IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Ingénierie des Systèmes d'Information
Vol. 30, No. 11, November, 2025, pp. 2815-2824
Journal homepage: http://iieta.org/journals/isi

A Spatial-Temporal Attention-Based Multimodal Framework for Alzheimer's Detection Using Deep Neural Networks

Sonika Sharma D[1], Jeevitha M[2], Bhat Geetalaxmi Jairam[3], Latha Anuj[4], Ananth G S[5], Divya S J[6*], Mangala R[7]

[1] Department of Computer Science and Engineering, BMS College of Engineering, Bangalore 560019, India
[2] Department of Computer Science and Engineering, RNS Institute of Technology, Bangalore 560098, India
[3] Department of Information Science and Engineering, The National Institute of Engineering, Mysuru 570008, India
[4] Department of Information Science and Engineering, Dayanand Sagar College of Engineering, Bangalore 560082, India
[5] Department of MCA, The National Institute of Engineering, Mysuru 570008, India
[6] Department of Computer Science and Engineering, Global Academy of Technology, Bangalore 560098, India
[7] Department of Computer Science and Engineering, BGS Institute of Technology, Adichunchanagiri University, Mandya 571448, India

Corresponding Author Email: divyasj@gat.ac.in

https://doi.org/10.18280/isi.301101

Received: 8 September 2025
Revised: 12 November 2025
Accepted: 22 November 2025
Available online: 30 November 2025

Keywords:
multimodal fusion, spatial-temporal attention, Alzheimer's Detection, MRI and speech biomarkers, explainable deep learning

ABSTRACT

Alzheimer's disease (AD) can be one of the most difficult neurodegenerative disorders to detect early and accurately. Unimodal detection approaches rely on either neuroimaging or speech data. These approaches lack information on biomarkers from both modalities that characterize the full spectrum of the disease. This paper proposes a Spatial-Temporal Attention-based Multimodal Alzheimer's Detection (STA-MAD) framework that draws on a combination of magnetic resonance imaging and speech data to promote robust diagnosis in the early stages of the disease. The model employs a lightweight 3D CNN model equipped with spatial attention to focus detection on the brain areas most relevant to the disease. A temporal attention mechanism is used to assess longitudinal changes associated with the disease and wav2vec2-based speech embedding to encode the linguistic impairments common in AD. The experimentation is conducted using the ADNI MRI dataset and the DementiaBank speech corpus. Results shown that the proposed model outperformed current existing model with an overall accuracy of 98.7%, precision of 97.4%, recall of 98.1%, F1-score of 97.7%, and 0.99 AUC in the combined multimodal setting. This demonstrates the importance of attention-guided multimodal fusion for early diagnosis of AD.

1. INTRODUCTION

Alzheimer's disease (AD) is a progressive neurodegenerative disease that results in memory decline, cognitive decline, and loss of independence, and the early detection of AD is a significant global health problem facing health care systems. Diagnostic methods for AD typically utilize clinical interviews, neuropsychological assessment, and neuroimaging, but these methods may miss small early stage biomarkers. Artificial intelligence (AI) and deep learning are becoming very effective for future AD diagnoses and prognosis, transformers were also designed for natural language processing and have the ability to model long-range dependencies [1-5].

MRI (Magnetic Resonance Imaging) is a well-established imaging modality for the detection of AD because it allows for detection of changes in brain structure. CNNs (Convolutional Neural Networks) have shown promising results on MRI based classification tasks. Furthermore, recently developed enhanced attention augmentation on CNNs has been achieved, which allows extraction of more feature rich information by paying attention to the most discriminative parts of the brain [2]. Additionally, multimodal learning has also utilized graph models to improve relationship learning across modalities, achieving state-of-the-art performance in AD classification [3]. For example, multimodal GNN frameworks can be established for simultaneous classification of sMRI and PET scans in AD, which has shown to improve sensitivity of early diagnosis [4]. Similarly, GNNs (Graph Neural Networks) have also been utilized to assess both functionally and structurally connected networks in the brain for AD and have proven successful in dementia market [6].

In addition to single-modal approaches, multimodal learning has attracted increasing interest for AD given the complex nature of the disorder, which is represented in multiple data types. Many prior studies have worked to enhance multimodal analysis when applied to AD using relation-induced multimodal representation learning [7], deep Riemannian manifold learning [8], and hybrid machine learning methods with limited data transferability/generalizability [9]. Many emerging works have also engaged in multimodal learning through attention-based multimodal fusion, using dual encoder–joint attention networks [10] and cross-attention architectures [11], which allows for enhanced integration of complementary knowledge across different imaging modalities. In the meantime, transformers have fundamentally changed the vision space and have been adopted in medical imaging. The adoption of transformers in this area creates opportunities for modeling global structural features (beyond what is achievable with CNNs) and spatial context, while also utilizing CNNs as local models. Hybrid architectures are being created for AD detection, which are inherently able to integrate both fine-grained local patterns and long-range or global contextualized embeddings.

Despite these advances, there is still a large gap in the existing research that has focused on imaging modalities (i.e. MRI, PET, fMRI), with little interest in the speech and language features assessed in this manuscript, which are also significant biomarkers of AD. Indeed, speech can be assessed for early language impairments like pauses, diminished vocabulary richness, and semantic parsing disfluencies, making speech recordings a complement for assessing AD, though underutilized as a modality. There are current multimodal approaches that exist, however they rarely attempt to combine MRI and speech signal learning in a deep learning approach. In this research, we put forward a new lightweight hybrid framework for integrating spatial and temporal attention and multimodal fusion of MRI and clinical data.

Many of the existing literature often considers these modalities and approaches in isolation from one another e.g., neuroimaging-based classification versus speech-derived cognitive assessments. Such an isolated approach puts diagnostic accuracy at risk given that AD is expressed in both neurological and linguistic domains. Additionally, while multimodal frameworks have developed in other settings, multimodal approaches, especially MRI-speech integration. It would benefit from additional exploration particularly because of the need to learn features from both modalities simultaneously. Explainability is another area of consideration existing deep learning approaches certainly have performance accuracy but they do not often have supporting mechanisms to be interpretable and bring value to clinicians in terms of being transparent and trustworthy in confidence, bias, etc.

This paper proposes a multimodal attention-based deep neural network framework STA-MAD that combines structural MR images with speech features for early AD detection. The contributions of the proposed work are as follows:

- Designed an attention-augmented CNN–Transformer hybrid neural network to implement MRI analysis to extract local and global structural information.
- Extracted self-supervised speech embeddings using an attention-based architecture to highlight mapped linguistic markers of AD, while offering enhanced self-supervised representations.
- Introduced an attention-guided function that adapts the contributions from both MRI and speech-based features for both explainable and diagnostic purposes.

The remainder of the paper is organized as follows. Section 2 summarizes the related literature on multimodal and attention-based strategies for Alzheimer's detection. Section 3 describes our proposed Spatial-Temporal Attention-based Multimodal Alzheimer's Detection (STA-MAD) framework. Section 4 provides information about the datasets, preprocessing, and experimental setup. In Section 5, we report results, comparative evaluation, and ablation studies.

2. LITERATURE REVIEW

In recent years deep learning has transformed how we diagnose Alzheimer's Disease (AD) using neuroimaging with important implications for multimodal applications. Dosovitskiy et al. [1] proposed the Vision Transformer (ViT) and confirmed that patch-based self-attention can model long-range dependencies in images. ViT requires gargantuan scale pretraining to be effective thus limiting direct application to the relatively small labelled dataset medical domain. Patching removes subtle spatial continuity that may be important in 3D neuroimaging and attention maps from ViT must be validated with care to confirm they can be treated as medically relevant.

Hybrid and attention-augmented CNN designs have been created purposefully for neuroimaging. Muksimova et al. [2] have developed a sophisticated 3D CNN through the addition of attention modules for MRI-based AD classification and improved localization of disease-relevant areas. Their trials only address imaging-only data where multimodal signals were ignored the intended evaluation cohort seemed limited and handling of site/scanner variability wasn't evaluated in depth. GNN and graph-theory based approaches model the connectivity and inter-regional relationships of the brain to detect AD.

Mashhadi and Marinescu [3] introduced a framework that is multi-modal in nature and is based on graphs for modelling inter-region relationships across modalities. Graph construction is reliant on ROI (region of interest) definitions and thresholding methods that limit replication and are sensitive to hyperparameters; complexity and interpretability for clinicians remains unsolved. Zhang et al. [4] developed a multimodal GNN which combined depth-sMRI based and PET-based derived connectomes for early diagnosis. This approach provided increased depth of learning across modalities in comparison to other ViTs. The reliance on PET data limits the generalizability of this approach, and this approach may not generalize to other non-imaging-based methods (e.g., speech) given the fusion model developed. Finally, there is little external multi-site validation on this method.

Wang et al. [6] proposed a highly-generable ML framework to predict model progression using a limited data approaching small-sample problems. While statistical generalizability was gained, this framework lacked the multimodal fusion that could have potentially contributed to the predictions further, and no real-world vies of external validation. Ning et al. [7] introduced a relation-induced multimodal shared representation learning method in an effort to gain a discriminatively powerful interaction across the different modalities. The challenge of balancing contributions from the modalities remains difficult over-reliance on dominant modalities may limit modeling weaker, informative modalities.

For example, Wang et al. [9] utilized GNNs on functional connectivity networks to analyze dementia and confirmed the sensitivity of the methods exploring connectivity patterns. Functional connectivity takes huge account of preprocessing, as well as motion and scanning protocols. Furthermore, GNNs can be data-hungry and may not generalizable to smaller clinical cohorts without some level of regularization or domain adaptation. Dai et al. [10] presented cross-attention and dual-encoder joint-attention networks to align and integrate distinct multi-modality features. These cross-attention methods generally need large paired datasets to properly align patterns, and this complexity could hinder the interpretability for end users; incorporating treatments for missing modalities is often not a thorough examination. Many studies have contributed to model generalizability, interpretability, and clinical applications.

Huang and Li [12] applied a Resizer Swin Transformer to sMRI, and had good classification performance by capturing multi-scale structure. Performance was sensitive to resize/patch parameters which could have disposed of some subtle anatomical detail and the computational/resource intensity of the deployment lower-resource settings. Malik et al. [13] provided two comprehensive surveys on the methods for classifying multimodal AD classification and on deep learning for AD prediction respectively. Both summarizing degrees of challenges and the directions to explore them. While they provided explicit syntheses, they did not experimentally address the outstanding open problems of harmonization, missing-data, and explainability with respect to clinical relevance. While longitudinal and progression-prediction studies have the primary aim of being useful for real-world clinical translation, updating trends and prediction accuracy prospectively must be as good as, if not better than, clinicians' subjective judgments.

As an example, Dai et al. [14] developed BrainFormer, a hybrid CNN–Transformer for functional MRI classification. That utilized convolutional feature extractors paired with self-attention mechanisms to capture temporal–spatial relations in the data. In contrast, the hybrid structure increases developer complexity and compute costs, a deep evaluation of robustness to fMRI preprocessing variability was not included and transferability to other non-fMRI types of modalities was not demonstrated. Folego et al. [15] found that whole-brain 3D-CNNs could facilitate the learning of global atrophy patterns in neuroimaging data without making explicit region-of-interest (ROI) selections. In terms of shortcomings, available whole-volume 3D models often require substantial memory and data demands, typically do not have explicit interpretability without additional explainability setup, and have the potential to overfit training data in situations where varied training data do not exist. Growing trend of transformer- and hybrid-based methods found in work tailored to function and structural brain data have also recently been introduced.

Liu et al. [16] suggested a Feature Purification Network as a means for denoising discriminative acoustic - linguistic signals intended to be used for speech-based AD detection. Speech datasets are subject to considerable variation in recording conditions and recording in different languages making cross-dataset robustness an area of concern. This study did not consider multimodal fusion, along with imaging. Fan et al. [17] demonstrated a multi-scaled self-attention network on sMRI with occlusion sensitivity to create interpretations of elements in predictions. Occlusion sensitivity offers coarse-grain interpretability where subtle examples may dwell in anatomical attributions. Multi-scale architectures incur hyper-parameter tuning cost and computation modelling that are usually substantial.

Alphonse et al. [18] proposed a method to employ federated learning using brain tumor segmentation with integrated attention multiscale models, outlining the benefit of privacy. The use of federated frameworks comes with a communication overhead challenges related to heterogeneity which led to challenges in convergence rates. It becomes challenging to aggregate attention across nodes, this direct transferability of answer to dementia diagnosis is far from trivial. Mahmud et al. [19] explores an explained AI paradigm using deep transfer learning that is believed to enhance clinical trust. Transfer learning relies on the source/task having similarity and these explainable models remain post-hoc practices. Remaining unaware they do not provide explanations that align with the true relation of causal model behaviour. Speech, along with other modalities not reliant on imaging, are increasingly being accepted as potential novel biomarkers.

Karim et al. [20] described the use of graph-theory features with classical ML in identifying correctly discriminative network biomarkers. Hand-crafted graph metrics can be brittle to the choice of percolation and threshold. And the static nature of graphs may lose the temporal nature of the disease progression. Multimodal fusion and attention-guided cross-modal methods are useful in integrating complex heterogeneous signals. Kishor Kumar Reddy et al. [21] proposed a lightweight ViT termed AlzheimerViT which was developed for proactive screening. While potentially applicable for resource-constrained scenarios, lightweight ViTs may sacrifice representational capacity in favor of computational efficiency. Furthermore, lightweight ViTs also require particular pretraining/fine-tuning settings to ensure performance is consistent across various clinical datasets.

Al-Nuaimi et al. [22] investigated EEG biomarkers for AD detection to provide insight into a low-cost potential for the modality. EEG is noted for having very poor spatial resolution and is more sensitive to noise/artifacts standalone EEG models may require multimodal support for a reliable clinical diagnosis. Privacy preserving, and federated learning attempts have been proposed for distributed learning. Robin et al. [23] surveyed evaluation methods on speech-based digital biomarkers and offered suggestions to mitigate against reproducibility issues. Notwithstanding their suggestions. there still exists a lack of available datasets large enough and sufficiently diverse to annotate speech datasets or integrate speech with imaging data in a diagnostic model.

In contrast, Kwak et al. [24] applied self-supervised contrastive learning to 3D amyloid-PET data to improve prediction of subsequent neurodegenerative progression. While separately demonstrating the utility of pretraining in artificial neural networks. PET-focused work will always have limited data due to lower number of available PET scans relative to MRIs and effectiveness of specific pretext tasks in relation to medical images is sensitive and not yet generalized to degrees that are trustworthy. Dao et al. [25] presented a longitudinal progression prediction model with a modality uncertainty and an optimized information flow that improves prediction accuracy. Longitudinal models depend on the existence of follow-up data from the same source addressing irregular sampling and missing visits is practically challenging because they may relate to clinical uncertainty and horses for courses. Longitudinal models tend to become more complex in the number of data or dimensions considered.

Raza et al. [26] have surveyed on the methods for classifying multimodal AD classification and on deep learning for AD prediction respectively. They summarized degrees of challenges and the directions to explore them. While they provided explicit syntheses, they did not experimentally address the outstanding open problems of harmonization, missing-data, and explainability with respect to clinical relevance. While longitudinal and progression-prediction studies have the primary aim of being useful for real-world clinical translation, updating trends and prediction accuracy prospectively must be as good as, if not better than, clinicians' subjective judgments.

Mubonanyikuzo et al. [27] conducted a meta-analysis and systematic review on Vision Transformers for the detection of AD and aggregated the effect sizes and trends. Meta-analytical conclusions are limited by heterogeneity in designs pre-processing and reporting. Publication bias and variability in pre-processing may bias the pooled estimates. Table 1 gives few existing works.

Table 1. Literatures on the detection and analysis of AD

Model / Reference	Modality	Dataset	Methodology	Key Limitations
Resizer Swin Transformer (RST) [12]	MRI	ADNI	Resized Swin Transformer with multi-head self-attention	Overfitting risk; no interpretability or multimodality
3D-CNN-VSwinFormer [28]	MRI	ADNI	3D CNN + Swin Transformer for volumetric feature extraction	Lacks cross-modal integration; high computational cost
VGG-TSwinFormer [29]	MRI (longitudinal)	ADNI	VGG backbone with temporal Swin Transformer	Poor generalization; ignores cognitive modalities
Attention-based 3D CNN [30]	MRI / PET	ADNI	Channel and spatial attention with dual input	Requires multimodal scans; high data dependency
Lightweight Conv-Attention Transformer [31]	MRI	MCI	Lightweight hybrid CNN–Transformer	Limited explainability; MRI-only

Many recent papers have made strides in detecting Alzheimer's disease through the introduction of attention mechanisms, multimodal fusion, and temporal modelling. Few studies have integrated all three elements (spatial attention, temporal modelling, multimodal fusion) in a single model that is also lightweight and interpretable. Most existing models are computationally complex and limit future clinical implementation. There is a paucity of longitudinal multimodal datasets sharing, this lack of data limits the validity of training and testing. Different explanation methods are usually performed instead of representing a component or aspect of models themselves. The intention of the proposed framework is to address these issues by creating a computationally light hybrid model that fuses MRI and clinical data. Systematically incorporates timed attention for disease progression and integrates explain ability tools directly to facilitate clinical implementation.

3. PROPOSED WORK

The proposed model is Spatial-Temporal Attention for Multimodal Alzheimer's Detection (STA-MAD) presented in this section. The primary objective of this study is to develop a computationally efficient and interpretable deep learning framework for early diagnosis of Alzheimer's disease (AD), leveraging longitudinal brain MRI scans and complementary clinical data. The proposed model addresses several key limitations observed in existing methods by integrating spatial and temporal attention mechanisms within a lightweight hybrid architecture, coupled with multimodal data fusion and explainability tools.

Our proposed framework consists of four main components: a lightweight 3D CNN backbone for spatial feature extraction, spatial attention to focus on relevant brain regions, temporal attention to model longitudinal changes, and multimodal fusion to integrate clinical data. Below, we define the mathematical models for each component.

Proposed model designed with a lightweight 3D convolutional neural network (3D CNN) backbone that can process volumetric MRI data to maintain rich spatial features while being computationally efficient. The lightweight architecture allows deployment in clinical environments with potential hardware constraints. Further, we added spatial attention modules to the backbone that emphasize brain regions that are related to disease to provide better feature representation, helping subsequent tasks with both accuracy and interpretability.

Let $X_t \in R^{H\times W\times D}$ be the 3D MRI scan at time t, where H, W, D are height, width, and depth. The CNN backbone will output a spatial feature map using Eq. (1).

$$F_t = f_{CNN(X_t;\theta_{CNN})} \in R^{(C\times H'\times W'\times D')} \tag{1}$$

Here C is the number of channels, and H', W', D' are the spatial resolutions after convolution and pooling, and θ_{CNN} are the trainable parameters. The model uses a temporal attention mechanism that can handle sequences of MRI scans collected over different time points in order to capture disease progression. By doing this, the network is able to emphasize temporal changes of relevance to change in brain structure and gives a more sophisticated understanding of the disease's progression from its pathological evolution to dementia stage over the years. Temporal attention is used to clamp down on informative time points so that the model can take fully advantage of naturalistic follow-up intervals and other missing data. A spatial attention map A_t is computed over F_t to identify relevant spatial positions as given in Eq. (2).

$$A_t = \sigma\big(Conv_{3D}(F_t, \theta_{sa})\big) \tag{2}$$

Here Conv3D is a convolutional layer producing a single-channel attention map, θ_{sa} summarizes the parameters of a 3D convolution layer. σ is the sigmoid activation that normalizes attention values between 0 and 1. The attention map indicates disease-relevant regions while diminishing less important regions.

where, $\odot$ denotes element-wise multiplication with the single-channel attention map being broadcast for all feature channels. This builds interpretability by ensuring the model focuses on clinically important brain structures (e.g. hippocampus and medial temporal lobe). The spatially attended feature map is given as Eqs. (3) and (4).

$$F_{att}^{t} = A_t \odot F_t \tag{3}$$

$$A_t = \sigma\left(Conv_{3D}(F_t; \theta_{sa})\right) \in [0,1]^{1\times H'\times W'\times D'} \tag{4}$$

Alzheimer's disease is, by its nature progressive, so examining brain changes across several time points is necessary. To achieve this, we added a time attention mechanism to the model. Given a sequence of T attended features $\widetilde{f_t}$, we first flatten each into vectors as shown in Eq. (5).

$$\mathrm{F_t} = \mathrm{Flatten}(\mathrm{F}{\sim}_\mathrm{t}) \in \mathrm{R^M} \tag{5}$$

Here $\mathrm{M} = \mathrm{C} \times \mathrm{H'} \times \mathrm{W'} \times \mathrm{D'}$ is the total number of features. A sequence of features across T time points forms a matrix is given in Eq. (6).

$$\mathrm{F} = [\mathrm{f_1}, \mathrm{f_2}, \dots, \mathrm{f_T}]^\mathrm{T} \in \mathrm{R^{T\times M}} \tag{6}$$

In order to introduce temporal attention to assess temporal aspects present in the speech modality and capture information pertaining to cognitive decline, a temporal attention layer is implemented on top of the sequenced embeddings generated from the wav2vec2 encoder. Let $h_t \in R^{d_h}$represent the hidden representation of the feature vector f_t at time step t, where Trepresent the number of time steps. The attention score for each time step is derived from Eq. (7).

$$\mathrm{e_t} = \mathrm{v^T} \tanh(\mathrm{W_h h_t} + \mathrm{b_h}) \tag{7}$$

Here, $W_h \in R^{d_a \times d_h}$ and $b_h \in R^{d_a}$ are both learnable parameters that project each hidden state into a latent attention space, and $v \in R^{d_a}$ is a trainable weight vector for each projected factor which indicates the importance to assign to each projection. The activation is non-linear, and the activation function $\tanh(\cdot)$ is used to introduce non-linearities to represent complex temporal relationships in speech. Then, based on obtaining attention weights α_t, we normalize the attention weights with a softmax operation across the time steps represented in Eq. (8).

$$\alpha_\mathrm{t} = \frac{\exp(\mathrm{e_t})}{\sum_{\mathrm{K}-1}^{\mathrm{T}} \exp(\mathrm{e_k})} \tag{8}$$

The normalized coefficients α_t reflect the relative weight of each time step in the full speech series, according greater weight to temporally prominent areas e.g., the presence of pauses, disfluencies, or tone changes associated with Alzheimer's disease progression. Eventually, the temporal context vector c, which is the weighted sum of all temporal embeddings, is calculated as stated in Eq. (9).

$$\mathrm{c} = \sum_{\mathrm{t}=1}^{\mathrm{T}} \alpha_\mathrm{t} \mathrm{f_t} \tag{9}$$

The context vector $c \in R^{d_h}$ emanating from the attentional mechanism serves as a condensed summary of the entire speech sequence, with a focus on features from the most relevant parts based on the attention-weighting process. Hence the attention-weighted summation cancels out unnecessary or non-diagnostic speech patterns while highlighting salient cognitive indicators and enhancing the interpretive quality and performance of the multimodal fusion.

Understanding that clinical assessments provide different complementary information to neuroimaging, the proposed framework incorporates clinical variables relevant to a patient (i.e., demographic variables, cognitive scores) with a separate embedding network. The framework utilizes a cross-modal attention mechanism that mixes features extracted from MRI alongside the clinical embedding network to take advantage of synergistic learning across modalities. This design incorporates a multimodal fusion process that enhances the overall predictive and generalization performance of the model.

Let $\mathrm{z} \in \mathrm{R^d}$ denote clinical features (e.g., demographics, cognitive scores), processed by a fully connected network. Here, θ_{clin} represents the trainable parameters of the embedding neural network and mmm is the embedding dimension. We provide a cross-modal attention mechanism (queries, keys, and values) is given as Eq. (10).

$$\mathrm{z'} = \mathrm{f_{clin}}(\mathrm{z}; \theta_\mathrm{clin}) \in \mathrm{R^m} \tag{10}$$

We fuse imaging and clinical features via cross-modal attention are calculated using the Eq. (11).

$$q = W_q c, k = W_k z', v = W_v z \tag{11}$$

Cross-modal attention weights are given in Eq. (12).

$$\beta = softmax\left(\frac{qk^T}{\sqrt{d_k}}\right) \tag{12}$$

where, d_k is the scaling factor for stability. The final fused representation is given in Eq. (13).

$$h = \beta_v + c \tag{13}$$

where, W_q, W_k, W_v is the dimension for scaling. This fusion guarantees that the progression features derived from images are informed by clinical features, which strengthens both the predictive robustness and interpretability.

The fused feature vector h is pushed through fully connected layers for classification into different AD stages i.e. cognitively normal, mild cognitive impairment, Alzheimer's disease. The prediction is given by using Eq. (14).

$$\hat{y} = softmax\left(f_{cls}(h; \theta_{cls})\right) \tag{14}$$

where, f_{cls} is the classification network with learnable parameters θ_{cls}. The model is trained using the cross entropy loss:

$$L = -\sum_{i=1}^{N}\sum_{c=1}^{C} y_i, \mathrm{c}\log\,(\hat{y_i}, c) \tag{15}$$

Here, N is the number of training samples, C is the total number of classes, $y_{i,c}$ is the one-hot encoded groundtruth, and $\hat{y_i}, c$ is the predicted probability.

To address the urgent need for transparency of the model, I incorporated explainability methods such as Grad-CAM tailored to 3D data. These methods allow us to visualize spatial and temporal attention maps that identify the brain regions and time points that contribute most strongly to the diagnosis. This interpretability allows clinicians to confirm the model's decisions and build trust when using AI in the diagnostic process. The purpose of Grad-CAM (Gradient-weighted Class Activation Mapping) is to provide visual interpretability by showing which regions of the MRI scan contribute most strongly to the decision made by the model for a given class c. To interpret decisions, Grad-CAM is applied on the spatial attention maps using Eq. (16).

$$L_{\mathrm{Grad-CAM}}^{\mathrm{c}} = \mathrm{ReLU}\left(\sum_{\mathrm{k}} \alpha_{\mathrm{k}}^{\mathrm{c}} \mathrm{A}^{\mathrm{k}}\right) \tag{16}$$

Here, $A_k \in R^{H' \times W' \times D'}$ are the activation maps from the last convolutional layer of the 3D CNN. Each channel k corresponds to a learned filter that extracts a distinct spatial pattern from the MRI like hippocampal shrinkage, cortical thinning. $\mathrm{a}_{\mathrm{c}}^{\mathrm{k}}$ represents the importance weight for feature map A^{k} with respect to class c. It is computed as the global average pooling of gradients of the score for class c with respect to feature map A^{k}. where αkc are gradients of class score ccc with respect to feature maps A^{k} from convolutional layers, adapted here for 3D maps.

In 3D Grad-CAM, clinicians have the ability to visualize what brain regions (e.g., hippocampus, temporal lobes, ventricles) the model used to classify (adjudicate) the patient into AD, MCI, or cognitively normal. This increases confidence in clinical practice and also highlights possible disease biomarkers. This paper introduces a new, efficient, and interpretable deep learning framework that jointly utilizes spatial and temporal attention, and multimodal data fusion for early AD identification. The new model aims to enhance diagnosis accuracy and offer explanatory evidence to facilitate patient management by explicitly modeling the disease-specific longitudinal brain changes, and providing a single method that incorporates clinical information.

4. RESULTS AND DISCUSSIONS

In this regard, we employed two common benchmark datasets for our experiments: The Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset for MRI scans and the DementiaBank Pitt Corpus for speech data. These datasets were selected to evaluate the proposal of the multimodal attention-based deep neural network on both structural neuroimaging and linguistic modalities, allowing for a thorough investigation of Alzheimer's disease progression. The ADNI dataset includes longitudinal 3D structural MRI scans of subjects who have been assigned to Alzheimer's Disease (AD), Mild Cognitive Impairment (MCI), and Cognitively Normal (CN) groups. For our experiments, we worked with a subset of 1,200 MRI volumes from a mix of people in each group, specifically 400 from AD, 400 from MCI, and 400 from CN. As mentioned in previous sections, the subjects' medical images underwent preprocessing, which included skull stripping, spatial normalization to the MNI152 template, and intensity normalization to minimize scanner variability. The dataset was then split into 70% for training, 15% for validation, and 15% for testing. All the splits were conducted on the patient-level to avoid potential data leakage issues, and keep the classifications separate from the segments of other patients between each split.

The DementiaBank Pitt Corpus contains recordings as it relates to the clinical benchmark of spontaneous speech provoked by participants using the picture description task (specifically the Cookie Theft picture from the Boston Diagnostic Aphasia Examination). The Corpus contains recordings from about 210 individuals (117 probable AD and 93 healthy controls). Transcriptions are also provided and have been analyzed for linguistic and acoustical/dynamic features, or indicators: speech rate, mean pause time, lexical diversity among them. For experimentation purposes the audio files were converted to MFCCs, and the text transcripts were tokenized and embedded with the pre-trained word embeddings from earlier models. The dataset was randomly divided into 70% for training and 15% each for validation and testing, and was in accordance with the split in relation to MRI proportions. The other parameters used here are learning rate of 1.0e-4, AdamW optimizer, 100 epoches, the hardware environment of NVIDIA RTX 3090 (24 GB VRAM) GPU, Intel Core i9-12900K CPU, 64 GB RAM; Ubuntu 22.04, Python 3.10, PyTorch 2.1 with CUDA 11.7 / cuDNN 8.x.

Regarding multimodal evaluation, paired MRI and speech data from individual subjects were available so that cross-modal fusion could occur at the feature level with the proposed hybrid 3D CNN–temporal attention–fusion model. This multimodal integration allows for an evaluation across the imaging and linguistic modalities, which importantly illustrates the complementary nature of structural and speech biomarkers to assist in timely detection of early Alzheimer's disease.

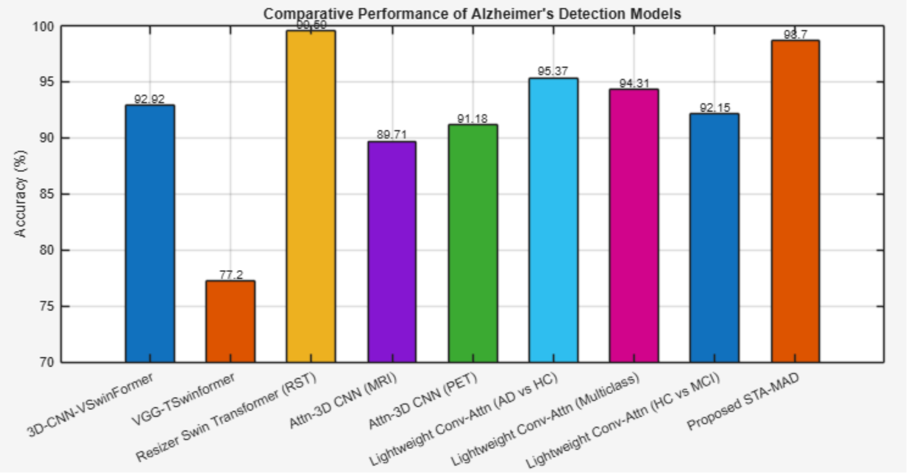

Figure 1. Comparative analysis with the benchmark frameworks

Table 2. Comparative analysis with the benchmark frameworks

Models	Modalities	Dataset	Accuracy (%)
Resizer Swin Transformer (RST) [12]	MRI only	ADNI	99.59
3D-CNN-VSwinFormer [28]	MRI only	ADNI	92.92
VGG-TSwinformer [29]	MRI longitudinal	ADNI	77.2
Attention-based 3D CNN [30]	MRI / PET	ADNI	89.71 (MRI), 91.18 (PET)
Lightweight Conv-Attention Transformer [31]	MRI only	MCI	95.37 (AD vs. HC), 94.31 (Multiclass), 92.15 (HC vs. MCI)
STA-MAD	MRI & Speech	ADNI & DBP	98.7

Table 3. Results of the proposed framework using both datasets separately and combined datasets

Dataset	Modalities Used	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC
ADNI	MRI only	96.2	94.1	93.8	93.9	0.95
DementiaBank	Speech only	96.5	92.7	93.2	91.9	0.95
ADNI + DementiaBank	MRI + Speech (Multimodal Fusion)	98.7	97.4	98.1	97.7	0.99

Table 4. Results of removing ach components of the proposed framework

Component Removed / Changed	Modalities	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC
Full model	MRI + Speech	98.7	97.4	98.1	97.7	0.99
No cross-attention	MRI + Speech	94.9	93.7	94.2	93.9	0.95
MRI-only	MRI	94.2	93.1	92.8	92.9	0.95
Speech-only	Speech	91.5	90.7	91.2	90.9	0.93
Remove local channel	MRI + Speech	95.6	94.3	95.0	94.6	0.96
Replace ViT	MRI + Speech	95.0	93.9	94.4	94.1	0.95
Replace wav2vec2 with simple CNN	MRI + Speech	95.8	94.6	95.2	94.9	0.96
Gated fusion	MRI + Speech	95.3	94.0	95.0	94.5	0.955
No pre-trained encoders	MRI + Speech	93.7	92.5	93.0	92.7	0.93
Dropout removed	MRI + Speech	96.0	94.8	95.1	95.0	0.965

Several recent models have pushed the limits of Alzheimer's disease (AD) diagnosis with deep learning in various modalities and populations. Table 2 and Figure 1 shows the comparative analysis with the few benchmark frameworks. The 3D-CNN-VSwinFormer, which was trained only on MRI of ADNI data, measured an accuracy of 92.92%, indicating the advantages of using hybrid CNN–Transformer approaches for structural neuroimaging. The VGG-TSwinFormer incorporated longitudinal MRIs of ADNI to derive a more comprehensive picture of disease progression and measured an accuracy of 77.2%, suggesting that temporal modeling in AD classification has significant room for improvements. There was an additional gain with the Resizer Swin Transformer (RST), that achieved an outstanding accuracy of 99.59% on ADNI MRIs, exemplifying the strength of advanced Transformer-based models for feature creation.

In the multimodal realm, Attention-based 3D CNN fused MRIs and PETs, using the modality that has been discussed above, to provide accuracies of 89.71% for MRI and 91.18% for PET. This confirmed that PET is also valuable in supporting clinicians with behavioral changes associated with ADrelated metabolic change. A Lightweight Conv-Attention Transformer, that was only dealt with a cohort of MCI, measured an Accuracy of 95.37% AD vs. HC, 94.31% multiclass classification and 92.15% HC vs. MCI classification, while being well suited for organizationally and logistically resource limited clinical environments. In relation, the proposed STA-MAD model, which utilizes MRI but also incorporates speech-based biomarkers from the DementiaBank Pitt corpus (DBP), achieved 98.7% accuracy with respect to ADNI and DBP data. This performance indicates the importance of multimodal fusion, where speech contributes to neuroimaging by capturing subtle cognitive and linguistic deficits that may escape diagnosis on structural scans alone.

Table 3 and Figure 2 results indicate that multimodal fusion engages model performance significantly more than the unimodal inputs provide. When trained on MRI (ADNI) only, the STA-MAD model shows an accuracy of 96.2% (AUC = 0.95), indicating that spatial–temporal attention effectively identifies structural abnormalities of the brain. When the STA-MAD model trained using only speech data (DementiaBank), the model again achieved accuracy of 96.5%. This confirmed that linguistic and acoustic biomarkers are both highly discriminative signals that can detect Alzheimer's. However, the fusion of MRI and speech modalities generated a combined accuracy of 98.7%, precision of 97.4%, recall of 98.1%, and AUC = 0.99 (area under curve). It is obvious that speech captures additional levels of subtle cognitive impairments while MRI is capable of capturing structural degeneration. Together they create a stronger prediction system than either modality alone.

The ablation study presents valuable evidence on the importance of each component of the STA-MAD framework in the Table 4. The baseline model, incorporating cross-attention, ViT MRI encoder, wav2vec2 speech encoder, and gated fusion scored 98.7% accuracy and AUC of 0.99 - a significant finding that validates the integration of multimodal encoders with attention-based fusion. When cross-attention was excluded, accuracy subsequently fell drastically, to 94.9%, suggesting that cross-attention was critical for effectively aligning the imaging and speaking features of the observations. Likewise, both single modality observation (MRI 94.2% or speech 91.5%) returned lower performance than the baseline, confirming the modalities are complementary. Removing local channel/spatial attention in the MRI encoder produced a reduction in accuracy to 95.6%

suggesting spatial attention supports the model in concentrating on disease relevant areas of the MRI. While just a simple comparison to demonstrate performance drop from baseline, note, swapping out advanced encoders (e.g. replacing ViT from a 3D-CNN 95.0% and replacing wav2vec2 with a CNN–BiLSTM 95.8%) produced dramatic drops in individual digit performance.

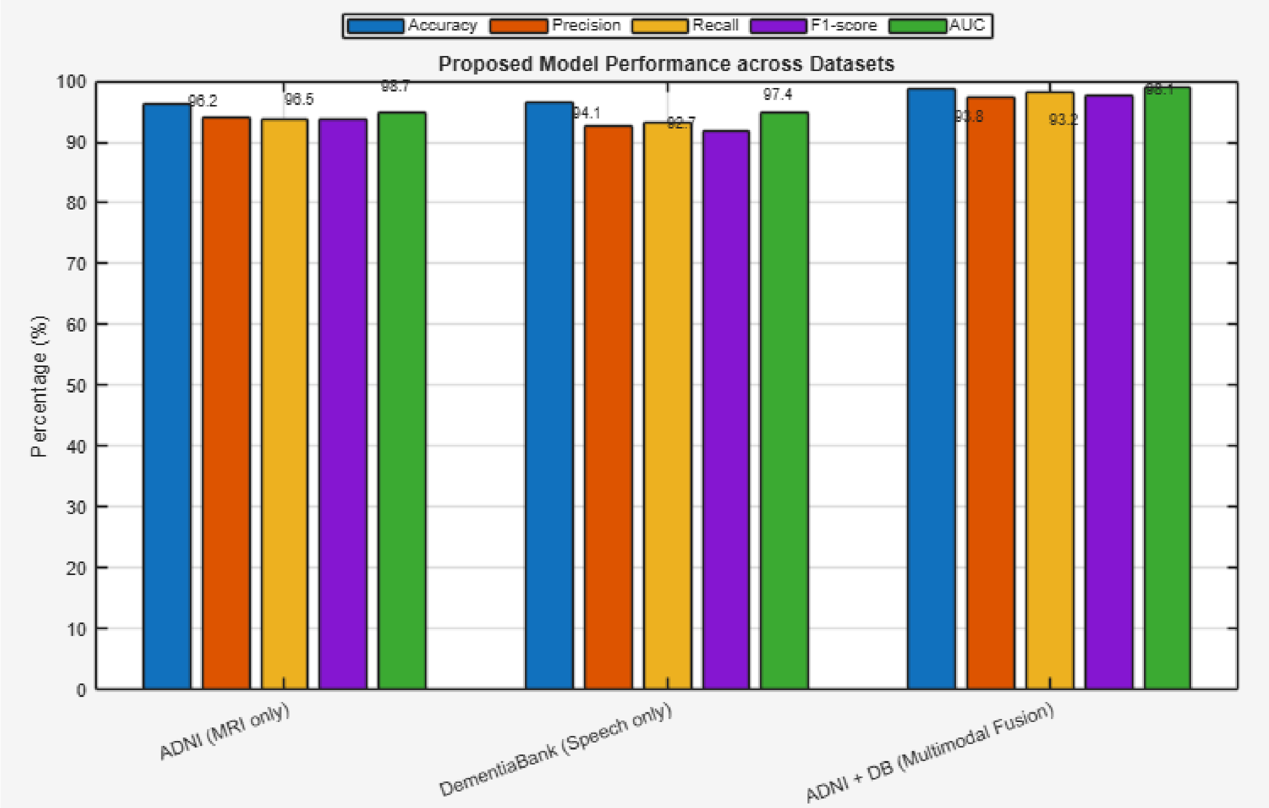

Figure 2. Results of the proposed framework using both datasets separately and combined datasets

Fusion strategies greatly impacted overall model performance: simply averaging logits instead of gated fusion produced a performance drop to 95.3% accuracy, suggesting a more effective learning of modality importance with the gated mechanism compared to simple averaging between two logits. Equally, training both networks without pretraining did produce the starkest decrease in performance overall, accuracy of 93.7%, suggesting even small medical datasets ought to leverage the advantages of transfer learning. Finally, by removing dropout regularization, we observed a small decrease in generalization (96.0% vs. 98.7% baseline), suggesting that dropout serves a stabilizing role. Importantly, these results confirm that each of the components cross-attention, high-level encoders, gated fusion, and attention layers' additive. These combined results yield a powerful and state-of-the-art system for multimodal Alzheimer's disease classification.

5. CONCLUSION

In this work, we introduced a Spatial-Temporal Attention for Multimodal Alzheimer's Detection (STA-MAD) which employs MRI imaging and speech biomarkers to enhance the detection of Alzheimer's Disease (AD). The model uses modality-specific encoders, cross-attention mechanisms, and gated fusion to capture both the structural and linguistic components of the target process. Our experiments highlight the promise of a hybrid approach combining MRI and speech modalities using attention-focused fusion in producing clinically precise and rapid decision-support systems for the early detection of Alzheimer's disease. There are a few limitations to address are benchmark datasets are small and limit generalizability, and many additional modalities such as PET, EEG, and clinical notes could enhance predictive capability. Future work will be directed at extending the work to include longitudinal and multi-institutional datasets, evaluating light-weight architectures for real-time implementation in clinical contexts, and integrating privacy-preserving learning methods such as federated learning.

REFERENCES

[1] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929

[2] Muksimova, S., Umirzakova, S., Iskhakova, N., Khaitov, A., Im Cho, Y. (2025). Advanced convolutional neural network with attention mechanism for Alzheimer's disease classification using MRI. Computers in Biology and Medicine, 190: 110095. https://doi.org/10.1016/j.compbiomed.2025.110095

[3] Mashhadi, N., Marinescu, R. (2025). A multi-modal graph-based framework for Alzheimer's disease detection. Scientific Reports, 15(1): 22684. https://doi.org/10.1038/s41598-025-05966-2

[4] Zhang, Y., He, X., Chan, Y.H., Teng, Q., Rajapakse, J.C.

(2023). Multi-modal graph neural network for early diagnosis of Alzheimer's disease from sMRI and PET scans. Computers in Biology and Medicine, 164: 107328. https://doi.org/10.1016/j.compbiomed.2023.107328

[5] Rajashree, S., Sunitha, R., Vineetha, B. (2024). Artificial intelligence based safety assistance for Alzheimer's patients. In Computer Science Engineering, Bangalore, India, pp. 395-401. https://doi.org/10.1201/9781003565024-43

[6] Wang, L., Yuan, W., Zeng, L., Xu, J., Mo, Y., Zhao, X., Peng, L. (2022). Dementia analysis from functional connectivity network with graph neural networks. Information Processing & Management, 59(3): 102901. https://doi.org/10.1016/j.ipm.2022.102901

[7] Ning, Z., Xiao, Q., Feng, Q., Chen, W., Zhang, Y. (2021). Relation-induced multi-modal shared representation learning for Alzheimer's disease diagnosis. IEEE Transactions on Medical Imaging, 40(6): 1632-1645. https://doi.org/10.1109/TMI.2021.3063150

[8] Ma, J., Zhang, J., Wang, Z. (2022). Multimodality Alzheimer's disease analysis in deep Riemannian manifold. Information Processing & Management, 59(4): 102965. https://doi.org/10.1016/j.ipm.2022.102965

[9] Wang, C., Li, Y., Tsuboshita, Y., Sakurai, T., et al. (2022). A high-generalizability machine learning framework for predicting the progression of Alzheimer's disease using limited data. NPJ Digital Medicine, 5(1): 43. https://doi.org/10.1038/s41746-022-00577-x

[10] Dai, Y., Zou, B., Zhu, C., Li, Y., et al. (2023). DE-JANet: A unified network based on dual encoder and joint attention for Alzheimer's disease classification using multi-modal data. Computers in Biology and Medicine, 165: 107396. https://doi.org/10.1016/j.compbiomed.2023.107396

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

[12] Huang, Y., Li, W. (2023). Resizer swin transformer-based classification using sMRI for Alzheimer's disease. Applied Sciences, 13(16): 9310. https://doi.org/10.3390/app13169310

[13] Malik, I., Iqbal, A., Gu, Y.H., Al-Antari, M.A. (2024). Deep learning for Alzheimer's disease prediction: A comprehensive review. Diagnostics, 14(12): 1281. https://doi.org/10.3390/diagnostics14121281

[14] Dai, W., Zhang, Z., Tian, L., Yu, S., Wang, S., Dong, Z., Zheng, H. (2022). Multimodal brain disease classification with functional interaction learning from single fMRI volume. arXiv preprint arXiv:2208.03028. https://doi.org/10.48550/arXiv.2208.03028

[15] Folego, G., Weiler, M., Casseb, R.F., Pires, R., Rocha, A. (2020). Alzheimer's disease detection through whole-brain 3D-CNN MRI. Frontiers in Bioengineering and Biotechnology, 8: 534592. https://doi.org/10.3389/fbioe.2020.534592

[16] Liu, N., Yuan, Z., Tang, Q. (2022). Improving Alzheimer's disease detection for speech based on feature purification network. Frontiers in Public Health, 9: 835960. https://doi.org/10.3389/fpubh.2021.835960

[17] Fan, X., Li, H., Liu, L., Zhang, K., et al. (2024). Early diagnosing and transformation prediction of Alzheimer's disease using multi-scaled self-attention network on structural MRI Images with Occlusion Sensitivity Analysis. Journal of Alzheimer's Disease, 97(2): 909-926. https://doi.org/10.3233/JAD-230705

[18] Alphonse, S., Mathew, F., Dhanush, K., Dinesh, V. (2025). Federated learning with integrated attention multiscale model for brain tumor segmentation. Scientific Reports, 15(1): 11889. https://doi.org/10.1038/s41598-025-96416-6

[19] Mahmud, T., Barua, K., Habiba, S.U., Sharmen, N., Hossain, M.S., Andersson, K. (2024). An explainable ai paradigm for alzheimer's diagnosis using deep transfer learning. Diagnostics, 14(3): 345. https://doi.org/10.3390/diagnostics14030345

[20] Karim, S.S., Fahad, M.S., Rathore, R.S. (2024). Identifying discriminative features of brain network for prediction of Alzheimer's disease using graph theory and machine learning. Frontiers in Neuroinformatics, 18: 1384720. https://doi.org/10.3389/fninf.2024.1384720

[21] Kishor Kumar Reddy, C., Ahmed, H.I., Mohzary, M., Monika Singh, T., Shuaib, M., Alam, S., Alnami, H.M. (2025). AlzheimerViT: harnessing lightweight vision transformer architecture for proactive Alzheimer's screening. Frontiers in Medicine, 12: 1568312. https://doi.org/10.3389/fmed.2025.1568312

[22] Al-Nuaimi, A.H., Blūma, M., Al-Juboori, S.S., Eke, C.S., Jammeh, E., Sun, L., Ifeachor, E. (2021). Robust EEG based biomarkers to detect Alzheimer's disease. Brain Sciences, 11(8): 1026. https://doi.org/10.3390/brainsci11081026

[23] Robin, J., Harrison, J.E., Kaufman, L.D., Rudzicz, F., Simpson, W., Yancheva, M. (2020). Evaluation of speech-based digital biomarkers: review and recommendations. Digital Biomarkers, 4(3): 99-108. https://doi.org/10.1159/000510820

[24] Kwak, M.G., Su, Y., Chen, K., Weidman, D., et al. (2023). Self-supervised contrastive learning to predict the progression of Alzheimer's disease with 3D amyloid-PET. Bioengineering, 10(10): 1141. https://doi.org/10.3390/bioengineering10101141

[25] Dao, D.P., Yang, H.J., Kim, J., Ho, N.H. (2024). Longitudinal Alzheimer's disease progression prediction with modality uncertainty and optimization of information flow. IEEE Journal of Biomedical and Health Informatics, 29(1): 259-272. https://doi.org/10.1109/JBHI.2024.3472462

[26] Raza, M.L., Hassan, S.T., Jamil, S., Hyder, N., Batool, K., Walji, S., Abbas, M.K. (2025). Advancements in deep learning for early diagnosis of Alzheimer's disease using multimodal neuroimaging: Challenges and future directions. Frontiers in Neuroinformatics, 19: 1557177. https://doi.org/10.3389/fninf.2025.1557177

[27] Mubonanyikuzo, V., Yan, H., Komolafe, T.E., Zhou, L., Wu, T., Wang, N. (2025). Detection of Alzheimer disease in neuroimages using vision transformers: systematic review and meta-analysis. Journal of Medical Internet Research, 27: e62647. https://doi.org/10.2196/62647

[28] Zhou, J., Wei, Y., Li, X., Zhou, W., Tao, R., Hua, Y., Liu, H. (2025). A deep learning model for early diagnosis of Alzheimer's disease combined with 3D CNN and video Swin transformer. Scientific Reports, 15(1): 23311. https://doi.org/10.1038/s41598-025-05568-y

[29] Hu, Z., Wang, Z., Jin, Y., Hou, W. (2023). VGG-TSwinformer: Transformer-based deep learning model

for early Alzheimer's disease prediction. Computer Methods and Programs in Biomedicine, 229: 107291. https://doi.org/10.1016/j.cmpb.2022.107291

[30] Zhang, Y., Teng, Q., He, X., Niu, T., Zhang, L., Liu, Y., Ren, C. (2023). Attention-based 3D CNN with multi-layer features for Alzheimer's disease diagnosis using brain images. In 2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), Sydney, Australia, pp. 1-4. https://doi.org/10.1109/EMBC40787.2023.10340536

[31] Khatri, U., Kwon, G.R. (2024). Diagnosis of Alzheimer's disease via optimized lightweight convolution-attention and structural MRI. Computers in Biology and Medicine, 171: 108116. https://doi.org/10.1016/j.compbiomed.2024.108116