying importance of each modality in different tasks which are critical aspects of multimodal learning and can increase interpretability. Most existing research simply concatenates or maximizes features extracted from multiple modalities [14–22], which may result in the loss of potentially important information and make it challenging to generate an effective cross-modal representation. Additionally, most existing studies used a limited number of modalities, or just simply used core genes, such as APOE4, to represent the complicated genetic modality [18,20,21].

To address the limitations of existing studies, we propose a Hierarchical Attention-Based Multimodal Fusion framework (HAMF). By considering the relative importance of different modalities for the target task, HAMF can simulate the actual decision-making process to prioritize relevant modalities. Additionally, the hierarchical attention mechanism allows us to learn cross-modal representations across all modalities. Performance of the proposed fusion approach will be evaluated via predicting the conversion of MCI to AD. The specific process is shown in Fig. 1. This paper offers the following innovations.

1) HAMF is proposed to adequately explore the cross-modal interactions and enhance relevant information expression, while

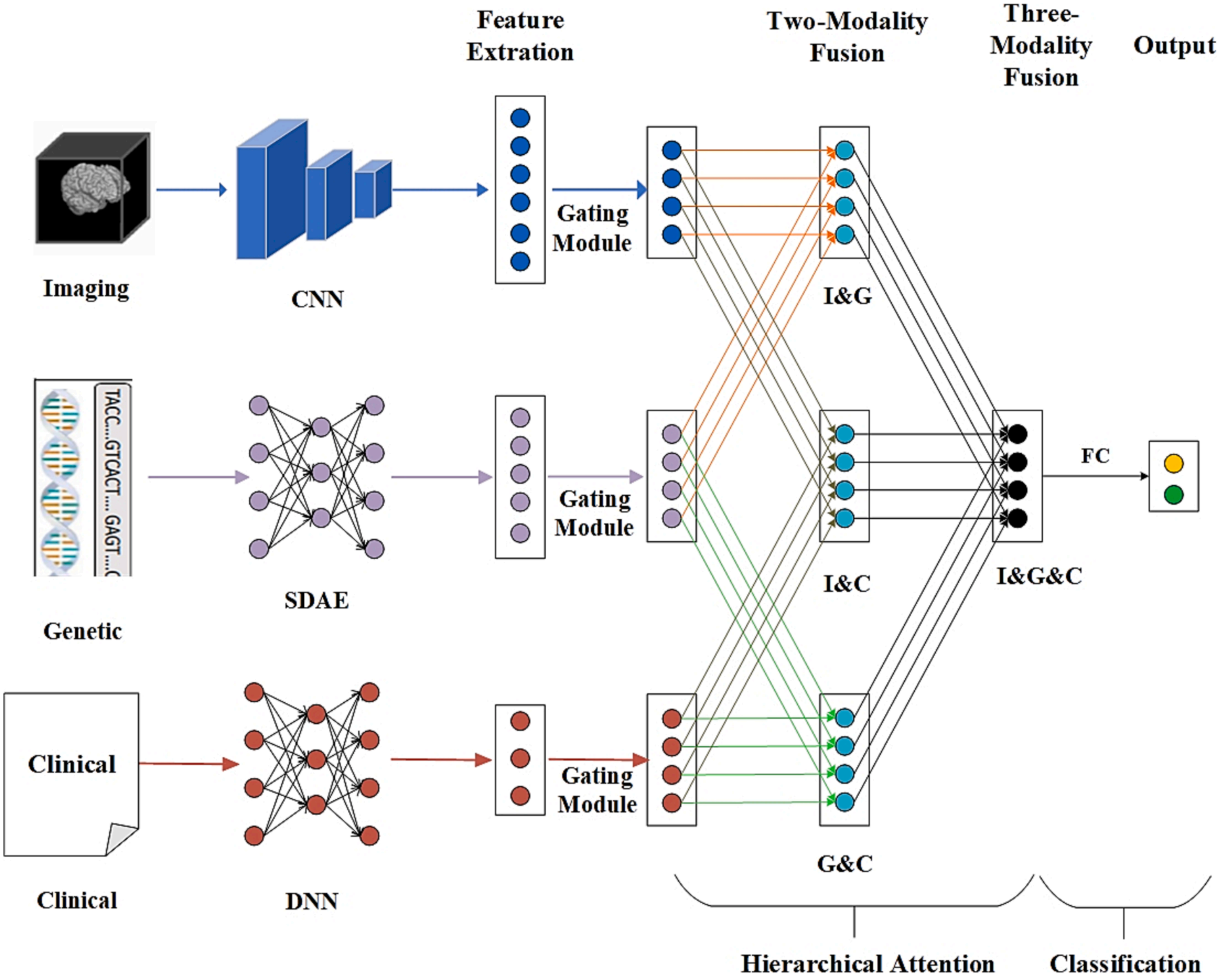


Fig. 1. The Structure of Hierarchical Attention based Multimodal Fusion framework (HAMF). Specifically, features of different modalities are extracted in the first step. The second step involves mapping the feature representations of different modalities into the same latent space using the nonlinear gating module. As a final step, using hierarchical attention, the model can infer both the optimal weights of different modalities dynamically and learn the cross-modal representation. CNN: Convolutional Neural Networks; SDAE: The Stacked Denoising Autoencoder; DNN: Deep Neural Network; I: Medical Imaging; G: Genetic; C: Clinical.

suppress redundant information, and finally achieve effective multimodal fusion.

2) Our methods achieved state-of-the-art performance compared to previous studies. We also show that multimodal fusion can have superior performance compared to unimodal models.
3) A comprehensive evaluation of the importance of each modality or combination for predicting MCI to AD conversion is presented, which identified the most appropriate modality and optimal combination for prediction and proved the necessity of multimodal fusion. To the best of our knowledge, this method represents a pioneering effort in accounting for the varying significance of each modality in predicting the conversion from MCI to AD.

This paper is structured as follows: Section 2 provides an overview of related works in the field, Section 3 delves into a detailed description of both the materials used for evaluation and our proposed methodology, Section 4 reports the experimental results, Section 5 presents in-depth discussions of our findings, and, finally, Section 6 outlines our conclusions. In addition to addressing the cause of AD, our research emphasizes early detection and intervention, with a specific focus on predicting the conversion of Mild Cognitive Impairment (MCI) to AD, which has become a central area of investigation in the ongoing quest to combat this debilitating disease.

2. Related work

Despite recent advances in medical and computer science, early detection of AD remains a challenge in the AD diagnosis research field. To improve the accuracy and reliability of AD detection, simple unimodal methods need to be augmented with multimodal approaches that combine multiple types of data and analysis techniques. A detailed explanation of AD detection techniques follows in the remainder of this section.

2.1. Single-modal based AD diagnosis

Early detection of AD is often based on medical imaging [23,24], electroencephalography (EEG) [25], clinical data [26], and genetic data [27,28]. Initially, machine learning-based AD detection models were often built based on unimodal data [9,29,30]. Most of these researchers first run feature selection methods to extract important features such as hippocampal volume, and surface area from the original data, and then establish a classification model to detect or predict AD [29,30]. However, there are some shortcomings in such studies. First, the model based on the feature selection method assumes in advance that the selected feature is the most informative, which may not cover the whole picture of the data. Secondly, since feature selection and classification are performed separately, it will cause further loss of useful information [31].

Deep learning has demonstrated remarkable results compared to traditional machine learning [32,33]. One advantage of deep learning model is to bypass the feature selection step which usually require specific domain knowledge. When combining with a tunable loss function, it can be used by experts in non-medical fields for their research or applications. For example, Convolutional Neural Network (CNN) has achieved good performance in the classification or prediction tasks of AD [23,24].

2.2. Multimodal based methods for AD diagnosis

With the development of deep learning, multimodal fusion has been developed and increased the accuracy of diagnosis and prediction. Despite some advancements and successes, there still have some limitations.

The first limitation with fusion is the neglect of cross-modal interaction and the importance of each modality in different tasks. Fusion data depicts instances more robustly and comprehensively by combining complementary information from multimodal information and highlighting key elements. There are two main multimodal fusion strategies; the first approach is average fusion [15], which gives equal weight to information from different sources. As a result of the average fusion approach, each source of information contributes equally to the target task, which may lead to some significant potential information being lost. Moreover, this is not consistent with the actual decision-making scenario as different information contains different amounts of information for the actual decision (i.e. PET provides more information than MRI for the detection of bone metastases [34]). The other fusion strategy is maximization fusion [35], which maximizes the most relevant information among all information. In this approach, only the most relevant information is considered, while other information is ignored. To achieve effective and adequate multimodal fusion, it is important to differentiate the weighting based on the importance of different modalities.

While many existing methods adopt a straightforward approach to integrate multimodal information by simply concatenating features extracted from different modalities, there is a growing interest in leveraging attention mechanisms to fuse multimodal data [36–38]. For instance, Zhang et al. introduced a novel multi-modal cross-attention framework for Alzheimer's Disease (AD) diagnosis [36]. However, a common challenge in these studies is the limited availability of input data for multimodal fusion. To illustrate, in the work by Zhang et al., the cross-attention framework was applied to fuse MRI, PET, and CSF data for AD diagnosis, yet this approach did not incorporate valuable clinical and genetic information [36]. Many multimodal fusion models for AD are based on the fusion of two modalities or only use the APOE4 to represent the full genetic information. For example, Kang et al. proposed a multimodal fusion framework based on transfer learning for AD prediction using two modalities of medical imaging data, MRI and diffusion tensor imaging [39]. Zhou et al. mapped two types of medical imaging data, MRI and PET to a common multimodal space using nonlinear mapping, and finally used SVM for the prediction of AD [40]. Similarly, Lahmiri et al. used MRI and cognitive assessment scores for classification prediction based on SVM [41]. Khatri and Kwon used MRI, cognitive scores, cerebrospinal fluid biomarkers, and APOE4 for AD prediction based on feature selection as well as extreme machine learning [21]. To better explore the performance of different multimodal fusions on AD prediction models, it is important that studies should incorporate more modalities with well-represented input features.

3. Materials and method

3.1. Dataset and data preprocessing

In this study, we used the ADNI (Alzheimer's Disease Neuroimaging Initiative) database, which is the largest and most widely used AD database. Launched in 2003 by several public and private organizations, the ADNI study aims to identify imaging, genomics, biological markers, and neuropsychological assessments that can be used to track the progression of AD [42]. More information can be accessed at http://adni.loni.usc.edu/. The ADNI data repository contains information from over 2,220 participants, including imaging, genetics, and clinical data. There are two types of imaging data: MRI and PET. The invasive nature of PET, combined with the low cost of MRI, led us to use MRI for this study. For genetics, we used SNP data, as previous genome-wide association studies have identified several genetic variants associated with an increased risk of AD [43,44].

The Clinical data in ADNI includes demographics, medication, biochemical data, and clinical tests (e.g., memory tests, and cognitive tests). Based on previous research [45,46], we chose 42 clinical features from clinical data for this study (as seen in Table 1 and Table 2). A detailed description of each clinical variable can be found in the supplementary material.

The definition of sMCI and pMCI is based on the DSM-V criteria,

Table 1
A summary of the continuous clinical variables of participants.

	sMCI(n = 297)	pMCI(n = 280)	Combined(n = 577)
Age	72.28 ± 7.43	73.92 ± 6.96	73.07 ± 7.25
Education years	16.04 ± 2.78	15.80 ± 2.79	15.92 ± 2.79
CDR	1.20 ± 0.65	1.92 ± 0.94	1.55 ± 0.88
ADAS11	8.48 ± 3.55	12.88 ± 4.35	10.62 ± 4.53
ADAS13	13.64 ± 5.45	20.84 ± 5.90	17.12 ± 6.72
MMSE	28.00 ± 1.70	26.93 ± 1.73	27.48 ± 1.79
RAVLT_I	37.72 ± 10.17	29.00 ± 7.51	33.48 ± 9.98
RAVLT_L	4.75 ± 2.40	2.99 ± 2.26	3.90 ± 2.49
RAVLT_F	4.35 ± 2.52	4.99 ± 2.22	4.66 ± 2.40
RAVLT_PF	50.36 ± 30.50	75.06 ± 28.76	62.34 ± 32.13
LDELTOTAL	7.03 ± 2.90	3.49 ± 2.91	5.31 ± 3.40
TRABSCORE	103.77 ± 49.81	122.23 ± 77.05	118.58 ± 67.69
FAQ	1.55 ± 2.77	5.40 ± 4.80	3.41 ± 4.34

*Data are mean ± standard deviation: CDR: Clinical Dementia Rating; ADAS: The Cognitive Subscale Alzheimer's Disease Assessment Scale; MMSE: Mini-Mental State Examination; RAVLT: The Rey Auditory Verbal Learning Test; LDELTOTAL: Logical Memory - Delayed Recall; TRABSCORE: Trails B score; FAQ: Functional Activities Questionnaire.

Table 2
A summary of the categorical clinical variables of participants.

		Participants, No (%)		
		sMCI	pMCI	Combined
Sex	Male	174(58.6)	172(61.4)	346(60.0)
	Female	123(41.4)	108(38.6)	231(40.0)
Marital Status	Married	215(72.3)	206(73.6)	421(73.0)
	Divorced	32(10.8)	17 (6.1)	49(8.5)
	Widowed	43(14.5)	30(10.7)	73(12.7)
	Never married	7(2.3)	4(1.4)	11(1.9)
Parents with Dementia	Yes	143(48.1)	172(61.4)	315(54.6)
	No	154(51.9)	104(37.1)	258(44.7)
Parents with AD	Yes	141(47.5)	147(52.5)	288(49.9)
	No	156(52.5)	133(47.5)	289(50.1)

where pMCI refers to MCI subjects who convert to dementia during the follow-up period, while sMCI is defined when the subjects do not fulfill these criteria [47]. While different studies use varying follow-up times, we have defined our follow-up period as 3 years. In other word, we defined sMCI as participants who do not convert to AD within 3 years, while pMCI as participants who do. Through the classification of sMCI and pMCI, we were able to predict the conversion of MCI to AD. This study only selected baseline data from participants to prevent data leakage. Our study included 577 MCI from 2200 data which all contain MRI. Our dataset included 297 sMCI and 280 pMCI and among them166 sMCI and 177 pMCI with MRI, SNP, and clinical data. Participants with three modalities and one or two modalities underwent multimodal fusion and feature extraction, respectively. Table 1 and Table 2 summarize the subjects in detail.

3.2. Data pre-processing

As mentioned above, our study consists of three modality data: MRI, SNP, and Clinical.

MRI data. The MRI data was preprocessed using FSL (https://fsl.fmrib.ox.ac.uk/). First, the N4 algorithm is used to correct bias fields [48]. In addition, affine linear alignment of onto the MIN152 atlas. Lastly, we stripped the skulls from each image and got images that had 129x145x129 voxels.

SNP data. SNP pre-processing consists of two steps to complete quality control and dimensionality reduction. Firstly, quality control was performed first on SNPs using Plink, retaining those with (i) genotype quality greater than 20; (ii) minor allele frequency (MAF) greater than 0.01; (iii) per-site missing rate less than 5 %, and (iv) Hardy–Weinberg equilibrium p-value greater than 0.05. There are still 7,610,179 SNPs in the VCF file after quality control, but only a few of them are associated with AD. Accordingly, we selected only SNPs that belong to the top AD gene candidates in AlzGene (https://www.alzgene.org/). Finally, we got 143,498 SNPs. Three groups were created from the final data: no alleles, just one allele, or two alleles.

Clinical data. This study covers 42 clinical features, including 38 continuous variables, 3 two-category variables (gender, whether the parent had dementia, whether the parent had AD), and 1 multi-category variable (marital status), as shown in Table 1, Table 2. Detailed information about these features can be found in the Supplementary File and ADNI (https://adni.loni.usc.edu). We used the average and the mode to interpolate missing values for the continuous and categorical variables, respectively. Each variables missing rate can be found in the Supplementary File. The continuous features were normalized using the Z-score method. One-hot encoding was performed for the categorical features.

3.3. Unimodal feature representation

In our study, we utilized various models to extract features from different modalities based on their characteristics. For MRI data, we opted for the 3D ResNet architecture, a Convolutional Neural Network (CNN) variant. CNNs have consistently demonstrated superior performance in processing medical imaging data due to their ability to capture spatial relationships effectively. When dealing with Single Nucleotide Polymorphism (SNP) data, characterized by high dimensionality and sparsity, we combined feature selection techniques with Stacked Denoising Autoencoders (SDAE) for feature extraction. This approach leverages the strengths of SDAE in handling high-dimensional, noisy data. Lastly, for structured clinical data, we employed a Deep Neural Network (DNN), a well-established method for representing clinical data structure.

3.3.1. MRI feature representation

In our previous study, we proposed a two-stage transfer learning method that combines transfer learning and contrastive learning for extracting medical imaging features [49]. Specifically, we used a 3D-ResNet pre-trained on a large public medical imaging dataset to identify common features, and then applied contrastive learning to extract more specific features from the target images. Our results indicated that our two-stage model improved performance compared to previous studies and outperformed benchmark models. In this study, we used the same method to extract MRI features.

3.3.2. SNP feature representation

In this study, the preprocessed SNPs contained 143,498 dimensions, but the sample size was still relatively small with over 300 subjects. Genomic data often contains a large number of irrelevant or redundant features, which can hinder the performance and efficiency of machine learning models [50]. Therefore, it is important to apply dimensionality reduction techniques to extract the most relevant and informative features from the genomic data.

Fig. 2 illustrates the genomic feature extraction method used in this study. The method consisted of two stages: feature selection and high-order feature extraction. In the feature selection stage, we proposed a hybrid multi-filter and wrapper method (HMFW) that combines the advantages of filter and wrapper methods while overcoming their limitations. The HMFW method first used multiple filter methods to quickly rank the features by their relevance to the target. To avoid the instability and loss of significant features that can occur with individual filter methods, we selected three representative filter methods (Information Gain, Fisher Score, and $\chi 2$ test) and jointly applied them to obtain the ranked feature set. Then, the wrapper method [51] is used to select the optimal subset of features that maximizes the target information, as shown in Fig. 2 (a). After feature selection, we obtained 450 genomic

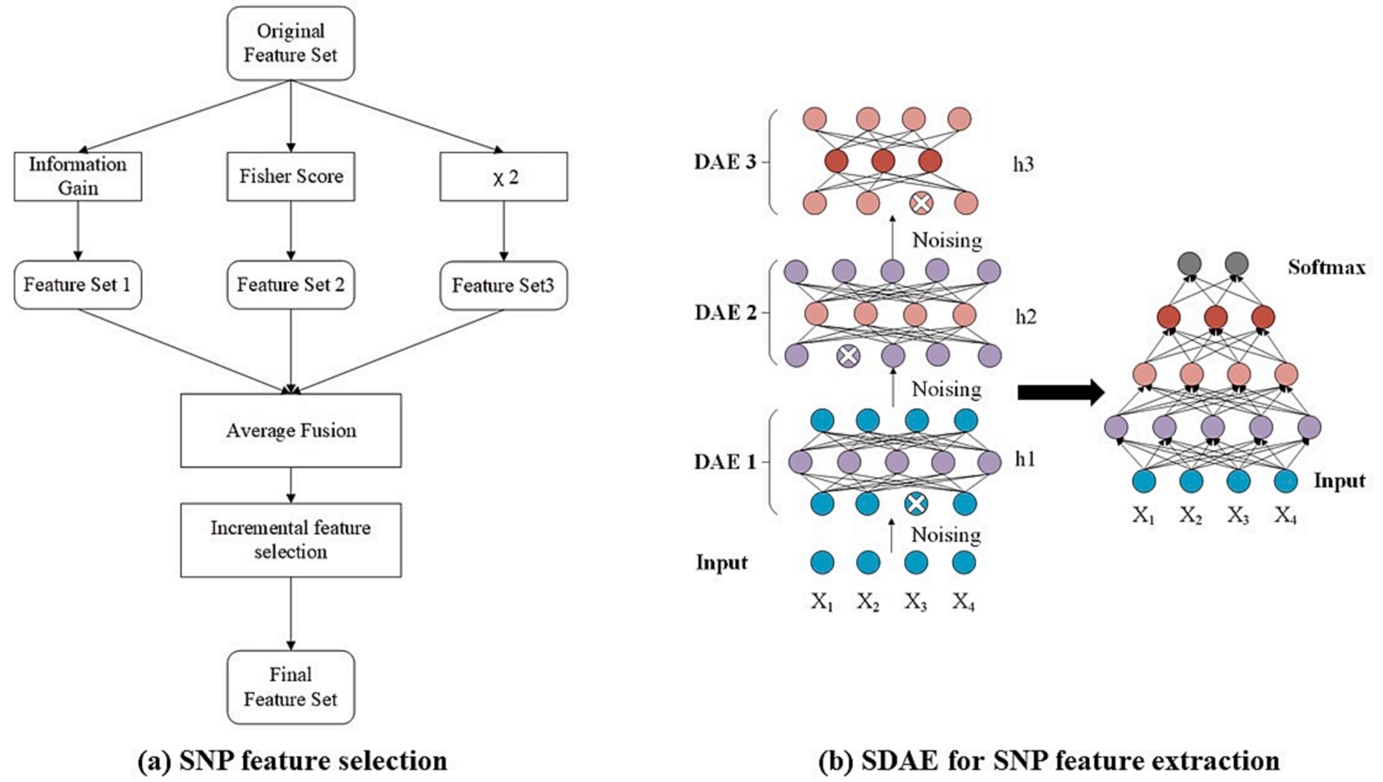


Fig. 2. SNP feature extraction architecture. Our study first used the (a) SNP feature selection method to reduce the features and then used (b) SDAE to capture the hidden relationships among features and extract high-order feature representations for fusion or prediction tasks. The (b) left is the unsupervised training process of the SDAE and the right is the supervised fine-tuning process of the SDAE.

features.

After feature selection, we used a stacked denoising autoencoder (SDAE) to extract high-order feature representations that capture the hidden relationships among the selected features. SDAE extracted deep features from the original input by combining multiple denoising autoencoders (DAE). Hidden layers of previous DAE serve as inputs to the next DAE. Fig. 2(b) shows the SDAE training process, which consists of unsupervised pre-training and supervised fine-tuning. Unsupervised training involves training each DAE individually. Zero-masking noise algorithm was used in this paper to add noise to the data, which resets some feature values to zero [52]. Through the encoder and decoder, the DAE reconstructed the original data. Following unsupervised training, all denoising autoencoders were connected and then combined with the SoftMax classifier where the parameters of the cascaded network were fine-tuned. Fig. 2 (b) illustrates the training process of the SDAE model.

3.3.3. Clinical feature representation

As discussed earlier, raw clinical data is typically sparse and cannot be directly fused with other modalities. To extract useful features from clinical data, we experimented with different network architectures and chose a deep neural network (DNN) which usually used as a clinical structure data representation method [53].

The DNN model had an input layer with 49 dimensions, two hidden layers with 64 and 32 dimensions, respectively, and an output layer with a single dimension. We used the ReLu activation function between the input and hidden layers and added BatchNorm1D to each layer. The output layer used the sigmoid activation function and the cross-entropy loss function. This architecture allows us to extract clinical features that can be fused with other modalities for improved AD prediction performance.

3.4. Hierarchical Attention-Based multimodal fusion

Using the above method, the medical imaging feature x^i, the genetic feature x^g, and the Clinical feature x^c are extracted respectively. Before fusion, the feature vector of MRI, SNP, and clinical data were projected into a shared embedding space with the same feature dimension. This is necessary because different modalities have different feature dimensions, which will affect the final fused feature representation. The existing projection methods include linear projection and nonlinear projection.

To enable the model to recalibrate each dimension according to its learned importance and to ensure that the dimensions are activated uniformly across the three modalities, we used nonlinear gating [54] which is defined as:

$$\widetilde{x^m} = \left(W_1^m x^m + b_1^m\right)^\circ \sigma\left(W_2^m\left(W_1^m x^m + b_1^m\right) + b_2^m\right) \tag{2-1}$$

$$\widetilde{x^g} = (W_1^g x^g + b_1^g)^\circ \sigma(W_2^g(W_1^g x^g + b_1^g) + b_2^g) \tag{2-2}$$

$$\widetilde{x^c} = \left(W_1^c x^c + b_1^c\right)^\circ \sigma\left(W_2^c\left(W_1^c + b_1^c\right) + b_2^c\right) \tag{2-3}$$

Where $\widetilde{x^m}$, $\widetilde{x^g}$ and $\widetilde{x^c}$ are the 64-dimensional vectors of imaging, genetic, and clinical data after nonlinear gating mapping, respectively. *W* and *b* are both learnable parameters for the weight vectors and bias, respectively. In this case, ° is a multiplication of the corresponding positions, and σ is the sigmoid activation function.

The nonlinear gating is inspired by the Gated Linear Unit (GLU)[55], which is formulated as follows:

$$\widetilde{x} = (W_1 x + b_1)^\circ \sigma(W_2 x + b_2) \tag{2-4}$$

Compared to GLU, nonlinear gating was chosen for this study mainly for the following reasons. Firstly, nonlinear gating allows the model to capture nonlinear interactions between features and intra-feature dependencies, which can improve the performance of the fusion model. Secondly, nonlinear gating allows the model to recalibrate the activation strength of each feature dimension based on its importance, which can improve the interpretability and robustness of the model. These benefits make nonlinear gating a suitable choice for multimodal fusion in the context of AD prediction.

To improve the performance and interpretability of the multimodal fusion model, we propose a hierarchical attention mechanism, as illustrated in Fig. 1. The model first fused each pair of unimodal feature vectors (MRI & SNP, MRI & Clinical, and SNP & Clinical) using an attention mechanism, which allows the model to focus on relevant information and suppress redundant information. The pairs of fused feature vectors were then subjected to a higher-level attention mechanism-based fusion to form the final fused representation of all three modalities (MRI & SNP & Clinical). This hierarchical approach enables the model to capture the complex relationships between the different modalities and their relevance to the target.

The specific attention mechanism in this study was based on the Keyless Attention mechanism [56]. Keyless Attention is defined as follows:

$$e_i = w^T x_i \tag{2-5}$$

$$\lambda_i = \frac{\exp(e_i)}{\sum_{j=1}^{n}\exp(e_j)} \tag{2-6}$$

$$Attention\ \left(x^1 \cdots x^M\right) = \sum_{i=1}^{M} \lambda_i x_i \tag{2-7}$$

Our hierarchical attention is defined as follows:

$$x^{ms} = Attention\left(\widetilde{x^m}, \widetilde{x^g}\right) \tag{2-8}$$

$$x^{mc} = Attention\left(\widetilde{x^m}, \widetilde{x^c}\right) \tag{2-9}$$

$$x^{cg} = Attention\left(\widetilde{x^c}, \widetilde{x^g}\right) \tag{2-10}$$

$$x^{fusion} = Attention\left(x^{mg}, x^{mc}, x^{cg}\right) \tag{2-11}$$

The fusion of three modalities is shown in Fig. 1. In the case of only two-modal fusion, only the general attention calculation (2–12) is used.

$$x^{ab} = Attention\left(\widetilde{x^a}, \widetilde{x^b}\right),\ where\, a, b \in (m, g, c) \tag{2-12}$$

3.5. Classification

After fusion, we applied a two-layer multilayer perceptron (MLP) to the fused feature representation. For the three-modal fusion, we use (2–11) to get the fused feature representation. For the case of only two-modal fusion, only (2–12) used to get the fused feature representation. The MLP has an input layer with the same number of nodes as the fused representation, a hidden layer with half the number of nodes, and an output layer with two nodes. The model is trained using the cross-entropy loss function to minimize the error between the predicted and actual disease labels. Once the model was trained, it can be used to predict the disease status of new subjects based on their multimodal data.

3.6. Experimental settings and evaluation

3.6.1. Experimental settings

The models have been implemented in Pytorch, a framework developed by Facebook with a GeForce GTX 3080Ti. To save computational resources, we used the pre-trained parameters of the MRI model from our previous study without updating them. The data used in this multimodal fusion study consisted of 343 samples (166 sMCI and 177 pMCI) with MRIs, SNPs, and clinical data, divided into training and test sets with a 3:1 ratio as show in Fig. S1. We applied five-fold cross-validation on the training set to select the optimal hyperparameters and the best models for each modality and the multimodal fusion. Finally, the model is applied to the test set for model evaluation. As for the unimodal model (MRI and clinical), we used all 577 samples to fully train the model which were divided into training and test sets with a 4:1 ratio. The data spilt are shown in Fig. S1. Table S2 shows the best hyperparameters for unimodal and multimodal models.

3.6.2. Evaluation

We evaluated our models using Accuracy (Acc), Sensitivity (Sens), Specificity (Spec), F1-Score (F1), and Area Under ROC Curve (AUC).

Lastly, we constructed confidence intervals (CIs) for the evaluation metrics of the models using non-parametric bootstrapping. As a result, each estimate was given a distribution and a 95 % CI estimated using the 2.5 and 97.5 percentiles of the bootstrap samples [57].

4. Results

4.1. Performance of baseline model

We first evaluated the performance of a single-modality model as the baseline for further comparisons. Fig. 3 summarizes the results and shows that Clinical has the best unimodal performance, while SNP has the worst. For the MRI modality, we used the results of our previous studies, and the best model achieved an accuracy of 81.9 %. For the clinical modality, as the method, we created 2 fully connected layers, and the best model achieved an accuracy of 83.7 %. For the SNP modality, we extracted the SNP features as described earlier and added a SoftMax classifier for prediction. The best SNP unimodal model achieved an accuracy of 67.4 %.

4.2. Performance of multimodal models

The following experiments were performed for predicting the conversion of MCI to AD using different combinations of modalities. These experiments were conducted to verify the effectiveness and necessity of multimodal fusion and to determine the most important modalities or combinations of modalities for disease prediction. This information can be useful for healthcare professionals in decision-making and for guiding future research.

1) Unimodality. Predict MCI to AD conversion based on a single modality as in section 4.1.
2) Two modalities. Predict MCI to AD conversion is based on any combination of the two modalities. The fused feature representation is obtained by fusing the two modalities using equations (2)–(12), which are then combined with a classification layer to predict Alzheimer's disease.
3) Three modalities. Predicting MCI to AD conversion using three modalities based on the HAMF.

The above models were evaluated for their ability to predict MCI to AD conversion. Table 3 lists all 4 evaluation metrics of models with different modality combinations, and the optimal outcome is bolded. As shown in Table 3A, clinical data achieved 83.7 % accuracy as the most accurate unimodal model. Compared with unimodality, the fusion of two modalities resulted in different degrees of improvement of accuracy in MCI to AD conversion prediction, and the optimal two-modality combination was MRI & Clinical with an accuracy of 87.2 %. Compared with the optimal MRI & Clinical, the accuracy of the three modality fusions of MRI&SNP &Clinical was not improved, which was also 87.2 %, but the accuracy was improved compared with the other two modalities fusions.

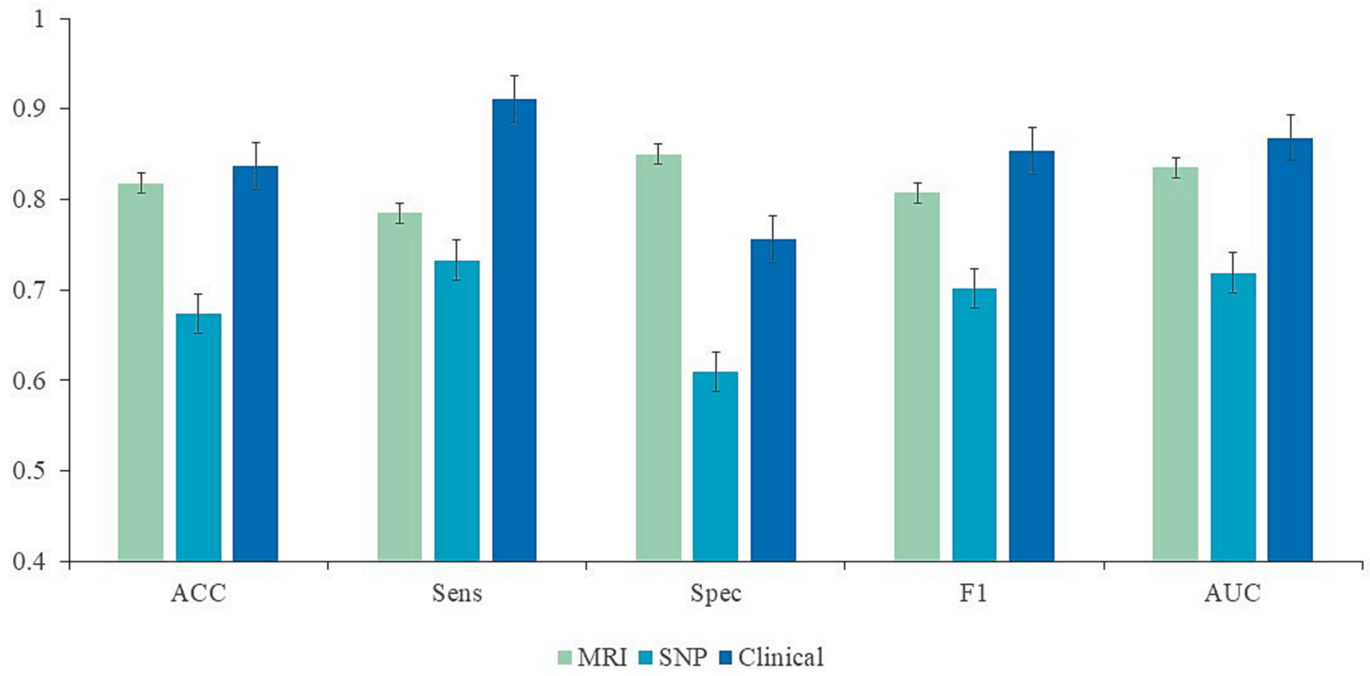


Fig. 3. Evaluation results of unimodal models. This graph shows all 5 evaluation metrics for imaging, genetics, and clinical models. Clinical models give the best overall performance, while genetic models give the lowest overall performance.

Table 3
Performance of different multimodal models for prediction.

		Acc	Sens	Spec	F1
Unimodal	MRI	0.818 (0.798, 0.841)	0.785 (0.754, 0.821)	0.850 (0.815,0.877)	0.807 (0.783, 0.834)
	SNP	0.663 (0.641, 0.686)	0.733 (0.697, 0.765)	0.586 (0.541, 0.630)	0.694 (0.674,0.717)
	Clinical	0.837 (0.819,0.863)	**0.911** **(0.890,0.931)**	0.756 (0.725,0.799)	0.854 (0.837,0.875)
Two Modality	MRI + SNP	0.826 (0.801,0.850)	0.844 (0.814,0.875)	0.805 (0.768,0.833)	0.835 (0.812,0.859)
	Clinical + SNP	0.837 (0.818,0.860)	0.866 (0.830,0.900)	0.805 (0.764,0.836)	0.848 (0.826,0.867)
	MRI + Clinical	**0.872** **(0.850,0.891)**	0.889 (0.860,0.911)	**0.854** **(0.827,0.882)**	0.879 (0.860,0.897)
Three Modality	MRI + SNP+ Clinical	**0.872** **(0.851,0.890)**	0.888 (0.857,0.909)	**0.854** **(0.824, 0.882)**	**0.884** **(0.859,0.896)**

The bold numbers denote the maximum value of each column. A 95% confidence interval is described in the parenthesis; ACC, Accuracy; F1, F1-score; Sens, Sensitivity; Spec, Specificity.

Regarding the AUC metric, the distribution of optimal modality combinations demonstrated similarities with accuracy, albeit with slight distinctions. In terms of accuracy, the fusion of the two modalities Clinical & SNP achieved the same accuracy of 83.7 % as the unimodal model (clinical data). However, in terms of AUC, Clinical & SNP improved by 0.1 % over Clinical. Similarly, two-modality fusion MRI & Clinical and three-modality fusion MRI& Clinical & SNP achieved the same accuracy of 87.2 %, but the AUC of MRI & Clinical &SNP improved by 0.6 % over MRI & Clinical to obtain the highest AUC value among all results, and the ROC curves are shown in Fig. 4(a).

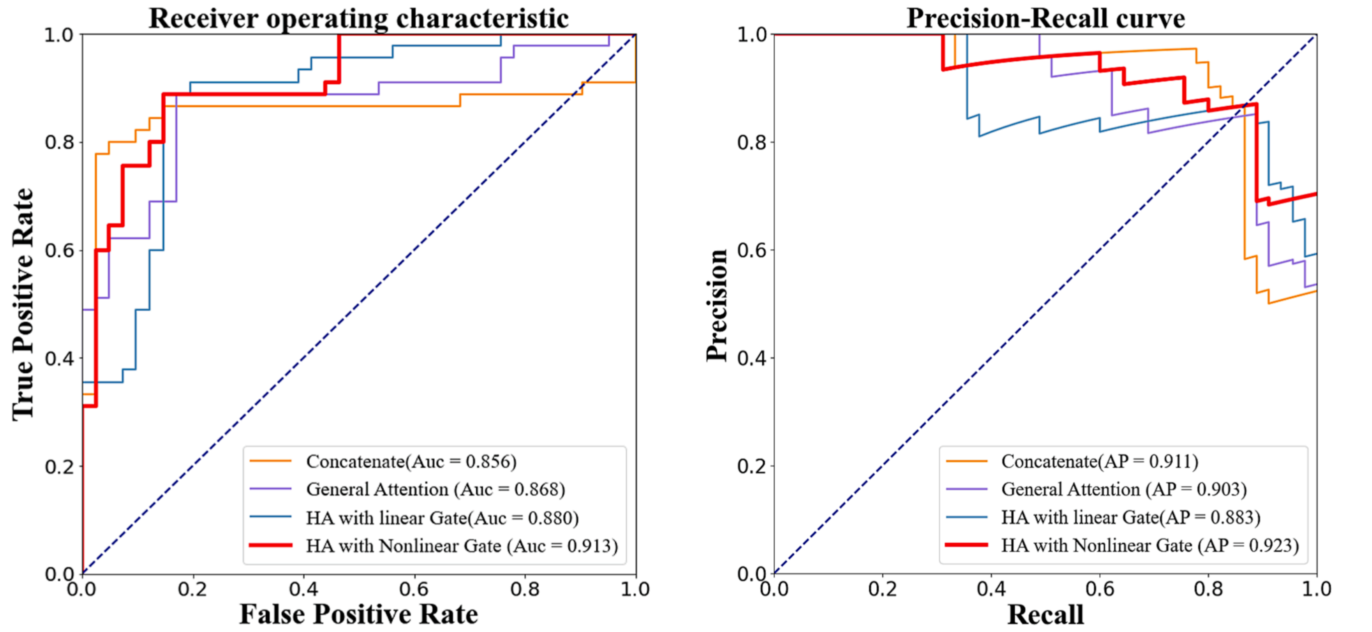
4.3. Ablation study for the HAMF module

The following comparative experiments were conducted to demonstrate the effectiveness of the HAMF proposed in this paper:

1) Concatenate. Concatenate features of different modalities directly to form the final fused feature.
2) General Attention. A general attention mechanism that directly performs the summation of the three attentions without hierarchy.
3) Hierarchical Attention with Linear Gating. The others are the same as our method, except that nonlinear gating is replaced by linear gating.
4) Hierarchical Attention with nonlinear gating (Our).

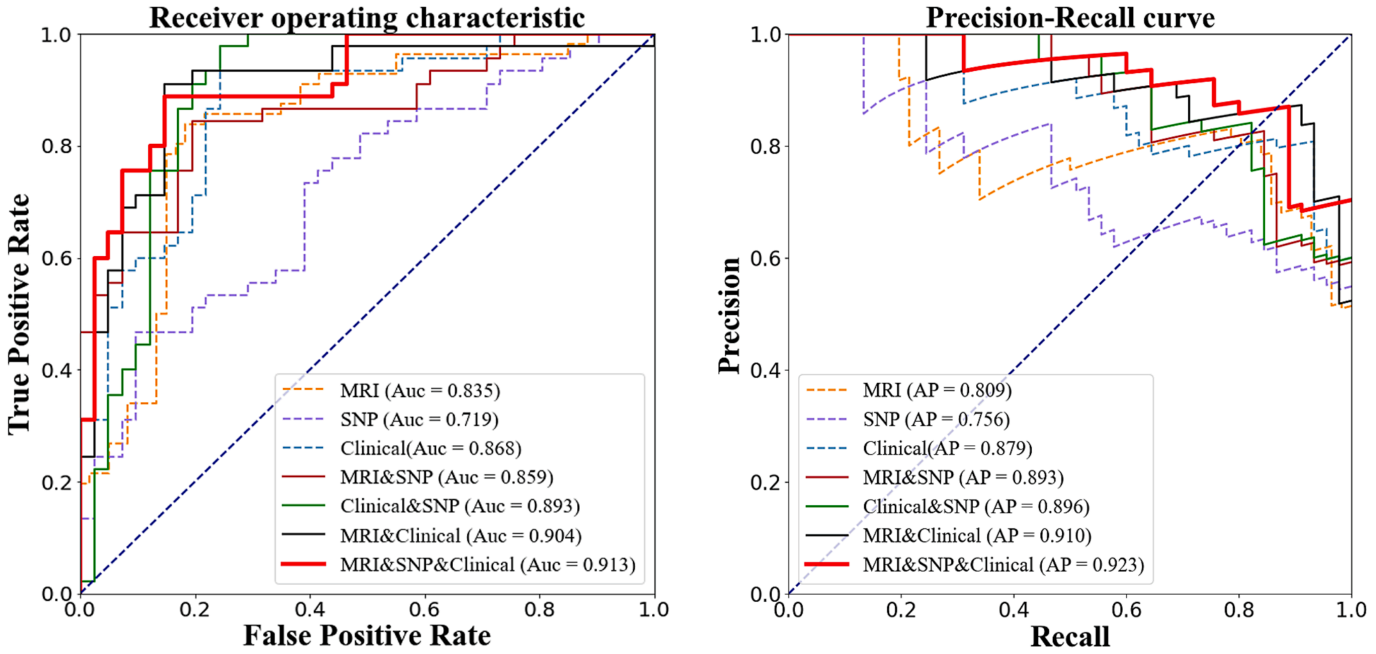
The effectiveness of the fusion strategy is evaluated based on the prediction results with different multimodal fusion strategies. The evaluation results of different multimodal fusion strategies are shown in Fig. 6. The multimodal fusion strategy based on hierarchical attention with nonlinear gating achieve the highest accuracy (ACC), sensitivity (Sens), specificity (Spec), F1 score, and AUC. Compared with the direct concatenate feature (Concatenate) model, the fusion strategies using the attention (general attention, hierarchical attention) all achieve better prediction results for conversion. Compared with the linear gating, the hierarchical attention with nonlinear gating achieves better performance as shown in Fig. 6.

As shown in Fig. 5 (a), the ROC curves of the four models indicate



(a) ROC curve of different multimodal fusion strategies **(b) PR curve of different multimodal fusion strategies**

Fig. 4. Curves of Different multimodal importance.



(a) ROC curve of different multimodal importance **(b) PR curve of different multimodal importance**

Fig. 5. Curves of Different multimodal Fusion strategies. HA, Hierarchical Attention.

that multimodal fusion based on hierarchical attention produces the best AUC. Direct concatenate obtain the poorest result with an AUC of 0.856, which is lower than the arbitrary model using the attention mechanism and lower than the clinical alone or any two-modality fusion model (as shown in Fig. 5). Hierarchical attention obtain a better result with AUC of 0.880(linear gating) and 0.913(nonlinear gating) than general attention with AUC of 0.868. Hierarchical attention with nonlinear gating (our model) get the best performance with AUC of 0.913.

Bold numbers indicate the column's maximum value; AUC: Area under the ROC Curve; ACC: Accuracy; CNN: convolutional neural network; SDAE: Stacked Denoising Auto-Encoder; Time, published time. ROI, Region of interest, using Freesufer or other tools to extract the brain feature. CSF, Cerebrospinal Fluid; Clinical, containing demographic information, neurological scale information, etc.; MRI, Magnetic Resonance Imaging; PET, Positron Emission Computed Tomography; SNP, Single Nucleotide Polymorphism; NA, not available.

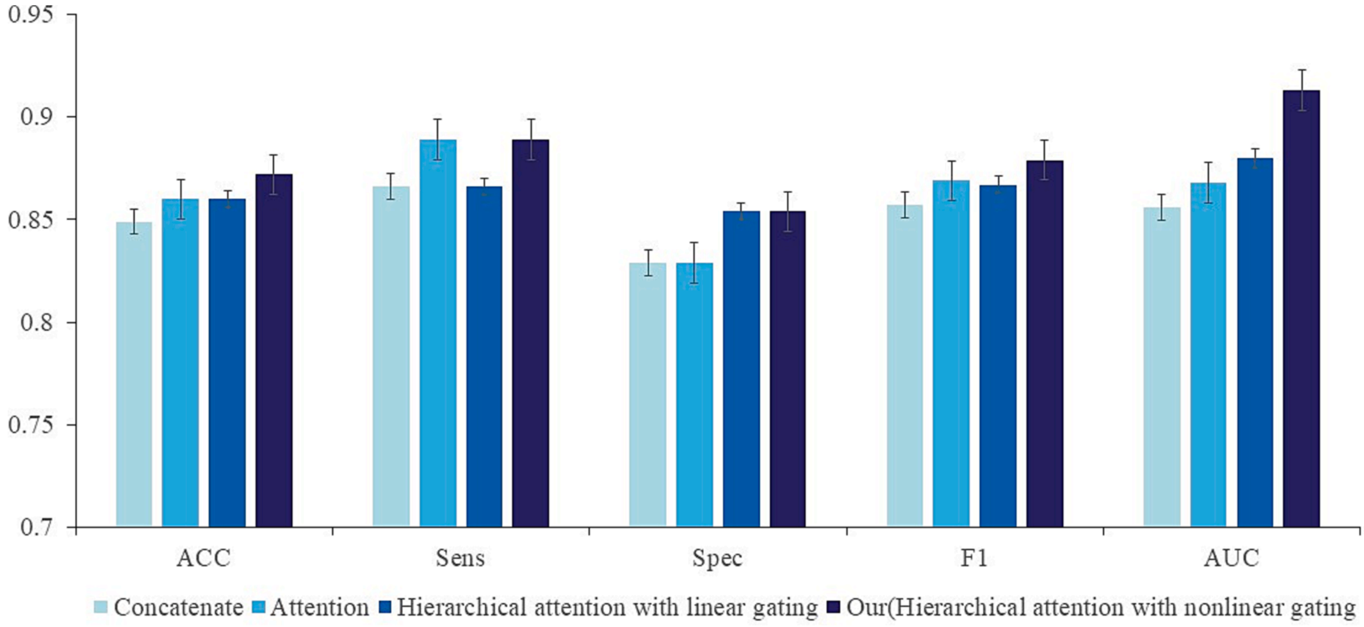


Fig. 6. Performance of different fusion strategies.

5. Model explainability

An advantage of using attention mechanisms is the interpretability of their results. For this reason, we visualize our Hierarchical Attention used in the HAMF model.

Fig. 7 illustrates the fusion of unimodal features using hierarchical attention. As shown in the figure, the weights (color-coded) of each modality are depicted. As expected, the network focuses primarily on clinical and MRI channels in the first layer, while the clinical-MRI pair in the first layer provides the highest score in the second layer.

Unimodal and multimodal model results indicated that Clinical features are important to the model's performance. We examined all clinical feature carefully to make sure none could give an unfair advantage to the model. In this study, we leveraged SHapley Additive exPlanations (SHAP) [62] to evaluate the importance of each input feature in our models. SHAP is a method that explains the output of any machine learning model by assigning feature importance values based on the concept of Shapley values from cooperative game theory. These values reflect the contribution of each feature to the model's output and enabled us to identify the most important features for predicting Alzheimer's disease. Additionally, SHAP facilitated our understanding of the underlying biological mechanisms of the disease by generating feature-level visualizations, which allowed us to investigate how each input feature influenced the model's predictions.

Based on SHAP explainers [62], we calculated each clinical feature contributions of clinical model. Fig. 8 shows the top ten clinical features in the model. The most influential feature is FDR followed by RAVIL-I (RAVIL- immediate). It is important to note that features with a longer tail to the right have a greater positive influence, and the reverse is also true. All top ten features are symmetric between the two classes. For example, low values of FAQ negatively affect pMCI predictions, but they positively influence sMCI predictions. Please refer to the supplement for important features of SNPs.

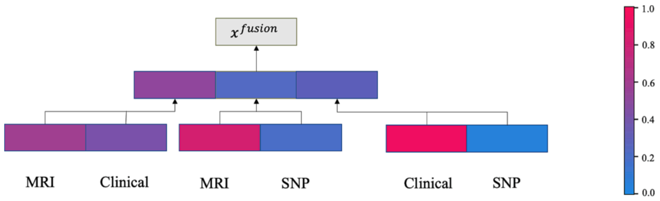
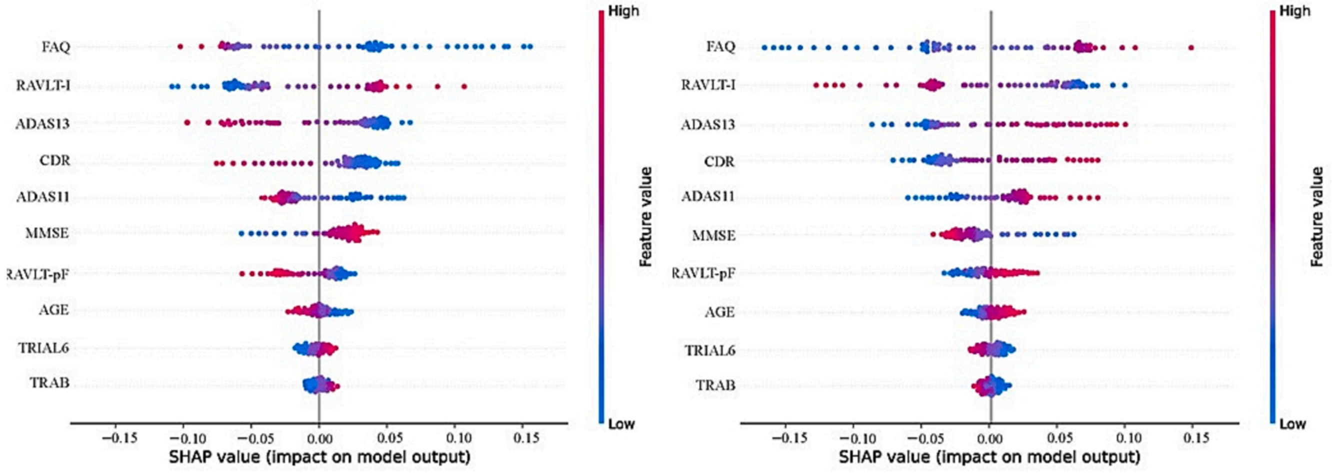


Fig. 7. A visual representation of HAMF's hierarchical attention fusion strategy. Both the pair-wise attention (first layer) and the higher layer scores of the clinical-MRI model are the highest.

6. Discussion

In this study, we propose a Hierarchical Attention-based Multimodal Fusion framework (HAMF) that uses three modalities, MRI, SNP, and Clinical, as the input of prediction tasks of conversion from MCI to AD. The model achieve excellent accuracy, sensitivity, specificity, and F1 score of 87.2 %, 93.3 %, 80.4 %, and 88.4 %, respectively, and an AUC of 91.1 %.

The simplest and most used approach for multimodal fusion is to directly concatenate or sum features from different modalities into the classifier [16–22]. For example, An et al combined MRI, PET, and cerebrospinal fluid markers directly into their AD classification model [14]. Venugopalan et al. used CNN and MLP to complete the dimensionality reduction of MRI, SNP, and Clinical, respectively. They fed the three reduced features into the classifier in series to complete the classification of AD [15]. However, different modalities contain various amounts of information for completing the task. In Fig. 4, the AUC of MCI to AD conversion prediction using direct concatenation fusion strategies of the three modalities is lower than that of the Clinical modal alone, indicating that direct concatenation or summation is not effective at highlighting critical information or suppressing redundant information, which reduces model performance in varying degrees. Additionally, different modalities have rich connections, so direct concatenation or summations cannot effectively explore the relations between different modalities, which affects the fusion effect. Moreover, as illustrated in Fig. 6, the attention-based multimodal fusion model outperforms the model based on general connection multimodal fusion. Using the attention mechanism, weights can be assigned to various modalities by back-propagation dynamic weighting, and the more important modalities are given larger weights. The model performance is improved by improving the expression of important information and suppressing redundant information to obtain a more accurate representation of fusion features [63]. As well as exploring the relationship between different modalities, hierarchical attention with nonlinear gating makes



(a) SHAP summary plot for class sMCI

(b)SHAP summary plot for class pMCI

Fig. 8. SHAP summary plot for the clinical unimodal model. The left figure and right figure show the sMCI and pMCI class respectively. Dots represent instances and their colors indicate feature values (red = high, blue = low).

optimal use of nonlinear connections among different modalities combinations. In contrast to general attention and linear gating hierarchical attention, our method combines multiple modalities more effectively, resulting in better outcomes. These results demonstrate the effectiveness of the hierarchical multimodal fusion method proposed in this study.

This study investigates the performance of different unimodal and multimodal MCI to AD conversion predictions and provides a comprehensive evaluation of the importance of different modalities (including different unimodal or multimodal). By comparing the evaluation of MCI to AD conversion prediction with different unimodal or multimodal, we have identified the following key findings. First, the optimal unimodal model was Clinical based with an accuracy of 83.7 %, which is consistent with several existing studies [18,64]. This study suggests that using non-invasive Clinical data for AD prediction may be a good option when budgets are limited, or initial clinical recruitment screenings have been performed. We also used SHAP to rank the importance of features in Clinical unimodal model to increase the interpretability of the model. As Fig. 8, the most influential feature is FDR followed by RAVIL-I and ADAS13 which is consistent with previous studies [45].

Additionally, multimodal fusion improves AD prediction model performance. Two- or three-modality fusion models achieved higher AUC values than single-modality models, with MRI&SNP& Clinical achieving the best AUC of 91.1 %. MRI &Clinical was the best two-modality combination, achieving an AUC of 90.4 %. This is consistent with existing studies [22,65–68] that different modalities describe AD from different perspectives, capturing the heterogeneity of the disease and improving MCI to AD conversion prediction. MRI highlights structural changes in the brain from the macroscopic aspect, while SNP explains AD heritability from a microscopic biology perspective, and Clinical describes functional changes in disease process. This suggests the necessity of multimodal fusion for the prediction of AD. Notably, MRI & Clinical & SNP achieved the optimal AUC but the same accuracy as MRI & Clinical, both at 87.2 %, indicating that the addition of SNP did not improve the model’s prediction accuracy. The accuracy of SNP for model evaluation was 66.6 %, the lowest among all unimodal models. The reason for this result may be that both Imaging and Clinical data are phenotypic features that are closely related to diagnostic labels, but SNPs are genetic features that indicate genetic variation predisposition of disease but not necessarily directly connected to a current disease condition represented by the diagnostic labels [14,69].

Lastly, this study summarizes the research on predicting MCI to AD conversion using multimodal fusion over the past three years as shown in Table 4. We selected studies that met the following criteria comparison: (1) Classification model of AD conversion prediction. (2)

Table 4
Comparison of state-of-the-art research on AD conversion prediction using multimodal fusion.

	Type of modality	Feature extraction Method	Fusion Method	AUC	ACC
Our	MRI + SNP + Clinical	CNN + SDAE + DNN	Hierarchical Attention	**0.91**	**0.87**
Wang et al., 2022 [57]	MRI + Clinical	CNN	Concatenate	0.91	0.85
Pena et al., 2022 [15]	MRI + Clinical	CNN Transfer learning	Concatenate	0.85	NA
Mirabnahrazam et al., 2022 [16]	MRI + SNP	ROI based	Concatenate	NA	0.74
Ma, Zhang, & Wang, 2022 [58]	MRI + fMRI	Riemannian Manifold	Self-Attention	NA	0.85
Ning, Xiao, Feng, Chen, & Zhang, 2021 [59]	MRI + PET	ROI based	Shared Representation	0.84	0.85
El-Sappagh, Alonso, Islam, Sultan, & Kwak, 2021 [17]	MRI + PET + APOE4 + Clinical	ROI based	Concatenate	0.88	0.87
Shen et al., 2021 [18]	MRI + PET + Clinical	ROI based	Concatenate	0.79	0.78
Yang et al., 2021 [19]	MRI + Clinical + APOE4	CNN	Concatenate	0.90	0.83
Shao, Peng, Zu, Wang, & Zhang, 2020 [60]	MRI + PET	ROI based	Hypergraph	0.70	0.75
Khatri & Kwon, 2020 [20]	MRI + APOE4 + CSF + Clinical	ROI based	Concatenate	0.83	0.83
Forouzannezhad et al., 2020 [21]	MRI + PET + Clinical	ROI based	Concatenate	NA	0.73

Containing at least two modalities. (3) Published in the last 4 years (preprint was not included). (4) The data were from ADNI. The models in this study achieved state-of-the-art results, outperforming existing studies. This is partially due to the hierarchical attention we have adopted, as well as the deep learning-based feature extraction approach. As shown in Table 4, several previous studies have used feature engineering or selected regions of interest (ROI) as the extracted features [17–19,21,22,60,61], ignoring other features in the model and sometimes leading to information loss. Instead, we extract all features from each modality using deep learning, which improves the model's performance. Reviewing the existing studies on multimodal fusion-based for predicting MCI to AD conversion reveals that MRI is the most used, while SNP is less commonly used. This may be related to the fact that SNP has limited value for the improvement of a AD prediction model. This limited usefulness of SNP may be attributed to the diverse nature of the disease, which comprises multiple distinct types. In comparison to studies based on time-series data, our study requires only a static dataset, which has the advantages of simplicity and low cost. Moreover, our model also produces better results than time series data [64,70]. Multimodal fusion methods based on hierarchical attention mechanisms are shown to be effective again by comparing them with other studies.

Although it has demonstrated promising results, the proposed method still has some limitations. First, our method combined three modalities (MRI, SNP, and Clinical), but the ADNI database includes other modalities, such as PET and CSF, and the clinical data contains many more features than the 44 we used. Second, we only tested our method on the two-class problem, but accurate diagnosis of patients at a particular stage of the disease is important. Lastly, the confidence intervals of our results show overlapping between the modalities. Future studies could use additional tests such as reclassification indices (e.g., absolute reclassification index, integrated discrimination index) or comparison of ROC curves with statistical tests to further validate our findings.

7. Conclusion

In this work, we propose a Hierarchical Attention-based Multimodal Fusion framework (HAMF) to predict MCI to AD conversion. Many existing multimodal models simply concatenate the features from each modality, regardless of their importance or cross-modality connections. To address this problem, we employ hierarchical attention-based multimodal fusion in which attention reinforces the most important features extracted from each modality, and hierarchical attention reinforces the relationship between the modalities. This results in an accuracy of 87.2 %, which is superior to existing multimodal fusion methods and defines the state-of-the-art for predicting MCI to AD conversion. We also investigate the importance of each modality and modality combination to inform decision-making and inspire future data collection efforts. We anticipate that our work will benefit clinical practice and provide insight into the powerful hierarchical attention-based models. In the future, we plan to explore our method with more modalities and investigate our method in multi-class classification, including AD/CN/sMCI/pMCI.

8. Contribution statement

Peixin Lu is responsible for designing the project, formal analysis, investigation, methodology and the original draft. Lianting Hu and Huiying Liang is responsible for the final manuscript. Mitelpunkt Alexis and Bhatnagar Surbhi is responsible for review and editing. Long Lu is responsible for conceptualization of the project and funding acquisition. All authors contributed to critically reviewing and editing the manuscript.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data are open access

Acknowledgment

This research was funded by Medical Science and Technology Foundation of Guangdong Province (A202311), The Excellent Young Scientists Fund (82122036), The Basic and Applied Basic Research Foundation of Guangdong Province (2022A1515110722), National Natural Science Foundation of China (61936013, 71921002), The National Social Science Fund of China (18ZDA325), National Key R&D Program of China (2019YFC0120003), Natural Science Foundation of Hubei Province of China (2019CFA025). The numerical calculations in this paper have been done on the supercomputing system in the Supercomputing Center of Wuhan University.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.bspc.2023.105669.

References

[1] A. Association, 2019 Alzheimer's disease facts and figures, Alzheimer's & Dementia 15 (2019) 321–387.
[2] R. Heun, M. Mazanek, K.-R. Atzor, J. Tintera, J. Gawehn, M. Burkart, M. Gänsicke, P. Falkaic, P. Stoeter, Amygdala-Hippocampal Atrophy and Memory Performance in Dementia of Alzheimer Type, Dement. Geriatr. Cogn. Disord. 8 (1997) 329–336.
[3] G. Livingston, A. Sommerlad, V. Orgeta, S.G. Costafreda, J. Huntley, D. Ames, C. G. Ballard, S. Banerjee, A. Burns, J. Cohen-Mansfield, C. Cooper, N.N. Fox, L. N. Gitlin, R. Howard, H.C. Kales, E.B. Larson, K. Ritchie, K. Rockwood, E. L. Sampson, Q.M. Samus, L.S. Schneider, G. Selbæk, L. Teri, N. Mukadam, Dementia prevention, intervention, and care, Lancet 390 (2017) 2673–2734.
[4] C.R. Jack, D.A. Bennett, K. Blennow, M.C. Carrillo, B. Dunn, S.L.B. Haeberlein, D. M. Holtzman, W.J. Jagust, F. Jessen, J. Karlawish, E. Liu, J.L. Molinuevo, T. J. Montine, C.H. Phelps, K.P. Rankin, C.C. Rowe, P. Scheltens, E.R. Siemers, H. M. Snyder, R.A. Sperling, NIA-AA Research Framework: Toward a biological definition of Alzheimer's disease, Alzheimers Dement. 14 (2018) 535–562.
[5] R.C. Petersen, Mild cognitive impairment: Transition between aging and Alzheimer's disease, Neurologia 15 (2000) 93–101.
[6] E.D. Roberson, Mucke L (2006) 100 Years and Counting: Prospects for Defeating Alzheimer's Disease, Science 314 (1979) 781–784.
[7] 2020 Alzheimer's disease facts and figures , Alzheimer's Dementia 16.
[8] S. Rathore, M. Habes, M.A. Iftikhar, A. Shacklett, C. Davatzikos, A review on neuroimaging-based classification studies and associated feature extraction methods for Alzheimer's disease and its prodromal stages, Neuroimage 155 (2017) 530–548.
[9] C. Salvatore, A. Cerasa, P. Battista, M.C. Gilardi, A. Quattrone, I. Castiglioni, A. D. Neuroimaging, Magnetic resonance imaging biomarkers for the early diagnosis of Alzheimer's disease: a machine learning approach, Front. Neurosci. 9 (2015).
[10] J. Zhang, B. Zheng, A. Gao, X. Feng, D.-P. Liang, X. Long, A 3D densely connected convolution neural network with connection-wise attention mechanism for Alzheimer's disease classification, Magn. Reson. Imaging (2021).
[11] M.A. Wajid, A. Zafar, Multimodal Fusion: A Review, Taxonomy, Open Challenges, Research Roadmap and Future Directions, 2021.
[12] B. Bouchey, J. Castek, J. Thygeson, Multimodal Learning. Innovative Learning Environments in STEM Higher Education, 2021.
[13] L. An, E. Adeli, M. Liu, J.B. Zhang, S.-W. Lee, D. Shen, A Hierarchical Feature and Sample Selection Framework and Its Application for Alzheimer's Disease Diagnosis, Sci. Rep. 7 (2017).
[14] J. Venugopalan, L. Tong, H. Hassanzadeh, M.D. Wang, Multimodal deep learning models for early detection of Alzheimers disease stage, Sci. Rep. 11 (2021).
[15] D. Pena, J. Suescun, M. Schiess, T.M. Ellmore, L. Giancardo, the ADNI, Toward a Multimodal Computer-Aided Diagnostic Tool for Alzheimer's Disease Conversion. Front. Neurosci., 15 (2022).
[16] G. Mirabnahrazam, D. Ma, S. Lee, K. Popuri, H. Lee, J. Cao, L. Wang, J.E. Galvin, M.F. Beg, Initiative the AND, Machine Learning Based Multimodal Neuroimaging Genomics Dementia Score for Predicting Future Conversion to Alzheimer's Disease, J. Alzheimer's Dis., 87 (2022) 1345–1365.

[17] S. El-Sappagh, J.M. Alonso, S.M.R. Islam, A.M. Sultan, K.S. Kwak, A multilayer multimodal detection and prediction model based on explainable artificial intelligence for Alzheimer's disease, Sci. Rep. 11 (2021) 2660.
[18] H.T. Shen, X. Zhu, Z. Zhang, S. Wang, Y. Chen, X. Xu, J. Shao, Heterogeneous data fusion for predicting mild cognitive impairment conversion, Inf. Fusion 66 (2021) 54–63.
[19] L. Yang, X. Wang, Q. Guo, S. Gladstein, D.W. Wooten, T. Li, W.Z. Robieson, Y. Sun, X. Huang, Deep Learning Based Multimodal Progression Modeling for Alzheimers Disease, Stat. Biopharm. Res. 13 (2021) 337–343.
[20] U. Khatri, Kwon G-R (2020) An Efficient Combination among sMRI, CSF, Cognitive Score, and APOE4 Biomarkers for Classification of AD and MCI Using Extreme Learning Machine, Comput. Intell. Neurosci. (2020).
[21] P. Forouzannezhad, A. Abbaspour, C. Li, C. Fang, U. Williams, M. Cabrerizo, A. Barreto, J. Andrian, N. Rishe, R.E. Curiel, D. Loewenstein, R. Duara, M. Adjouadi, A Gaussian-based model for early detection of mild cognitive impairment using multimodal neuroimaging, J. Neurosci. Methods 333 (2020), 108544.
[22] S.E. Spasov, L. Passamonti, A. Duggento, P. Liu, N. Toschi, A parameter-efficient deep learning approach to predict conversion from mild cognitive impairment to Alzheimer's disease, Neuroimage 189 (2019) 276–287.
[23] J. Wen, E. Thibeau-Sutre, M. Diaz-Melo, J. Samper-González, A. Routier, S. Bottani, D. Dormont, S. Durrleman, N. Burgos, O. Colliot, Convolutional neural networks for classification of Alzheimer's disease: Overview and reproducible evaluation, Med. Image Anal. 63 (2020), 101694.
[24] K.D. Tzimourta, V. Christou, A.T. Tzallas, N. Giannakeas, L.G. Astrakas, P. A. Angelidis, D.G. Tsalikakis, M.G. Tsipouras, Machine Learning Algorithms and Statistical Approaches for Alzheimer's Disease Analysis Based on Resting-State EEG Recordings: A Systematic Review, Int. J. Neural Syst. 2130002 (2021).
[25] S. Kumar, I. Oh, S.E. Schindler, A.M. Lai, P.R.O. Payne, A. Gupta, Machine learning for modeling the progression of Alzheimer disease dementia using clinical data: a systematic literature review, JAMIA Open 4 (2021).
[26] S. Palmqvist, P. Tideman, N.C. Cullen, H. Zetterberg, K. Blennow, J.L. Dage, E. Stomrud, S. Janelidze, N. Mattsson-Carlgren, O. Hansson, Prediction of future Alzheimer's disease dementia using plasma phospho-tau combined with other accessible measures, Nat. Med. (2021).
[27] H. Zhang, Y. Wang, D. Lyu, Y. Li, W. Li, Q. Wang, Q. Qin, X. Wang, M. Gong, H. Jiao, W. Liu, J. Jia, Cerebral blood flow in mild cognitive impairment and Alzheimer's disease: A systematic review and meta-analysis, Ageing Res. Rev. 71 (2021).
[28] C. Möller, Y.A.L. Pijnenburg, W.M. van der Flier, A. Versteeg, B.M. Tijms, J.C. de Munck, A. Hafkemeijer, S.A.R.B. Rombouts, J. van der Grond, J.C. van Swieten, E. G.P. Dopper, P. Scheltens, F. Barkhof, H. Vrenken, A.M. Wink, Alzheimer Disease and Behavioral Variant Frontotemporal Dementia: Automatic Classification Based on Cortical Atrophy for Single-Subject Diagnosis, Radiology 279 (3) (2016) 838–848.
[29] I. Beheshti, H. Demirel, Probability distribution function-based classification of structural MRI for the detection of Alzheimer's disease, Comput. Biol. Med. 64 (2015) 208–216.
[30] X. Hao, G. Zhang, S. Ma, Deep Learning, Int. J. Semantic Comput. 10 (2016) 417-.
[31] A. Abrol, Z. Fu, M.S. Salman, R.F. Silva, Y. Du, S. Plis, V. Calhoun, Deep learning encodes robust discriminative neuroimaging representations to outperform standard machine learning, Nat. Commun. 12 (2021).
[32] L. Alzubaidi, J. Zhang, A.J. Humaidi, A. Al-dujaili, Y. Duan, O. Al-Shamma, J. Santamaría, M.A. Fadhel, M. Al-Amidie, L. Farhan, Review of deep learning: concepts, CNN architectures, challenges, applications, future directions, J. Big Data 8 (2021).
[33] N.M. Bruckmann, J. Kirchner, L. Umutlu, W.P. Fendler, R.P. Seifert, K. Herrmann, A.-K. Bittner, O. Hoffmann, S. Mohrmann, C. Antke, L. Schimmöller, M. Ingenwerth, K. Breuckmann, A. Stang, C. Buchbender, G. Antoch, L.M. Sawicki, Prospective comparison of the diagnostic accuracy of 18F-FDG PET/MRI, MRI, CT, and bone scintigraphy for the detection of bone metastases in the initial staging of primary breast cancer patients, Eur. Radiol. 31 (2021) 8714–8724.
[34] R. Delgado-Escaño, F.M. Castro, N. Guil, V.S. Kalogeiton, M.J. Marín-Jiménez, Multimodal Gait Recognition Under Missing Modalities. In: ICIP 2021 (2021).
[35] J. Zhang, X. He, Y. Liu, Q. Cai, H. Chen, L. Qing, Multi-modal cross-attention network for Alzheimer's disease diagnosis with multi-modality data, Comput. Biol. Med. 162 (2023), 107050.
[36] Y.-H.-H. Tsai, S. Bai, P.P. Liang, J.Z. Kolter, L.-P. Morency, R. Salakhutdinov, Multimodal Transformer for Unaligned Multimodal Language Sequences, Proc. Conf. Assoc. Comput. Linguist Meet. 2019 (2019) 6558–6569.
[37] Y. Dai, B. Zou, C. Zhu, Y. Li, Z. Chen, Z. Ji, X. Kui, W. Zhang, DE-JANet: A unified network based on dual encoder and joint attention for Alzheimer's disease classification using multi-modal data, Comput. Biol. Med. (2023).
[38] L. Kang, J. Jiang, J. Huang, T. Zhang, Identifying Early Mild Cognitive Impairment by Multi-Modality MRI-Based Deep Learning, Front. Aging Neurosci. 12 (2020).
[39] T. Zhou, K.-H. Thung, M. Liu, F. Shi, C. Zhang, D. Shen, Multi-modal latent space inducing ensemble SVM classifier for early dementia diagnosis with neuroimaging data, Med. Image Anal. 60 (2020), 101630.
[40] S. Lahmiri, A. Shmuel, Performance of machine learning methods applied to structural MRI and ADAS cognitive scores in diagnosing Alzheimer's disease, Biomed. Signal Process. Control 52 (2019) 414–419.
[41] C.R. Jack, M.A. Bernstein, N.C. Fox, P. Thompson, G. Alexander, D. Harvey, B. Borowski, P.J. Britson, J.L. Whitwell, C. Ward, A.M. Dale, J.P. Felmlee, J. L. Gunter, D.L.G. Hill, R. Killiany, N. Schuff, S. Fox-Bosetti, C. Lin, C. Studholme, C. S. DeCarli, G. Krueger, H.A. Ward, G.J. Metzger, K.T. Scott, R. Mallozzi, D. Blezek, J. Levy, J.P. Debbins, A.S. Fleisher, M. Albert, R. Green, G. Bartzokis, G. Glover, J. Mugler, M.W. Weiner, J. Study, The Alzheimer's Disease Neuroimaging Initiative (ADNI): MRI methods, J. Magn. Reson. Imaging 27 (2008) 685–691.
[42] D.H. Harold, R. Abraham, P. Hollingworth, R. Sims, A. Gerrish, M.L. Hamshere, J. S. Pahwa, V. Moskvina, K. Dowzell, A.J. Williams, N. Jones, C. Thomas, A. Stretton, A.R. Morgan, S. Lovestone, J. Powell, P. Proitsi, M.K. Lupton, C. Brayne, D. C. Rubinsztein, M. Gill, B.A. Lawlor, A. Lynch, K. Morgan, K.S. Brown, P. Passmore, D. Craig, B. McGuinness, S. Todd, C. Holmes, D.M.A. Mann, A.D. Smith, S. Love, P. G. Kehoe, J. Hardy, S. Mead, N.C. Fox, M.N. Rossor, J. Collinge, W. Maier, F. Jessen, B. Schürmann, R. Heun, H. van den Bussche, I. Heuser, J. Kornhuber, J. Wiltfang, M. Dichgans, L. Frölich, H. Hampel, M. Hüll, D. Rujescu, A.M. Goate, J. S.K. Kauwe, C. Cruchaga, P. Nowotny, J.C. Morris, K. Mayo, K. Sleegers, K. Bettens, S. Engelborghs, Deyn P.P. De, Broeckhoven C. Van, G. Livingston, N.J. Bass, H.M. D. Gurling, A. McQuillin, R. Gwilliam, P. Deloukas, A. Al-Chalabi, C.E. Shaw, M. Tsolaki, A.B. Singleton, R. Guerreiro, T.W. Mühleisen, M.M. Nöthen, S. Moebus, K.-H. Jöckel, N. Klopp, H.-E. Wichmann, M.M. Carrasquillo, V.S. Pankratz, S. G. Younkin, P.A. Holmans, M.C. Donovan, M.J. Owen, J. Williams, Genome-wide association study identifies variants at CLU and PICALM associated with Alzheimer's disease, Nat Genet 45 (2013) 712.
[43] J. Chung, L.A. Farrer, A.D.N. Initiative, G.R. Jun, Genome-wide association study in different clinical stages of Alzheimer's disease, Alzheimer's & Dementia 11 (2015).
[44] C. James, J.M. Ranson, R. Everson, D.J. Llewellyn, Performance of Machine Learning Algorithms for Predicting Progression to Dementia in Memory Clinic Patients, JAMA Netw. Open 4 (2021) e2136553–e.
[45] R.V. Marinescu, N.P. Oxtoby, A.L. Young, E.E. Bron, A.W. Toga, M.W. Weiner, F. Barkhof, N.C. Fox, P. Golland, S. Klein, D.C. Alexander, TADPOLE Challenge: Accurate Alzheimer's disease prediction through crowdsourced forecasting of future data, Predict Intell Med 11843 (2019) 1–10.
[46] A.P. Association, Diagnostic and Statistical Manual of Mental Disorders (2022).
[47] N. Tustison, B. Avants, P. Cook, Y. Zheng, A. Egan, P. Yushkevich, J. Gee, N4ITK: Improved N3 Bias Correction, IEEE Trans. Med. Imaging 29 (2010) 1310–1320.
[48] P. Lu, L. Hu, N. Zhang, H. Liang, T. Tian, L. Lu, A Two-Stage Model for Predicting Mild Cognitive Impairment to Alzheimer's Disease Conversion, Front. Aging Neurosci. (2022) 14.
[49] R.M. Aziz, C.K. Verma, N. Srivastava, Dimension reduction methods for microarray data: a review (2017).
[50] H. Liu, R. Setiono, Incremental Feature Selection, Appl. Intell. 9 (2004) 217–230.
[51] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, P.-A. Manzagol, Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion, J. Mach. Learn. Res. 11 (2010) 3371–3408.
[52] Y. Si, J. Du, Z. Li, X. Jiang, T.A. Miller, F. Wang, W.J. Zheng, K. Roberts, Deep Representation Learning of Patient Data from Electronic Health Records (EHR): A Systematic Review, J. Biomed. Inform. 103671 (2020).
[53] A. Miech, I. Laptev, J. Sivic, Learnable pooling with Context Gating for video classification (2017), *ArXiv* abs/1706.0.
[54] Y. Dauphin, A. Fan, M. Auli, D. Grangier, Language Modeling with Gated Convolutional Networks, in: ICML (2017).
[55] X. Long, C. Gan, G. Melo, X. Liu, Y. Li, F. Li, S. Wen, Multimodal Keyless Attention Fusion for Video Classification, Proceedings of the AAAI Conference on Artificial Intelligence, 32 (2018).
[56] B. Efron, R. Tibshirani, An Introduction to the Bootstrap (1993).
[57] C. Wang, Y. Li, Y. Tsuboshita, T. Sakurai, T. Goto, H. Yamaguchi, Y. Yamashita, A. Sekiguchi, H. Tachimori, C.Y.T.W.L. Goto, C. Wang, Y. Li, A high-generalizability machine learning framework for predicting the progression of Alzheimer's disease using limited data, NPJ Digit. Med. (2022) 5.
[58] J. Ma, J. Zhang, Z. Wang, Multimodality Alzheimer's Disease Analysis in Deep Riemannian Manifold, Inf. Process. Manag. 59 (2022), 102965.
[59] Z. Ning, Q. Xiao, Q. Feng, W. Chen, Y. Zhang, Relation-Induced Multi-Modal Shared Representation Learning for Alzheimer's Disease Diagnosis, IEEE Trans. Med. Imaging 40 (2021) 1632–1645.
[60] W. Shao, Y. Peng, C. Zu, M. Wang, D. Zhang, Hypergraph based multi-task feature selection for multimodal classification of Alzheimer's disease, Comput. Med. Imaging Graph. 80 (2020), 101663.
[61] S.M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, Adv. Neural Inf. Process Syst. (2017) 30.
[62] A. Vaswani, N.M. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, Attention is All you Need (2017). *ArXiv* abs/1706.0.
[63] G. Lee, K. Nho, B. Kang, K. Sohn, D. Kim, Predicting Alzheimer's disease progression using multi-modal deep learning approach, Sci. Rep. 9 (2019).
[64] T. Zhou, M. Liu, K.-H. Thung, D. Shen, Latent Representation Learning for Alzheimer's Disease Diagnosis With Incomplete Multi-Modality Neuroimaging and Genetic Data, IEEE Trans. Med. Imaging 38 (2019) 2411–2422.
[65] N. Amoroso, D. Diacono, A. Fanizzi, M. La Rocca, A. Monaco, A. Lombardi, C. Guaragnella, R. Bellotti, S. Tangaro, A.D. Neuroimaging, Deep learning reveals Alzheimer's disease onset in MCI subjects: Results from an international challenge, J. Neurosci. Methods 302 (2018) 3–9.
[66] B. Lei, N. Cheng, A.F. Frangi, E.-L. Tan, J. Cao, P. Yang, A. Elazab, J. Du, Y. Xu, T. Wang, Self-calibrated brain network estimation and joint non-convex multi-task learning for identification of early Alzheimer's disease, Med. Image Anal. 61 (2020), 101652.

[67] J. Peng, X. Zhu, Y. Wang, L. An, D. Shen, Structured sparsity regularized multiple kernel learning for Alzheimer's disease diagnosis, Pattern Recogn. 88 (2019) 370–382.

[68] J. Peng, L. An, X. Zhu, Y. Jin, D. Shen, Structured sparse kernel learning for imaging genetics based alzheimer's disease diagnosis. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 9901 LNCS, (2016) 70–78.

[69] G. Lee, B. Kang, K. Nho, K. Sohn, D. Kim, MildInt: Deep Learning-Based Multimodal Longitudinal Data Integration Framework, Front. Genet. 10 (2019).