Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

# Longitudinal Alzheimer's Disease Classification from 2.5D MRI and Clinical Data via Adaptive Gated Fusion

**Sharvari Hemant Mohokar**
National Institute of Technology Karnataka, Surathkal
**Sowmya Kamath S.**
National Institute of Technology Karnataka, Surathkal
**Santosh Rai P V**
santosh.rai@manipal.edu
Kasturba Medical College Mangalore, Manipal Academy of Higher Education
**Malathi S Bhat**
Kasturba Medical College Mangalore, Manipal Academy of Higher Education

**Research Article**

**Keywords:**

**Posted Date:** November 18th, 2026

**DOI:** https://doi.org/10.21203/rs.3.rs-11605408/v1

**License:** This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

**Additional Declarations:** No competing interests reported.

# Longitudinal Alzheimer's Disease Classification from 2.5D MRI and Clinical Data via Adaptive Gated Fusion

Sharvari Hemant Mohokar[1], Sowmya Kamath S.[1], Santosh Rai P V[2*], Malathi S Bhat[2]

[1]Department of Information Technology, National Institute of Technology Karnataka, Surathkal, Mangalore, 575025, Karnataka, India.

[2*]Department of Radiodiagnosis and Imaging, Kasturba Medical College Mangalore, Manipal Academy of Higher Education, Mangalore, 575001, Karnataka, India.

*Corresponding author(s). E-mail(s): santosh.rai@manipal.edu;

**Abstract**

Reliable discrimination between cognitively stable and impaired individuals remains a clinically important challenge in Alzheimer's disease (AD) assessment. Existing multimodal deep learning approaches have predominantly relied on computationally intensive 3D convolutional networks, while relatively few studies jointly model longitudinal imaging and non-imaging data or explicitly account for patient-specific modality contributions. In this work, a multimodal longitudinal framework is proposed that combines 2.5D MRI feature extraction with bidirectional LSTM-based longitudinal sequence encoding. A residual gated fusion mechanism adaptively weights imaging and clinical features for each patient, enabling joint learning of longitudinal and multimodal representations. Structural MRI and clinical data from 204 patients across three longitudinal visits from the Alzheimer's Disease Neuroimaging Initiative (ADNI) were evaluated using patient-level 5-fold cross-validation to prevent information leakage across visits. The proposed framework achieved a mean AUC of $0.870 \pm 0.035$, balanced accuracy of 0.802, sensitivity of 0.756, and specificity of 0.847, consistently outperforming unimodal baselines and alternative architectural variants. Ablation studies demonstrated that longitudinal modeling, the 2.5D representation, and residual gated fusion each contributed independently to the overall performance. Quantitative Grad-CAM analysis further revealed cross-visit spatial consistency

and significantly greater medial activation in cognitively impaired individuals than in stable individuals, indicating that the model focuses on neuroanatomically plausible regions associated with neurodegenerative change. Our work underscores the potential of multimodal longitudinal deep learning to improve AD assessment while leveraging routinely collected imaging and clinical data.

**Keywords:** Alzheimer's disease, Longitudinal classification, Multimodal fusion, Bidirectional LSTM, Attention mechanisms, Explainability

# 1 Introduction

Alzheimer's disease (AD), characterized by progressive neurodegeneration, remains the most common cause of dementia worldwide [1]. Central to this progression is the accumulation of amyloid-$\beta$ plaques and tau tangles, which together drive the cognitive decline associated with this disease [2]. Between 1990 and 2021, the global prevalence of AD surged by 161%, rising from 21.8 million to 56.9 million cases, and is projected to reach a staggering 152.8 million by 2050 [3]. With AD expected to become the second leading cause of death by 2040 [3], the development of accurate and timely diagnostic tools has become a pressing global health priority. Early diagnosis is therefore critical, as molecular changes, including alterations in chromatin accessibility and neural stem cell vulnerability, occur long before clinical symptoms appear, during a preclinical stage when targeted interventions could still help preserve cognitive function [4]. However, current diagnostic methodologies often identify AD only after substantial disease progression [5]. Structural MRI primarily detects patterns of brain atrophy that may overlap with normal aging [2], while PET imaging is costly and remains relatively inaccessible [5]. Biomarker-based approaches may also be invasive or exhibit limited sensitivity and specificity [5]. These limitations highlight the need for diagnostic approaches that can achieve reliable early detection while reducing dependence on expensive, invasive or computationally intensive methods.

Deep learning has demonstrated considerable potential for early AD diagnosis, with CNN-based approaches using structural MRI achieving strong performance in distinguishing AD patients from cognitively normal individuals [6]. The integration of complementary information, including genetic, clinical and cognitive data, has further improved diagnostic performance compared with single-modality approaches [6]. Multimodal fusion has been explored at different levels, including raw feature-level, intermediate feature-level and decision-level fusion [7], as well as through cross-modal attention [8] and gating mechanisms [9] to model relationships between modalities. However, existing gating approaches primarily focus on cross-modal interactions and consistency, without explicitly accounting for the reliability of individual modalities. This is important in clinical settings, where the quality and diagnostic value of different modalities may vary across individuals.

In parallel, longitudinal approaches have sought to exploit the progressive nature of AD by incorporating information from multiple time points, with several studies reporting improved performance over single-timepoint models [10, 11]. Although

longitudinal information has also been incorporated into multimodal frameworks, most studies have focused on combining multiple imaging modalities rather than jointly modelling imaging and non-imaging information. Moreover, existing multimodal longitudinal approaches have predominantly relied on 3D CNNs for MRI feature extraction [11–13]. While 3D representations can capture volumetric information, they are computationally expensive and can be prone to overfitting, particularly when trained on relatively small clinical datasets. In contrast, 2.5D representations provide a computationally lighter alternative while retaining spatial information from multiple anatomical views, making them more suitable for efficient and potentially deployable diagnostic models [14, 15].

These observations reveal three underexplored directions in multimodal AD diagnosis: adaptive fusion mechanisms that account for the reliability of individual modalities, joint modelling of longitudinal imaging and non-imaging information, and the integration of computationally efficient 2.5D representations within longitudinal multimodal frameworks. Existing approaches often treat different modalities uniformly, do not fully exploit longitudinal information, or rely on computationally intensive 3D imaging representations. In this work, a multimodal longitudinal framework is proposed that combines 2.5D structural MRI representations with clinical, demographic and biomarker information through a reliability-aware residual gated fusion mechanism. The framework adaptively regulates the contribution of each modality based on its reliability while capturing temporal changes across longitudinal assessments. The resulting multimodal representation is used for binary classification of cognitively normal and cognitively impaired individuals.

The remainder of this paper is organised as follows. Section 2 reviews existing multimodal and longitudinal approaches for AD diagnosis. Section 3 presents the proposed reliability-aware multimodal longitudinal framework, including the 2.5D MRI representation and residual gated fusion mechanism. Section 4 presents the experimental evaluation and discusses the results. Section 5 concludes the paper and outlines directions for future work.

## 2 Background

Structural MRI has been widely used in deep learning-based approaches for AD diagnosis because it can capture structural changes in the brain associated with disease progression. Grueso and Viejo-Sobara [6] conducted a systematic review covering 116 machine learning studies on MCI-to-AD progression prediction, and found MRI to be the most frequently used single modality, appearing in 76 studies. They also reported that CNN-based methods achieved a mean accuracy approximately 3% higher than ML-based approaches, suggesting that deep learning can better capture the complex patterns present in MRI data [6]. Several studies have since used CNN-based architectures for MRI-based classification of different AD stages, using both models trained from scratch and transfer learning, with promising results [16, 17]. However, MRI captures only one aspect of the disease. Information from clinical assessments, cognitive tests, genetic data and biomarkers can provide additional information that may not be evident from structural changes alone. This has led to increasing interest in multimodal

approaches that combine imaging with complementary non-imaging information for AD diagnosis.

Despite the effectiveness of MRI-based approaches, incorporating complementary information such as genetic data, neuropsychological test scores and demographic characteristics has been shown to improve performance over single-modality models, with deep learning approaches achieving the strongest overall results [6, 18]. This has motivated a shift from single-modality classifiers towards multimodal architectures that learn complementary representations from heterogeneous data sources. Venugopalan et al. [7], for example, proposed a multimodal framework using structural MRI, electronic health records (EHR) and single nucleotide polymorphism (SNP) data from the ADNI cohort to classify subjects into CN, MCI and AD groups. Modality-specific representations were learned using 3D CNNs for MRI and stacked denoising autoencoders for clinical and SNP data, followed by feature concatenation for classification. The study showed that deeper models consistently outperformed shallow baselines, while multimodal deep architectures further improved performance over their single-modality counterparts.

A key design consideration in multimodal diagnosis is therefore how information from different modalities should be integrated. Venugopalan et al. [7] evaluated raw feature-level, intermediate feature-level and decision-level fusion and found intermediate feature-level fusion to provide the best performance. The advantage was attributed to its ability to learn relationships between representations from different modalities. More recent approaches have moved beyond simple concatenation towards mechanisms that explicitly model cross-modal interactions [19, 20]. Zhang et al. [8] took a different approach with MCAD, a framework that applies cross-modality attention to integrate structural MRI, FDG-PET and CSF biomarkers. A multimodal interaction module based on cross-attention was used to capture relationships between imaging and CSF representations, rather than independently extracting and concatenating modality-specific features. The proposed approach outperformed simple concatenation across the evaluated diagnostic tasks, with the improvement becoming more pronounced for more challenging classification tasks. These findings indicate that explicitly modelling cross-modal interactions can better exploit complementary information than conventional concatenation-based fusion. However, such approaches primarily focus on learning inter-modality relationships, without explicitly considering whether the information provided by each modality is reliable for a given subject. This motivates the need for fusion mechanisms that can adapt the contribution of individual modalities according to their reliability.

A related approach to multimodal fusion is the use of gating mechanisms, which control the contribution of different modalities rather than treating all modality features equally. To jointly address AD classification and cognitive score estimation, Tang et al. [9] proposed CGMF-Net, a muti-task learning framework built around cross modal gating of MRI and FDG-PET data. The framework introduces a Cross-Modal Similarity Gate (CMSGate) that enhances feature responses where structural and metabolic information is consistent across the two modalities. Specifically, cross-modal similarity is used to residually amplify regions with high structural-metabolic agreement before multimodal fusion. Their ablation experiments showed a drop in

performance when CMSGate was removed, indicating the importance of cross-modal consistency in learning discriminative representations. However, gating mechanisms remain relatively underexplored in multimodal AD diagnosis, and existing approaches have primarily used gating to capture cross-modal consistency rather than to assess the reliability of individual modalities. This distinction is important in clinical datasets, where the availability and quality of different modalities can vary across subjects, with some clinical assessments or biomarkers missing for certain patients. A fusion mechanism that can adaptively regulate the contribution of each modality based on its reliability could therefore provide a more robust approach to multimodal AD diagnosis.

Most of the multimodal approaches discussed in this review treat AD diagnosis as a single-timepoint classification problem, using data from an individual visit to make a prediction. This does not fully capture the progressive nature of AD, where changes in brain structure, cognitive function and other clinical measures over time can provide important information about disease progression. To address this limitation, recent studies have explored longitudinal models that incorporate data from multiple visits and have reported improvements over single-timepoint approaches [10, 11]. Jomeiri et al. [12] proposed a hybrid DenseNet-BiLSTM framework that uses longitudinal MRI data collected over four years to capture both spatial features and temporal changes. Their results highlight the value of modeling sequential MRI data for improving diagnostic performance and capturing patterns associated with disease progression. Longitudinal modeling has also been extended to multimodal settings, although most existing studies have focused on combining multiple imaging modalities across visits rather than integrating imaging with non-imaging information [10, 11]. The joint modelling of longitudinal imaging and clinical, demographic or other non-imaging information therefore remains relatively underexplored. A notable recent example is the work of Li et al. [13], which combines a pre-trained Brain LDM with a Longitudinal Tabular Transformer and a mixture-of-experts (MoE) fusion module for AD diagnosis. The framework uses irregularly sampled longitudinal structural MRI together with heterogeneous tabular clinical attributes and demonstrates the potential of jointly modelling longitudinal imaging and clinical information for AD classification. However, such approaches remain computationally demanding, particularly due to the use of volumetric 3D imaging representations, motivating the exploration of more computationally efficient representations for longitudinal multimodal AD diagnosis.

Another limitation of existing multimodal longitudinal studies is their reliance on 3D convolutional architectures for extracting features directly from volumetric MRI data [11–13]. Although 3D CNNs preserve the complete spatial context of MRI volumes, they require substantial computational resources and can be prone to overfitting when trained on the relatively small cohorts typical of AD datasets. Developing more computationally efficient representations is therefore important for building models that are practical for resource-constrained clinical settings. A review by Li et al. [14], synthesizing findings from 31 studies, found that among the 21 studies comparing 2.5D and 3D architectures, 12 reported better performance with 2.5D models. Soelistiono [15] further demonstrated the potential of 2.5D representations for longitudinal classification through a modified ResNet architecture that uses multiple adjacent slices to capture spatial context while reducing the computational cost of full 3D processing.

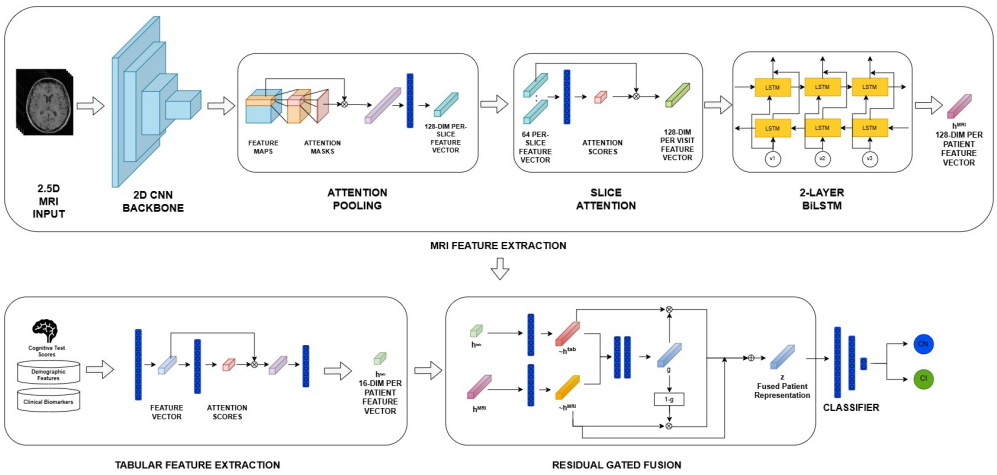



**Fig. 1** Proposed multimodal architecture for Alzheimer's Disease classification

Thus, 2.5D representations can provide a useful balance between spatial information and computational efficiency. Despite this potential, the use of 2.5D representations in longitudinal multimodal AD diagnosis remains largely unexplored, particularly when combined with fusion mechanisms that can account for differences in modality reliability. Taken together, the literature highlighted three important gaps: the continued reliance on computationally intensive 3D architectures in multimodal longitudinal models, the limited integration of longitudinal imaging with non-imaging information, and the lack of fusion mechanisms that adaptively regulate modality contributions based on their reliability. In this work, a multimodal longitudinal framework is proposed for binary AD classification that combines 2.5D structural MRI representations with demographic, clinical and biomarker information through a reliability-aware gated fusion mechanism. Our work aims to classify individuals as cognitively normal or cognitively impaired while providing a computationally efficient approach to integrating heterogeneous longitudinal information.

## 3 Materials and Methods

The overall architecture of the proposed framework is illustrated in Fig. 1. The framework is designed to jointly model longitudinal structural MRI and non-imaging information for binary classification of cognitively normal and cognitively impaired individuals. It consists of two parallel branches: an MRI branch that extracts spatial and temporal representations from longitudinal 2.5D MRI data, and a clinical branch that encodes longitudinal demographic, clinical and biomarker information. The representations learned from the two branches are subsequently integrated using a reliability-aware residual gated fusion mechanism, which adaptively regulates the contribution of each modality before the final classification layers.

## 3.1 Dataset Specifics

All experiments were conducted using data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) [21]. Patients were included if T1-weighted structural MRI scans were available across at least three longitudinal visits, together with corresponding demographic, clinical and biomarker information (Table 1), resulting in a cohort of 204 patients. To establish a consistent temporal structure across patients, visit timing was discretized based on the elapsed time from each patient's baseline MRI and assigned to annual bins. When multiple observations occurred within the same yearly bin, the earliest observation was retained. Patients were further required to have observations spanning at least three distinct yearly bins to ensure sufficient longitudinal information for temporal modeling.

**Table 1** Clinical and Demographic Features Used

| Sl. No. | Variable | Description |
|---|---|---|
| 1 | AGE | Age |
| 2 | SEX | Biological Sex |
| 3 | EDUCATION | Years of Education |
| 4 | MMSE | Mini-Mental State Examination |
| 5 | ADAS-13 | AD Assessment Scale (13-item) |
| 6 | CDRSB | Clinical Dementia Rating Sum of Boxes |
| 7 | FAQ | Functional Activities Questionnaire |
| 8 | APOE4 | Number of $\epsilon 4$ alleles of APOE |
| 9 | RAVLT Immediate | Rey Auditory Verbal Learning Test |
| 10 | RAVLT Learning | Rey Auditory Verbal Learning Test |
| 11 | RAVLT Forgetting | Rey Auditory Verbal Learning Test |
| 12 | RAVLT % Forgetting | Rey Auditory Verbal Learning Test |
| 13 | MOCA | Montreal Cognitive Assessment |
| 14 | AB42 | Amyloid Beta 42 |
| 15 | TAU | Total Tau Protein |
| 16 | PTAU | Phosphorylated Tau Protein |
| 17 | HCI | Hippocampal Composite Index |
| 18 | SROI_AD | Structural ROI (Alzheimer's Disease) |
| 19 | SROI_MCI | Structural ROI (Mild Cognitive Impairment) |

Each patient was assigned a single diagnostic label based on their ADNI-designated research group, namely, cognitively normal (CN), Mild Cognitive Impairment (MCI), or Alzheimer's Disease (AD). This label reflects each patient's overall diagnostic classification within ADNI and was applied uniformly across all three of their longitudinal visits. The three visits provide repeated, correlated imaging and clinical evidence of a single, fixed underlying diagnostic category for each patient. Patients were categorized as cognitively normal (CN; $n = 129$) or cognitively impaired (MCI or AD; $n = 75$, comprising 59 MCI and 16 AD patients). The MCI and AD categories were merged into a single label to mitigate the severe class imbalance that would otherwise arise from treating the three diagnostic categories separately. This binary formulation targets the clinically relevant task of distinguishing cognitively normal individuals from

those exhibiting Alzheimer's disease-related cognitive impairment, regardless of severity stage, using repeated longitudinal observations to inform a single, patient-level diagnostic classification.

## 3.2 Data Preprocessing

Missing values were handled after the train-test split to avoid information leakage. For selected clinical, biomarker and less frequently recorded cognitive assessment variables, missing observations were first addressed using intra-patient forward and backward filling to preserve temporal continuity. For patients with incomplete biomarker series that could not be resolved through this procedure, median imputation was performed using the training data. Biomarker variables were additionally winsorized using the training-set interquartile range, with values clipped to $[Q1 - 1.5, IQR, , Q3 + 1.5, IQR]$ to reduce the influence of extreme observations. All continuous tabular features were subsequently standardized using a scaler fitted on the training data. These steps produced temporally consistent tabular sequences while preventing information from the test set from influencing preprocessing.

Structural T1-weighted MRI scans were obtained in DICOM format and converted to compressed NIfTI format using dcm2niix [22]. Skull stripping was performed using FSL BET [23] with a fractional intensity threshold of 0.5 to remove non-brain tissue while retaining the cortical boundaries. The resulting volumes were spatially normalized to the MNI152 standard space using FSL FLIRT [24]. Bias field correction was then performed using FSL FAST [25] to reduce intensity inhomogeneity arising from MRI acquisition. The corrected volumes were reoriented to the standard anatomical convention using `fslreorient2std`. To reduce the computational cost associated with full 3D convolution while retaining information from neighboring slices, a 2.5D representation was used for MRI feature extraction. For each scan, five adjacent slices were considered at each spatial position. The resulting volumes were Z-score normalized and center-cropped to 64 positions, producing a tensor of size $(64, 5, 109, 91)$ for each MRI scan. This representation provides local volumetric context through adjacent slices while substantially reducing the computational requirements of processing the complete 3D volume.

## 3.3 MRI Feature Extraction

The proposed framework processes each patient's longitudinal multimodal data through two parallel branches: an MRI branch and a clinical branch. The MRI branch extracts spatial features from the 2.5D representations at each visit and models their temporal evolution using a bidirectional LSTM, while the clinical branch encodes longitudinal tabular features using visit-level attention. The resulting representations capture complementary information from imaging and non-imaging data and are subsequently combined through a residual gated fusion module, followed by fully connected layers for classification. Both branches share parameters across visits, allowing consistent feature extraction across the longitudinal sequence while limiting the number of trainable parameters.

For each visit, the preprocessed 2.5D MRI tensor of size $(64, 5, 109, 91)$ was processed through a spatiotemporal feature extraction pipeline. The pipeline first uses a 2D CNN backbone to extract spatial features from the adjacent-slice groups, followed by attention-based pooling to aggregate spatial information. Slice attention is then used to assign greater importance to informative slices within each 2.5D representation. The resulting visit-level features are arranged in chronological order and passed through a bidirectional LSTM to capture temporal dependencies across the three visits. The individual components of this pipeline are described below.

***CNN Backbone.***

A four-block 2D CNN backbone was adopted for spatial feature extraction. Each block consisted of two $3 \times 3$ convolutional layers with batch normalization and ReLU activation, followed by max pooling and dropout layers. The channel depth progressively increased from 32 to 256 across the blocks, enabling hierarchical feature extraction ranging from low-level tissue boundaries in the early layers to high-level structural representation in the deeper layers. Each of the 64 positions was represented as a stack of 5 adjacent slices and was treated as input channels to the CNN backbone, allowing the network to incorporate local spatial context without the computational burden on full 3D convolutions.

***Attention Pooling***

The spatial feature map computed by the CNN backbone was passed through an attention pooling module. Two successive $1 \times 1$ convolutions were applied to generate a single-channel spatial score map $s_{ij}$, which was normalized across all spatial locations using the softmax function to produce attention weights $\alpha_{ij}$. The attended feature vector was computed as a weighted sum over all spatial locations as follows:

$$\alpha_{ij} = \frac{\exp(s_{ij})}{\sum_{i'j'} \exp(s_{i'j'})}, \qquad \mathbf{f} = \sum_{ij} \alpha_{ij} \cdot F_{ij} \tag{1}$$

where $F_{ij}$ is the feature vector at spatial location (i, j) and $s_{ij}$ is the corresponding scalar score. This allowed the model to give more importance to regions associated with early Alzheimer's atrophy while suppressing non-discriminative background regions. Subsequently, a fully connected layer projected each attended vector to 128 dimensions, producing a slice-level representation $\mathbf{f}_t^{(s)} \in \mathbb{R}^{128}$ for each of the 64 slices, where $s$ denotes the slice index and $t$ denotes the visit index.

***Slice Attention***

To identify the most discriminative levels within each volume, a slice attention mechanism was applied over the 64 slice feature vectors. A fully connected layer mapped each slice representation to a scalar importance score, which was normalized using the softmax function across all slices to obtain attention weights $\beta_s$. The visit-level MRI representation was computed as a weighted aggregation of the slice features as follows:

$$\beta_s = \frac{\exp(\mathbf{w}^\top \mathbf{f}_t^{(s)})}{\sum_{s'} \exp(\mathbf{w}^\top \mathbf{f}_t^{(s')})}, \qquad \mathbf{v}_t = \sum_s \beta_s \cdot \mathbf{f}_t^{(s)} \tag{2}$$

where $\mathbf{w} \in \mathbb{R}^{128}$ is a learned weight vector and $\mathbf{f}_t^{(s)}$ is the feature vector of slice $s$ at visit $t$. This allowed the model to emphasize anatomically relevant slices while suppressing less informative regions. Each MRI volume per visit was represented by $\mathbf{v}_t \in \mathbb{R}^{128}$, where $t \in \{1, 2, 3\}$ denotes the visit index. Across three visits, this yielded the sequence $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$ for temporal modeling.

***BiLSTM***

To aggregate the three per-visit MRI representations into a single, more robust patient-level descriptor, the sequence $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$ was processed by a two-layer bidirectional LSTM with a hidden size of 64 for each direction. The two-layer architecture enables hierarchical encoding of the longitudinal sequence, with the lower layers capturing local visit-to-visit dependencies while the upper layers integrate information across the full sequence. The forward and backward hidden states from the final time step of the last layer were concatenated, resulting in a 128-dimensional MRI temporal representation, $\mathbf{h}^{\text{MRI}} \in \mathbb{R}^{128}$, per patient. Since the diagnostic label does not change across visits, the BiLSTM functions as a mechanism for aggregating correlated, repeated measurements of the same underlying condition into a single, richer representation that is more robust than any individual visit alone.

## 3.4 Tabular Feature Extraction

Along with structural MRI, each patient visit was associated with a 19-dimensional vector of clinical tabular features $\mathbf{x}_t \in \mathbb{R}^{19}$ comprising demographic information, cognitive assessment scores, and clinical biomarkers. These features capture complementary disease indicators that are not directly observable from MRI intensity patterns alone. The tabular features at each visit were passed through a shared fully connected layer with ReLU activation to produce a visit-level representation $\mathbf{e}_t \in \mathbb{R}^{32}$. The same encoder parameters were used across all visits to maintain a consistent representation of clinical features over time and limit the number of trainable parameters. Dropout was applied to the encoded representations to reduce overfitting, which is particularly important given the relatively small cohort size.

Not all visits contribute equally to the diagnostic signal, as the clinical significance of a particular timepoint may vary across patients depending on when meaningful changes occur. A visit-level attention mechanism was therefore used to assign learnable importance weights $\gamma_t$ to the encoded visit representations $\mathbf{e}_t$ and combine them into a single patient-level clinical representation, as defined in Eq. (3). Here, $\mathbf{u} \in \mathbb{R}^{32}$ denotes a learned context vector. The attention weights $\gamma_1, \gamma_2, \gamma_3$ form a probability distribution over the three visits, allowing the model to assign greater weight to visits that provide more discriminative clinical information for the diagnostic task. The resulting weighted representation $\mathbf{c}$ was subsequently passed through a fully connected layer with ReLU activation to obtain the final tabular representation, as defined in

Eq. (4), where $\mathbf{W}_p \in \mathbb{R}^{16\times 32}$. The resulting $\mathbf{h}^{\text{tab}} \in \mathbb{R}^{16}$ provides a compact representation of the patient's longitudinal clinical profile, complementing the structural information captured by $\mathbf{h}^{\text{MRI}}$.

$$\gamma_t = \frac{\exp(\mathbf{u}^\top \mathbf{e}_t)}{\sum_{t'=1}^{3} \exp(\mathbf{u}^\top \mathbf{e}_{t'})}, \qquad \mathbf{c} = \sum_{t=1}^{3} \gamma_t \cdot \mathbf{e}_t \tag{3}$$

$$\mathbf{h}^{\text{tab}} = \text{ReLU}(\mathbf{W}_p \mathbf{c} + \mathbf{b}_p) \tag{4}$$

## 3.5 Residual Gated Fusion

The MRI representation $\mathbf{h}^{\text{MRI}}$ and tabular representation $\mathbf{h}^{\text{tab}}$ encode complementary information about disease progression: structural MRI captures the volumetric atrophy patterns while tabular features capture the cognitive and biomarker trajectories. Naive concatenation assigns equal weight to both modalities regardless of their relative reliability for a given patient. We therefore adopted a residual gated fusion mechanism that learns to modulate the contribution of each modality adaptively based on the joint representation, while guaranteeing minimum contribution from the MRI stream. Both representations were first projected into a common 128-dimensional space using linear layers followed by layer normalization, as defined in Eq. (5). Here, $\mathbf{W}_m \in \mathbb{R}^{128\times 128}$ and $\mathbf{W}_t \in \mathbb{R}^{128\times 16}$ denote the learned projection matrices, and LN represents layer normalization. The projected representations were then concatenated and used to compute the gating vector in Eq. (6). The gating network consists of two linear transformations, with $\mathbf{W}_{g1} \in \mathbb{R}^{128\times 256}$ and $\mathbf{W}_{g2} \in \mathbb{R}^{128\times 128}$, followed by a sigmoid activation $\sigma$. This constrains each element of $\mathbf{g}$ to $(0, 1)$, allowing the network to adaptively regulate the contribution of the two modalities during fusion.

$$\tilde{\mathbf{h}}^{\text{MRI}} = \text{LN}(\mathbf{W}_m \mathbf{h}^{\text{MRI}}), \qquad \tilde{\mathbf{h}}^{\text{tab}} = \text{LN}(\mathbf{W}_t \mathbf{h}^{\text{tab}}) \tag{5}$$

$$\mathbf{g} = \sigma\Big(\mathbf{W}_{g2}\, \text{ReLU}\Big(\mathbf{W}_{g1}\Big[\tilde{\mathbf{h}}^{\text{MRI}};\, \tilde{\mathbf{h}}^{\text{tab}}\Big] + \mathbf{b}_{g1}\Big) + \mathbf{b}_{g2}\Big) \tag{6}$$

The fused representation is computed as given in Eq. (7), where $\odot$ denotes element-wise multiplication and $\mathbf{z}$ represents the resulting patient-level representation. The formulation assigns an effective weight in $(1, 2)$ to the MRI representation and a weight in $(0, 1)$ to the tabular representation. Thus, the MRI branch always retains a higher contribution to the fused representation, while the gate adaptively adjusts the relative contribution of the two modalities based on the learned representations. This asymmetric design reflects the clinical importance of structural MRI as a direct measure of neurodegenerative changes that remain associated with cognitive decline across MCI and dementia stages [26]. The residual connection also provides a direct gradient pathway to the MRI branch, supporting stable optimisation during training. The gate $\mathbf{g}$ therefore does not determine the absolute contribution of each modality independently, but adaptively regulates their relative balance while maintaining MRI as the dominant modality. This provides a controlled form of reliability-aware fusion in which the contribution of the tabular representation can be increased or reduced according to the learned gate without allowing it to dominate the MRI representation.

$$\mathbf{z} = \mathbf{g} \odot \tilde{\mathbf{h}}^{\text{MRI}} + (1 - \mathbf{g}) \odot \tilde{\mathbf{h}}^{\text{tab}} + \tilde{\mathbf{h}}^{\text{MRI}} \tag{7}$$

**Table 2** Performance of the proposed multimodal architecture with residual gated fusion across 5-fold cross-validation

| Fold no. | AUC | Sens | Spec | Prec | F1 CI | F1 Macro | Acc | Bal Acc |
|---|---|---|---|---|---|---|---|---|
| Fold 1 | 0.8972 | 0.7143 | 0.9474 | 0.8824 | 0.7895 | 0.8447 | 0.8644 | 0.8308 |
| Fold 2 | 0.8108 | 0.7143 | 0.7895 | 0.6522 | 0.6818 | 0.7463 | 0.7627 | 0.7519 |
| Fold 3 | 0.9118 | 0.8500 | 0.8684 | 0.7727 | 0.8095 | 0.8507 | 0.8621 | 0.8592 |
| Fold 4 | 0.8605 | 0.6500 | 0.9474 | 0.8667 | 0.7429 | 0.8159 | 0.8448 | 0.7987 |
| Fold 5 | 0.8711 | 0.8500 | 0.6842 | 0.5862 | 0.6939 | 0.7350 | 0.7414 | 0.7671 |
| Mean ± Std | 0.870 ± 0.035 | 0.756 ± 0.080 | 0.847 ± 0.100 | 0.752 ± 0.117 | 0.744 ± 0.050 | 0.799 ± 0.049 | 0.815 ± 0.052 | 0.802 ± 0.040 |

### 3.6 AD Classification

The fused representation $\mathbf{z}$ was passed through a three-layer classifier comprising fully connected layers with ReLU activations and dropout regularization of rate 0.3, with output dimensions 128, 64, and 2, respectively. A softmax activation produced the final probability distribution over the cognitively normal and cognitively impaired classes.

## 4 Experiments and Results

### 4.1 Experimental Setup

All experiments were conducted using PyTorch on a single NVIDIA Tesla T4 GPU. The proposed model has a relatively low model complexity of 1.53M trainable parameters, which is particularly relevant for clinical applications where computational resources may be limited. The model was optimized using AdamW with a learning rate of $1 \times 10^{-4}$ and weight decay of $3 \times 10^{-4}$. A cosine annealing learning rate scheduler was used to gradually reduce the learning rate to $1 \times 10^{-6}$ during training. To address class imbalance, class-weighted cross-entropy loss was used, with weights of 0.5 and 2.5 assigned to the cognitively normal and cognitively impaired classes, respectively. Label smoothing of 0.1 was additionally applied to reduce overconfident predictions and improve model calibration. Training was performed for a maximum of 60 epochs, with early stopping based on validation AUC and a patience of 20 epochs.

Model performance was evaluated using stratified five-fold cross-validation. For each fold, the training portion was further divided into training and validation subsets using an 80:20 split, with the validation set used for early stopping and model selection. The primary evaluation metric was the area under the receiver operating characteristic curve (AUC-ROC), as it provides a threshold-independent measure of discrimination and is suitable for imbalanced datasets. Sensitivity, specificity, balanced accuracy, precision and F1-score were also reported to provide a more comprehensive assessment of classification performance, particularly for the cognitively impaired class.

### 4.2 Results and Discussion

Table 2 shows the per-fold and aggregate performance of the proposed model under 5-fold cross-validation. The model achieved a mean AUC of 0.870 ± 0.035, with the low standard deviation indicating consistent generalization across partitions despite the limited cohort size. The mean balanced accuracy of 0.802 ± 0.040 confirms that the model maintains discriminative capacity across both classes. The mean sensitivity

**Table 3** Ablation study results across 5-fold cross-validation

| Variant | AUC | AUC Std | Sens | Spec | Prec | Acc | Bal Acc |
|---|---|---|---|---|---|---|---|
| **Full Model** | **0.870** | **0.035** | **0.756** | **0.847** | **0.752** | **0.815** | **0.802** |
| w/o Attention Pooling | 0.842 | 0.046 | 0.620 | 0.811 | 0.707 | 0.743 | 0.715 |
| w/o Slice Attention | 0.889 | 0.081 | 0.719 | 0.775 | 0.624 | 0.702 | 0.719 |
| w/o BiLSTM | 0.845 | 0.083 | 0.695 | 0.811 | 0.676 | 0.771 | 0.753 |
| w/o Residual Gated Fusion | 0.858 | 0.090 | 0.892 | 0.584 | 0.585 | 0.693 | 0.738 |
| w/o Residual Fusion | 0.854 | 0.066 | 0.729 | 0.716 | 0.593 | 0.719 | 0.722 |

of 0.756 ± 0.080 and specificity of 0.847 ± 0.100 reflect that despite class-weighted loss favoring cognitively impaired class, cognitively normal cases remain comparatively easier to discriminate. The mean F1-score for the cognitively impaired class (0.744 ± 0.050) supports this, demonstrating the model's capacity to retrieve true cognitively impaired cases with reasonable precision despite their under-representation in training. Thus, it can be concluded that the proposed multimodal longitudinal architecture captures complementary diagnostic signals from both structural MRI and clinical trajectories, yielding robust and clinically meaningful discriminative performance.

To evaluate the contribution of each architectural component in the pipeline, ablation studies were performed by removing and replacing individual modules from the full model. Table 3 reports the mean performance of each variant across the 5-fold cross-validation. We noticed that replacing attention pooling with standard max pooling resulted in a drop in sensitivity, from 0.756 to 0.620. This indicates that the model loses focus on diagnostically relevant regions without spatial attention, highlighting its importance for identifying cognitively impaired individuals. A marginally higher AUC was observed on replacing the slice attention module with mean slice pooling. However, this came at the cost of substantially greater variance and lower balanced accuracy, indicating instability across folds and sensitivity to patient distribution within each fold. This showcases the importance of slice attention for producing stable and reliable model by selectively weighting diagnostically informative slices.

Replacing the BiLSTM with a unidirectional LSTM reduced AUC from 0.870 to 0.845 and balanced accuracy from 0.802 to 0.753. This indicates that the bidirectional LSTM extracts more informative evidence from the three-visit sequence than a unidirectional LSTM, likely because the backward direction allows earlier visits to be contextualized by later observations, rather than being encoded in isolation before later visits are seen. Replacing the residual gated fusion with simple concatenation resulted in the largest drop in specificity, despite the marginal gains in sensitivity. This reflects that without adaptive modality weighting, the model becomes overconfident in predicting the cognitively impaired class, generating higher number of false positives. The residual gated fusion helps in balancing sensitivity and specificity, yielding more reliable predictions across both the classes. Removing only the residual MRI connection while preserving the learned gate reduced AUC from 0.870 to 0.854, specificity from 0.847 to 0.716, precision from 0.752 to 0.593, and balanced accuracy from 0.802 to 0.722, while AUC variance nearly doubled from 0.035 to 0.066. One might expect simple concatenation and the gated-without-residual variant to behave similarly, since neither guarantees a minimum MRI contribution, but they do not behave the same way at all. Replacing the fusion mechanism with simple concatenation led to

an increase in sensitivity from 0.756 to 0.892, while specificity fell sharply from 0.847 to 0.584. In contrast, removing only the residual term produced a smaller decrease in sensitivity, while the specificity drop was considerably smaller than under simple concatenation. This indicates that while simple concatenation causes the model to generate substantially more false positives, the gated-without-residual variant instead produces a more conservative model. The residual term therefore helps keep the gate's behavior predictable and stable, as without it, the AUC variance across folds was found to nearly double, indicating that the model's behavior becomes considerably less consistent from one patient split to the next. Taken together, it can be concluded that each proposed component in the proposed pipeline contributes meaningfully to the overall model performance.

## 4.3 Comparison with Unimodal Baselines

To assess the contribution of multimodal fusion, two unimodal baselines were evaluated under identical 5-fold cross-validation conditions and training hyperparameters. The MRI-only model retained the full imaging branch with BiLSTM temporal modeling but completely removed the clinical branch, while the clinical-only model retained the tabular branch with visit attention but removed the MRI branch. As can be seen from Table 4, the MRI-only baseline achieved only 0.693 AUC with a high standard deviation of 0.177 and a balanced accuracy of 0.672, reflecting substantial instability across folds. In contrast, the clinical-only model performed considerably better, achieving an AUC of 0.862 with a much lower standard deviation, indicating that longitudinal clinical trajectories alone provide strong and stable discriminative capacity. Although the clinical-only model achieved a comparable AUC and lower variance than the proposed multimodal framework, the full model still outperformed it on specificity, precision, and balanced accuracy. While MRI features are insufficient as a standalone predictor, they contribute complementary information when combined with clinical data, helping the model reduce false positives that the clinical-only model alone could not resolve.

**Table 4** Comparison with Unimodal Baselines

| Variant | AUC | AUC Std | Sens | Spec | Bal Acc | Prec |
|---|---|---|---|---|---|---|
| **Full Model** | **0.870** | **0.035** | **0.756** | **0.847** | **0.802** | **0.752** |
| MRI-only | 0.693 | 0.177 | 0.745 | 0.600 | 0.672 | 0.620 |
| Clinical-only | 0.862 | 0.018 | 0.737 | 0.763 | 0.750 | 0.632 |

## 4.4 Effect of MRI Input Representation

To evaluate the effect of MRI input representation, the 2.5D MRI representation, which incorporates five adjacent slices to capture local volumetric context, was replaced with a 2D MRI representation, comprising a single 2D slice for each location. All other architectural components were held constant. From Table 5, it can be observed that

**Table 5** Effect of 2.5D MRI Input Representation

| Variant | AUC | AUC Std | Sens | Spec | Bal Acc | Prec |
|---|---|---|---|---|---|---|
| **Our model** | **0.870** | **0.035** | **0.756** | **0.847** | **0.802** | **0.752** |
| 2D MRI | 0.819 | 0.109 | 0.708 | 0.737 | 0.722 | 0.593 |

**Table 6** Effect of Longitudinal Modeling

| Variant | AUC | AUC Std | Sens | Spec | Bal Acc | Prec |
|---|---|---|---|---|---|---|
| **Our model** | **0.870** | **0.035** | **0.756** | **0.847** | **0.802** | **0.752** |
| Static classifier | 0.761 | 0.060 | 0.666 | 0.626 | 0.646 | 0.496 |

every metric was lower for the 2D variant compared to the proposed 2.5D model, with mean AUC falling from 0.870 to 0.819 and balanced accuracy from 0.802 to 0.722. Beyond this overall decline, the 2D variant also exhibited substantially greater variability across folds, with the standard deviation for AUC being roughly three times higher than that of the 2.5D model. Thus, removing the five adjacent slices from the MRI representation does not merely lower average performance but also destabilizes the model across different patient partitions, likely because a single slice provides insufficient local context for consistent feature extraction. This is the advantage that 2.5D representations offer over plain 2D slices, providing richer spatial context while remaining substantially more computationally efficient than full 3D volumetric processing.

## 4.5 Effect of Longitudinal Modeling

To assess the importance of longitudinal modeling, a static classifier was developed containing all architectural components except the BiLSTM, using only the baseline visit for each patient. Table 6 presents the results, where it is clear that performance dropped across every metric for the static classifier, with mean AUC falling from 0.870 to 0.761 and balanced accuracy from 0.802 to 0.646, while the standard deviation for AUC nearly doubled, indicating higher instability across folds. Notably, the drop in specificity and precision was considerably larger than the drop in sensitivity, indicating an increase in false positives, with the static classifier predicting cognitively normal individuals as cognitively impaired. This suggests that, without longitudinal modeling, the model was unable to reliably distinguish stable, cognitively normal patients from those reflecting genuine cognitive impairment based on a single visit alone. This confirms the importance of the BiLSTM in aggregating repeated, correlated evidence across visits into a single, more reliable representation of a patient's diagnostic condition than any one visit could provide in isolation.

## 4.6 Interpretability Analysis

To assess whether the model's predictions are grounded in neuroanatomically meaningful spatial features, Grad-CAM visualizations were generated for representative

samples. Fig. 2 illustrates activation maps for a representative true positive cognitively impaired case across three longitudinal visits. Across all three visits, activations were consistently localized to the inferior and medial temporal regions, encompassing the hippocampus and entorhinal cortex, structures known to exhibit early neurodegeneration in Alzheimer's disease [26]. While illustrative rather than conclusive on its own, this pattern was further examined quantitatively across the full test cohort. First, cross-visit spatial consistency was measured by computing the cosine similarity between Grad-CAM activation maps across visit pairs for each patient. The mean cross-visit consistency was $0.826 \pm 0.062$, indicating that the model attends to stable structural regions across longitudinal visits rather than visit-specific artifacts. This is consistent with the intended role of the BiLSTM, which aggregates repeated evidence of the same underlying condition across visits rather than tracking a changing outcome.

The spatial centroid of high-activation regions, defined as the top 20% of Grad-CAM activations, was compared between cognitively normal and cognitively impaired patients. Cognitively impaired patients exhibited a significantly lower mean W-centroid ($31.9 \pm 8.1$ pixels) than cognitively normal individuals ($39.3 \pm 10.1$ pixels, $p < 0.001$), indicating that activations were situated significantly more medially, toward the midline of the brain slice, for cognitively impaired patients. No significant difference was observed along the height axis ($p = 0.389$), suggesting that the medial-lateral distinction specifically, rather than a general shift in activation location, is the

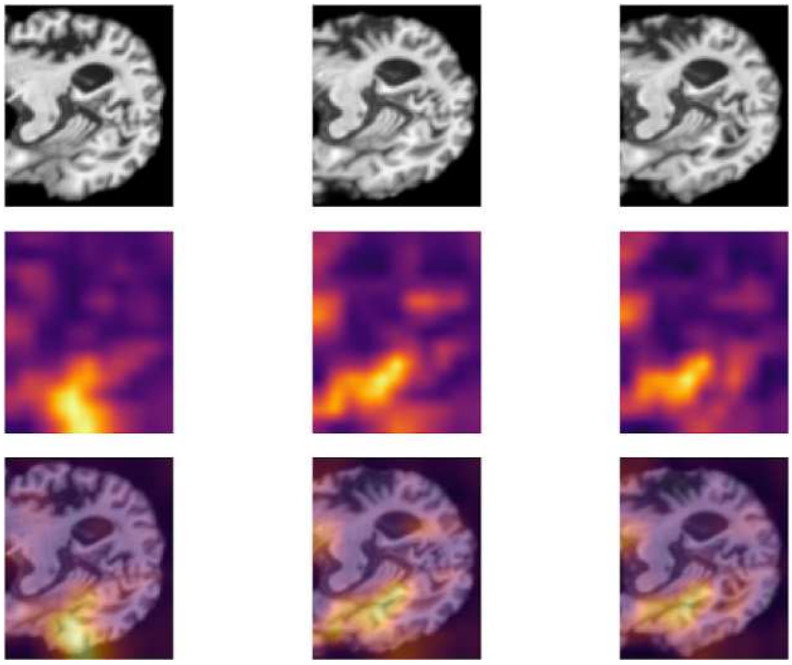


**Fig. 2** Grad-CAM visualization for a true positive cognitively impaired case across three longitudinal visits. *Top row:* input MRI slices. *Middle row:* Grad-CAM activation maps. *Bottom row:* overlay on original slice.

**Table 7** Positioning of the proposed method against representative ADNI-based studies for AD diagnosis. (Note: The reported results from previous work are shown only for reference, as differences in preprocessing, split design, and task formulation limit direct comparison).

| Work | Task | Input Data | Longitudinal | Accuracy | AUC |
|---|---|---|---|---|---|
| ***Contemporary AD classification: AD vs. CN*** | | | | | |
| Zhang et al. (2023) [8] | AD vs. CN | MRI + FDG-PET + CSF | No | 91.07 | 94.07 |
| Tang et al. (2026) [9] | AD vs. CN | MRI + FDG-PET | No | 92.08 | 97.74 |
| Jomeiri et al. (2024) [12] | AD vs. CN | MRI | Yes | 95.28 | 96.49 |
| ***Fine-grained disease classification: CN vs. MCI vs. AD*** | | | | | |
| Li et al. (2026) [13] | CN vs. MCI vs. AD | MRI + Clinical | Yes | 78.83 | 93.84 |
| Soelistiono (2026) [15] | CN vs. MCI vs. AD | MRI (2.5D) | Yes | 47.50 | 52.00 |
| ***Cognitive impairment detection: CN vs. CI*** | | | | | |
| **Proposed work** | **CN vs. CI** | **MRI (2.5D) + Clinical** | **Yes** | **81.5** | **87** |

primary spatially discriminative signal captured by the model. This medial shift is consistent with the known preferential atrophy of medial temporal lobe structures, such as the hippocampus, in Alzheimer's disease, providing further evidence that the model's learned representations align with established disease pathology rather than an arbitrary decision boundary.

## 4.7 Benchmarking against State-of-the-art Work

Table 7 compares the proposed model's performance alongside prior multimodal and longitudinal AD diagnosis works. Direct comparison across these studies is complicated due to differences in task formulation, cohort, and modality composition. Non-longitudinal studies presented in the Table 7 report considerably higher accuracy on the simpler AD vs. CN task compared to the proposed model's performance on the more challenging task of distinguishing cognitively normal and impaired patients. These results are not directly comparable to the present work, both because AD vs. CN is a substantially easier discrimination than CN vs. CI, and because none of these studies merge MCI into the positive class as done here. The most architecturally related prior work is that of Li et al. [13] and Soelistiono [15], both of which adopt longitudinal modeling of structural MRI, though evaluated on a three-way CN vs. MCI vs. AD classification rather than the binary CN vs. CI task used here. While neither result is directly comparable to the binary task addressed here, the proposed model's performance suggests that combining a lightweight 2.5D MRI representation with longitudinal clinical data via a reliability-aware gated fusion mechanism can be competitive with, and in some respects more parameter-efficient than, considerably more complex longitudinal multimodal architectures. Given the substantial differences in task formulation, cohort size, and modality composition across studies, these comparisons are intended to contextualize rather than directly benchmark the proposed model's performance.

# 5 Conclusion

In this article, a longitudinal, multimodal framework for classification of cognitively normal individuals from impaired patients was presented. A 2.5D MRI input representation was integrated with a CNN backbone, alongside spatial and slice attention for efficient feature extraction. A BiLSTM was incorporated for aggregation of repeated, correlated diagnostic signal across visits into a single, reliable representation of the patient's diagnostic condition. A residual gated fusion module was further utilized for adaptive combination of the imaging and clinical representations. When evaluated on 204 ADNI patients under 5-fold cross-validation, the model achieved AUC of $0.870 \pm 0.035$ and balanced accuracy of $0.802 \pm 0.040$, demonstrating that reliability-aware fusion of longitudinal imaging and clinical data is feasible without the computational burden of 3D architectures.

However, several limitations remain to be addressed. Our study relied on a relatively small cohort, which may increase sensitivity to specific cohort patterns. We additionally assume three uniformly spaced longitudinal visits, which does not reflect the irregular follow-up intervals common in real-world clinical settings. Most importantly, our evaluation was restricted to the ADNI dataset. Future work will focus on evaluating the framework on independent cohorts such as AIBL or OASIS to assess generalizability across acquisition protocols and demographics. We also plan to explore temporal architectures capable of handling irregular visit intervals and variable sequence lengths.

# Declarations

**Ethics approval and consent to participate.** This study was an analysis of de-identified data obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database under the ADNI Data Use Agreement. The ADNI study was conducted in accordance with the Declaration of Helsinki. ADNI protocols were approved by the institutional review boards of all participating sites, and written informed consent was obtained from all participants or their authorized representatives at the time of enrollment. Because this study involved secondary analysis of de-identified data available to approved researchers, no new data were collected from human participants by the authors, and no separate institutional ethics approval was required.

**Consent for publication.** Not applicable.

**Availability of data and materials.** The datasets analyzed during the current study are available from the ADNI repository [21], http://adni.loni.usc.edu. Access is restricted and requires an application and agreement to the ADNI Data Use Agreement, so the data are not publicly distributable by the authors.

**Competing interests.** The authors declare that they have no competing interests.

**Funding.** This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

**Authors' contributions.** SHM conceptualized the study, developed the methodology and software, performed the experiments and analysis, and wrote the original draft. SKS supervised the study and the methodology and reviewed and edited the manuscript. SRPV and MSB provided clinical and radiological expertise, validated the interpretability analysis, and reviewed and edited the manuscript. All authors read and approved the final manuscript.

**Acknowledgements.** Data collection and sharing for the Alzheimer's Disease Neuroimaging Initiative (ADNI) is funded by the National Institute on Aging (National Institutes of Health Grant U19AG024904). The grantee organization is the Northern California Institute for Research and Education. In the past, ADNI has also received funding from the National Institute of Biomedical Imaging and Bioengineering, the Canadian Institutes of Health Research, and private sector contributions through the Foundation for the National Institutes of Health (FNIH) including generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics.

# References

[1] World Health Organization: Dementia. https://www.who.int/news-room/fact-sheets/detail/dementia. Accessed 18 July 2026

[2] Blennow, K., de Leon, M.J., Zetterberg, H.: Alzheimer's disease. The Lancet **368**(9533), 387–403 (2006) https://doi.org/10.1016/S0140-6736(06)69113-7

[3] Wang, R.: Global burden of alzheimer's disease and other dementias during 1990–2021: A global burden of disease and risk factors study 2021-based study. Neuroepidemiology **60**(1), 16–27 (2025) https://doi.org/10.1159/000543578 https://karger.com/ned/article-pdf/60/1/16/4326176/000543578.pdf

[4] Disouky, A., Sanborn, M.A., Sabitha, K.R., Mostafa, M.M., Ayala, I.A., Bennett, D.A., Lu, Y., Zhou, Y., Keene, C.D., Weintraub, S., Gefen, T., Mesulam, M.-M., Geula, C., Maienschein-Cline, M., Rehman, J., Lazarov, O.: Human hippocampal neurogenesis in adulthood, ageing and alzheimer's disease. Nature **652**(8112), 1264–1273 (2026) https://doi.org/10.1038/s41586-026-10169-4

[5] Monteiro, A.R., Barbosa, D.J., Remião, F., Silva, R.: Alzheimer's disease: Insights and new prospects in disease pathophysiology, biomarkers and disease-modifying

drugs. Biochemical Pharmacology **211**, 115522 (2023) https://doi.org/10.1016/j.bcp.2023.115522

[6] Grueso, S., Viejo-Sobera, R.: Machine learning methods for predicting progression from mild cognitive impairment to alzheimer's disease dementia: a systematic review. Alzheimer's Research & Therapy **13** (2021) https://doi.org/10.1186/s13195-021-00900-w

[7] Venugopalan, J., Tong, L., Hassanzadeh, H.R., Wang, M.: Multimodal deep learning models for early detection of alzheimer's disease stage. Scientific Reports **11**, 3254 (2021) https://doi.org/10.1038/s41598-020-74399-w

[8] Zhang, J., He, X., Liu, Y., Cai, Q., Chen, H., Qing, L.: Multi-modal cross-attention network for alzheimer's disease diagnosis with multi-modality data. Computers in Biology and Medicine **162**, 107050 (2023) https://doi.org/10.1016/j.compbiomed.2023.107050

[9] Tang, C., Shi, L., Xu, S.: Integrating structural and metabolic neuroimaging biomarkers for alzheimer's disease diagnosis and cognitive score estimation via cross-modal gated learning. Biology **15**(13) (2026) https://doi.org/10.3390/biology15131091

[10] Wang, S., Zhang, H., Kong, W.: Multimodal classification of alzheimer's disease using longitudinal data analysis and hypergraph regularized multi-task feature selection. Bioengineering **12**, 388 (2025) https://doi.org/10.3390/bioengineering12040388

[11] Muksimova, S., Umirzakova, S., Baltayev, J., Cho, Y.: Multi-modal fusion and longitudinal analysis for alzheimer's disease classification using deep learning. Diagnostics **15**, 717 (2025) https://doi.org/10.3390/diagnostics15060717

[12] Jomeiri, A., Navin, A.H., Shamsi, M.: Longitudinal mri analysis using a hybrid densenet-bilstm method for alzheimer's disease prediction. Behavioural Brain Research **463**, 114900 (2024) https://doi.org/10.1016/j.bbr.2024.114900

[13] Li, Z., Gao, L., Dong, L., Huang, H.: Longitudinal multimodal modeling for alzheimer's disease with pre-trained brain latent diffusion and mixture-of-experts fusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, pp. 4343–4350 (2026)

[14] Li, M., Zhou, C., Cao, S.: 2d, 2.5d, or 3d? comparing dimensional approaches in deep neural networks for 3d medical image analysis. Journal of Imaging Informatics in Medicine (2026) https://doi.org/10.1007/s10278-025-01827-6

[15] Soelistiono, S.: A longitudinal and explainable 2.5d deep learning framework for alzheimer's disease progression using adni mri. NeuroImage: Reports **6**(2), 100342 (2026) https://doi.org/10.1016/j.ynirp.2026.100342

[16] Murugan, S., Chandran, V., M G, S., Gao, X.-Z., Elakkiya, B., Muthuramalingam, A., Subramanian, D.M.: Demnet: A deep learning model for early diagnosis of alzheimer diseases and dementia from mr images. IEEE Access **9**, 90319–90329 (2021) https://doi.org/10.1109/ACCESS.2021.3090474

[17] Raza, N., Naseer, A., Tamoor, M., Zafar, K.: Alzheimer disease classification through transfer learning approach. Diagnostics **13** (2023) https://doi.org/10.3390/diagnostics13040801

[18] Yu, Z., Mulholland, A., Huang, T., Liu, Q.: Multimodal ai for alzheimer's disease diagnosis: A systematic review of datasets, models and modalities (preprint). Journal of Medical Internet Research **28** (2025) https://doi.org/10.2196/85414

[19] Shivahare, B., Rajadurai, H., Kanmani, D., Kansal, S., Kumar M, S., S.K.B, S.: An attention-augmented multimodal classification of alzheimer's disease and parkinson's disease vs healthy controls using mri, eeg, and snp data. Scientific Reports **16** (2026) https://doi.org/10.1038/s41598-025-32274-6

[20] Lu, P., Hu, L., Mitelpunkt, A., Bhatnagar, S., Ju, L., Liang, H.-Y.: A hierarchical attention-based multimodal fusion framework for predicting the progression of alzheimer's disease. Biomedical Signal Processing and Control **88**, 105669 (2024) https://doi.org/10.1016/j.bspc.2023.105669

[21] Alzheimer's Disease Neuroimaging Initiative: Alzheimer's Disease Neuroimaging Initiative (ADNI). http://adni.loni.usc.edu. Accessed 14 February 2026

[22] Li, X., Morgan, P.S., Ashburner, J., Smith, J., Rorden, C.: The first step for neuroimaging data analysis: Dicom to nifti conversion. Journal of Neuroscience Methods **264**, 47–56 (2016) https://doi.org/10.1016/j.jneumeth.2016.03.001

[23] Smith, S.M.: Fast robust automated brain extraction. Human Brain Mapping **17**(3), 143–155 (2002) https://doi.org/10.1002/hbm.10062

[24] Jenkinson, M., Smith, S.: A global optimisation method for robust affine registration of brain images. Medical Image Analysis **5**(2), 143–156 (2001) https://doi.org/10.1016/S1361-8415(01)00036-6

[25] Zhang, Y., Brady, M., Smith, S.: Segmentation of brain mr images through a hidden markov random field model and the expectation-maximization algorithm. IEEE Transactions on Medical Imaging **20**(1), 45–57 (2001) https://doi.org/10.1109/42.906424

[26] Jack, C.R., Knopman, D.S., Jagust, W.J., Shaw, L.M., Aisen, P.S., Weiner, M.W., Petersen, R.C., Trojanowski, J.Q.: Hypothetical model of dynamic biomarkers of the alzheimer's pathological cascade. The Lancet Neurology **9**(1), 119–128 (2010) https://doi.org/10.1016/S1474-4422(09)70299-6