Rel-SA: Alzheimer's Disease Detection using Relevance-augmented Self Attention by Inducing Domain Priors in Vision Transformers

Madhumitha V*[1], Sunayna Padhye*[1], Shanawaj S Madarkar*[1,2], Susmit Agrawal[1], Konda Reddy Mopuri[1]
[1]Department of Artificial Intelligence, Indian Institute of Technology Hyderabad, [2]Indian Navy

Abstract

Neurodegenerative diseases like Alzheimer's Disease (AD) present unique clinical challenges due to complex, progressive brain atrophy[1] patterns. Structural Magnetic Resonance Imaging (sMRI) is a critical tool for diagnosis of such neurodegenerative diseases. However, current methods often lack explainability and fail to highlight clinically meaningful regions from a clinician's perspective. Identifying key biomarkers that effectively distinguish patients with AD from healthy individuals using 3D sMRI scans thus remains a central challenge. To address this, we propose Relevance-augmented Self-Attention (Rel-SA), a neuroclinical knowledge–informed attention mechanism for Vision Transformers (ViTs). Rel-SA introduces a Relevance Bias (Rel-Bias), integrating insights from the AAL3 and JHU WM brain atlases to guide the model toward regions implicated in AD progression. Through qualitative and quantitative evaluations, we demonstrate that Rel-SA not only boosts diagnostic accuracy over ViT-base by $\sim 4\%$ but also enhances model interpretability efficiently with an addition of only 24 parameters.

Our work highlights the importance of incorporating clinical priors into model design and provides an effective approach to embed domain knowledge into existing architectures, resulting in more robust and interpretable deep learning solutions for neuroimaging.

1. Introduction

"Mapping the elusive patterns of the mind" has been a timeless quest. Linking subtle structural changes in the brain tissue with early signs of neurodegenerative diseases remains a significant challenge in neuroscience. Dementia currently affects more than 60 million people worldwide, a figure expected to soar to 139 million by 2050 [3]. Alzheimer's disease (AD) unfolds gradually over decades, progressing from a currently undetectable preclinical phase to severe dementia and ultimately death [44]. Given AD's prolonged progression and diagnostic challenges that span clinical, behavioural, psychological, and pathological assessments, no single study has been able to document the full evolution of the disease [44], leaving its timeline partially hypothetical. The progression of Alzheimer's involves volumetric brain atrophies, with structural MRI (sMRI) images offering a reflective representation of these brain changes [4]. Therefore, sMRI has long been recommended for clinical assessments [4, 46].

Deep learning has revolutionized computer vision, with CNN-based approaches excelling on smaller datasets due to their strong inductive biases [17, 27]. Of late, Vision Transformers (ViTs) [17] have outperformed CNNs on various benchmarks by effectively capturing long-range dependencies between image regions and offering expansive receptive fields [26, 54]. However, adapting ViTs from 2D to 3D is nontrivial, as it involves increased complexity in 3D data representation [12], a lack of inherent spatial priors (making them data-hungry), and limited availability of pre-trained models for transfer learning. These challenges are more pronounced in neuroimaging, where data is high-dimensional yet scarce [25]. For neurodegenerative diseases like AD, these shortcomings are particularly consequential: diagnosis hinges on identifying subtle, stage-specific atrophy patterns that evolve across neuroanatomically ordered regions. For example, clinical neuroscience research shows that early atrophy localizes to the hippocampus and amygdala, impairing memory and emotional processing [4, 44]. As AD progresses, degeneration expands to the entorhinal cortex, middle temporal gyrus (impacting semantic memory) [9], and eventually the striatum, thalamus, and widespread cortical regions (middle frontal, cingulate, parietal, and insular cortices) [40, 56]. Diffusion tensor imaging further reveals white matter degradation, with decreased integrity in AD-vulnerable tracts [55]. Current computer vision approaches often overlook these neuroscientific insights. Indeed, recent transformer-based works, while delivering good performance on small-scale datasets, remain largely data-driven [30, 56] black boxes.

*Equal contribution

[1]Brain atrophy refers to the loss of neurons and connections between them, leading to a reduction in brain size and function.

They achieve high accuracy on small datasets but often fail to provide meaningful interpretations, potentially attending to irrelevant brain regions to make predictions rather than considering known biomarkers.

A natural question arises: Can we incorporate crucial domain-specific knowledge, both clinical and empirical, as an inductive bias for ViT-based models to learn effectively? And how can we do that? To address this, we propose **Brain Atlas**[2]**-based clinical knowledge induction** within the model, such as ViTs, embedding clinical insights directly via our novel **Relevance Augmented Self Attention (Rel-SA)** module. In this context, **relevance** denotes the explicit clinical prioritization of disease-associated brain regions or biomarkers. Our experiments show that this explicit anatomical guidance enhances the ability of transformers to differentiate AD and CN subjects while concurrently ensuring that the models agree with clinical knowledge. We demonstrate the effectiveness of our approach through qualitative visualizations and quantitative results (§4) on two benchmark datasets (ADNI & AIBL). We achieve performance comparable to non-interpretable methods while simultaneously providing interpretations better aligned with clinical insights.

Our contributions can be summarized as follows:

- **Rel-SA, encoding clinical priors using Relevance Bias**: We encode the clinical prominence of structural markers critical in AD within the attention matrix in a ViT through an additive bias term, which we call **Relevance Bias** (abbreviated as **Rel-Bias**). We derive these markers from the meta-analysis of neuroscience literature. During training, the model learns to balance evidence-based priors from neuroscience (the Rel-Bias term) with data-driven attention scores. Importantly, this lightweight module adds only 24 parameters in ViT-Base and significantly improves both performance and interpretability.
- We unify two clinically validated brain atlases: 1) the **AAL3v1 atlas** [47], parcellating the brain into 166 cortical/subcortical regions [3], and 2) the **JHU White Matter atlas** [1], segmenting white matter into 48 tracts. This integration ensures comprehensive coverage of anatomically distinct regions. Each 3D MRI volume is mapped to the combined Atlas parcels. This parcellation serves as the spatial prior for our novel attention mechanism. To the best of our knowledge, this is the first work to integrate domain priors from combined brain atlases into a ViT architecture.
- **Qualitative Evaluation through Leave-One-Out Analysis**: We derive ten region sets ranked by their relevance (high to low) to AD diagnosis and evaluate model interpretability using a Leave-One-Out strategy (§5.2). We identify regions critical to prediction by masking each set and measuring changes in entropy, accuracy, and AUC. Additionally, a Reverse Leave-One-Out analysis (§A.3), retaining only one region set at a time, was conducted for completeness. Results confirm that Rel-SA consistently localizes attention to clinically meaningful areas. This highlights the model's strong alignment with neuroscientific insights.

[2]A brain atlas is a detailed map of the brain, dividing it into parcels. A parcel in brain research refers to a defined region of the brain, grouped based on structural and functional similarity.

[3]Cortical Region: It refers to a specific area of the brain's outer layer (the cortex) also called as grey matter, is responsible for key functions like thinking, memory, sensation, and movement. A subcortical region: It is an area beneath the brain's outer layer (cortex). White matter: It consists of nerve fibres (axons) coated with a fatty substance called the myelin sheath.

2. Related Work

The general landscape of architectures for medical image analysis, particularly for ADNI classification, can be divided into approaches that prioritize performance and those that emphasize interpretability, often with a trade-off between the two. Below, we discuss the two paradigms.

Black-Box Approaches: Earlier works using CNN-based architectures handle 3D sMRI scans as a sequence of 2D image slices [14, 19, 43, 45, 48]. These works overlook changes in volumetric attributes over spatially distant regions, a key indicator of numerous neurodegenerative diseases. Additionally, 3D CNNs require greater depth to fully capture the global features necessary for understanding 3D images, which is computationally expensive and memory-intensive [33]. More recently, Vision Transformer [17] has been used in 3D-input-based neurodegenerative disease diagnosis to overcome CNN limitations in modelling global context. They achieve high accuracy in tasks like AD diagnosis, primarily through data-driven learning of volumetric features [30, 37, 53, 56]. However, their reliance on purely data-driven learning often results in a lack of interpretability [38]. This limits their utility in clinical settings, where understanding ***why*** a model makes a prediction is as important as the prediction itself. The next frontier is marrying these powerful models with wisdom from clinical neuroscience – explicitly directing them to disease biomarkers.

Interpretable approaches: Some recent works have attempted to balance performance and interpretability by focusing on localized region-of-interest (ROI) analyses or leveraging complex frameworks to integrate domain knowledge [2, 20, 51]. While interpretable, these methods often fail to capture the widespread and diffuse patterns of neurodegeneration characteristic of AD. Therefore, models must balance attention across *both* high and low relevance regions to holistically capture evolving biomarkers. Rel-SA strives to achieve such balance, preserving the transformer's ability to attend to novel patterns across the entire brain while nudging attention to prioritize clinically relevant regions. This dual mechanism (anatomical and clinical prior, alongside data-driven attention) enables the model to

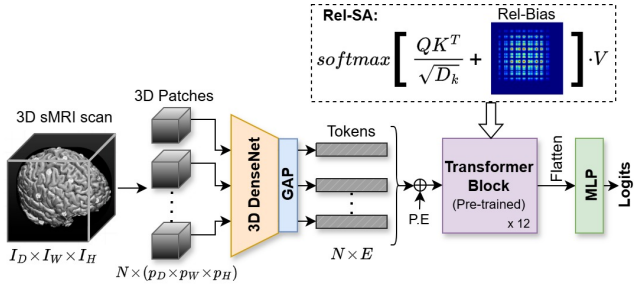

Figure 1. 3D Densenet+ViT Architecture showing the introduction of Rel-Bias in the Relevance-augmented Self-Attention module (Rel-SA) of the Transformer blocks. (GAP: Global Average Pooling, P.E: Position Embeddings, MLP: Multi-Layer Perceptron)

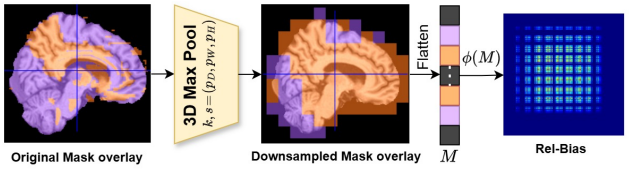

Figure 2. Rel-Bias calculation using relevance regions identified using Atlas. The downsampled masks are precomputed and stored before training. During training, the mask M is sent to each transformer block where its respective learnable parameters are used to calculate Rel-Bias.

1) leverage established biomarkers and 2) discover patterns that may be overlooked by traditional markers.

3. Architecture Design

Consider a subject's sMRI scan represented as a greyscale 3D volume $\mathbf{I} \in \mathbb{R}^{1 \times I_D \times I_W \times I_H}$ after preprocessing (§4). We divide the 3D scan into N non-overlapping patches, each patch $\mathbf{P} \in \mathbb{R}^{p_D \times p_W \times p_H}$. Each patch is embedded into a E dimensional token vector using a 3D DenseNet-like architecture. Note that each patch is embedded independently of all other patches, similar to patchification in 2D ViT models. Learnable positional embeddings are added to encode spatial information effectively. This list of tokens, $\mathbf{X} \in \mathbb{R}^{N \times E}$, is then sent to a Transformer model. The output embeddings are then flattened and sent to a single MLP layer for binary classification. We use the standard cross-entropy loss for the training. Figure 1 describes this architecture in detail.

3.1. Rel-SA: Relevance-augmented Self-Attention

We propose *Rel-SA* as a straightforward yet powerful way to inject neuroscientific domain knowledge from brain atlases into the ViT self-attention module. Concretely, we add a *relevance bias* to the attention matrix prior to the softmax operation, guiding the model to give preferential focus to the regions of the brain known to be critical for Alzheimer's

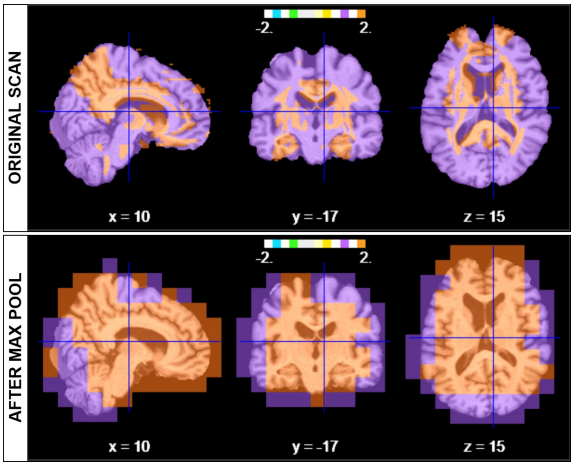

Figure 3. Visualization of high relevant (in orange) and low relevant (in violet) brain regions by overlaying the $3D$ binary masks, M_{high} and M_{low} in all three planes (Sagittal, Coronal, and Axial).

diagnosis. Formally, the Rel-SA attention scores are computed as

$$\text{Rel-SA}(Q, K, V) = \sigma\Big(\frac{QK^T}{\sqrt{d}} + \phi(M; w_{high}, \alpha)\Big)V \quad (1)$$

where $\mathbf{Q} = \mathbf{X}W_Q$, $\mathbf{K} = \mathbf{X}W_K$, $\mathbf{V} = \mathbf{X}W_V$ are the query, key, and value projections of the token sequence $\mathbf{X} \in \mathbb{R}^{N \times E}$, and $\sigma(\cdot)$ is the softmax function. The term $\phi(\cdot)$ is our *Relevance Bias* (Rel-Bias) derived from atlases of known neuroanatomical markers of Alzheimer's pathology as described below.

Rel-Bias Derivation. We use two well-studied brain atlases — AAL3v1 and JHU WM — to identify parcels strongly associated with AD. Specifically, we compile:

- **50 parcels** from the AAL3v1 Atlas [4, 11, 15, 16, 18, 19, 22, 23, 31, 32, 34, 39, 41, 42, 44],
- **31 parcels** from the JHU WM Atlas [6, 40, 56].

These parcels are deemed *high relevance* because they frequently exhibit pathological changes in AD. We create two 3D binary masks per scan: one for these *high-relevance* regions and another for the *low-relevance* (remaining) regions. To align with our patch-wise tokens (see Figure 2, 3), we perform **3D Max-Pooling** (kernel and stride = (p_D, p_W, p_H)) on these masks. This downsampling step assigns an entire patch as *high relevance* if it contains *any* portion of a high-relevance parcel. Such a conservative approach mitigates the risk of ignoring crucial areas near parcel boundaries, e.g., if the Frontal Superior Gyrus abuts the background, the relevant patch still receives a strong bias toward this region. Figure A7 depicts volumetric coverage of high and low relevant regions pre and post-downsampling. Although the sagittal view in Figure 3 may visually suggest that a large fraction of the brain is marked *high rel-*

evance, the coronal slices reveal a more nuanced picture: many cortical areas are excluded, and only central regions known to be pivotal in AD diagnosis remain highlighted. Hence, while we prioritize capturing critical regions, we do not simply label most of the brain as "high relevance". After pooling, each mask is reshaped to yield a flattened vector $M \in \mathbb{R}^{N\times 1}$ where $M = M_{\text{high}} + M_{\text{low}}$. A *score vector* $S \in \mathbb{R}^{N\times 1}$ is then obtained via

$$S(M; w_{high}, \alpha) = w_{\text{high}} \cdot M_{\text{high}} \; + \; w_{\text{low}} \cdot M_{\text{low}}, \quad (2)$$

$$w_{\text{low}} = \alpha \cdot w_{\text{high}}, \quad \alpha \in (0, 1) \quad (3)$$

ensuring that $w_{\text{low}} < w_{\text{high}}$, where w_{high}, α are learnable parameters. Here, w_{high} and w_{low} are the scores of high and low relevant regions, respectively. The final *Rel-Bias matrix* $\phi(M)$ has dimensions $\mathbb{R}^{N\times N}$ and is computed as

$$\phi(M; w_{high}, \alpha) = \beta \; \cdot \; \left\| S(M)\, S(M)^T \right\|_F \quad (4)$$

where β is a hyperparameter controlling the bias scale. Our calculations reveal that the Rel-Bias term can modify the probability value of a given patch by up to 20%.

Domain-aware Inductive Bias in Self-Attention. By incorporating these relevance-informed masks into the self-attention matrix, Rel-SA effectively embeds *domain-aware inductive biases*. It preferentially elevates interactions among tokens that correspond to crucial neuroanatomical parcels, thereby enhancing both *performance* (focusing model capacity on disease-critical regions) and *interpretability* (making attention maps reflect clinically grounded priorities). Empirically, we show that this leads to higher classification accuracy and more meaningful attention distributions for AD detection (see §4).

4. Experiments and Results

4.1. Experimental Setup

In this section, we present our experimental design to evaluate the proposed Rel-SA module for AD diagnosis from 3D sMRI data. We compare against several baselines and state-of-the-art methods, reporting classification accuracy and area under the ROC curve (AUC). Our experiments span two major public datasets, **ADNI** and **AIBL**, to assess both in-domain performance and generalizability. Details of model training configuration and hyperparameters are elaborated in Table A4 in the Supplementary Material.

4.2. Datasets and Preprocessing

ADNI (Alzheimer's Disease Neuroimaging Initiative) [35]. We use T1-weighted sMRI scans from the ADNI collection, which comprises both 1.5T and 3T acquisitions taken at intervals of 6 months. We obtain 4,178 scans from 1,042 subjects. The data distribution is split into **Control Normal (CN)** scans, containing 3,006 scans from 719 subjects, and scans of patients with AD, consisting of 1,172 scans from 323 subjects.

AIBL (Australian Imaging, Biomarker, and Lifestyle Flagship Study) [24]. AIBL provides T1-weighted sMRI scans from 596 subjects, for a total of 1,097 scans. After the same preprocessing pipeline as ADNI, we obtain 940 CN scans from 497 subjects, and 157 AD scans from 104 subjects. Subjects range in age from 55–93 years.

We follow standard preprocessing: skull stripping, bias field correction, registration to MNI-152 space, and Z-score intensity normalization. The final input to the model has dimension $1 \times 182 \times 218 \times 182$, where the single channel indicates grayscale sMRI data. We use a stratified 5-fold subject-wise cross-validation with an 80:20 train-test split.

4.3. Baselines

We consider multiple baseline architectures to examine both the impact of different patch-encoding backbones and the effect of the Rel-SA module in isolation. In particular, our baselines consist of three ViT variants, each differing in how 3D patches are tokenized to create patch embeddings: (1) **Standard ViT**, which uses a linear projection; (2) **ResNet-ViT**, which replaces the linear projection with a 3D ResNet [28]; and (3) **DenseNet-ViT**, which uses a 3D DenseNet [29]. These three baselines isolate the choice of patch encoder, allowing us to verify that any performance benefits arise from *Rel-SA* rather than from improved patch feature extraction. Results in Table 1 clearly demonstrate that the inclusion of Rel-SA in the baseline variants boosts performance across all classification metrics consistently. Further, we also present β variations with the best-performing model, DenseNet-ViT. Results demonstrate that β=10 gives the best accuracy and AUC gains. We utilize the same model in §5 for interpretability analysis.

4.4. Comparison with Existing State-of-the-Art Methods for AD Classification

We compare our approach with state-of-the-art models for AD classification in Table 2. Two of the methods, MedicalNet [13] and M3T [33], utilize $3D$ scans as input; however, they do not fully align with clinical evidence as they miss out on considering important regions linked with AD [33]. Additionally, M3T presents results for low-resolution data input of shape $128 \times 128 \times 128$. The other two methods, ADDformer [37] and Khatri et.al [36], utilize slices of the 3D scans to predict the classification, which may cause them to miss out on volumetric information. None of these methods provides any mechanisms to incorporate domain knowledge, and the interpretations offered only partially highlight relevant regions at best. In contrast, Rel-SA is explicitly designed to embed domain knowledge into the attention mechanism of ViT models, leading to more clinically aligned and interpretable predictions. These compar-

Table 1. Ablation studies for inclusion of Rel-SA on different model architectures and β values in 2-class (AD & CN) classification task. This table shows the comparison of the first three rows (w/o Rel-SA) using three different model architecture settings with the next three rows (w/ Rel-SA). The last five rows show the impact of varying β values in Rel-SA. All ViT models utilized pre-trained weights from Hugging-face's ViT-base (for only the transformer layers)

Model	β	Accuracy (%) (↑)	AUC (%) (↑)	Precision (%) (↑)	Recall (%) (↑)	F1-score (%) (↑)	Loss(↓)
MLP + ViT	-	77.29	75.51	70.42	71.03	70.7	0.89
3D ResNet + ViT	-	78.00	75.37	73.86	69.15	70.68	0.64
3D DenseNet + ViT	-	79.86	81.42	76.77	70.60	72.52	0.53
MLP + ViT + Rel-SA	10	80.02 (↑ 2.73)	82.05 (↑ 6.54)	75.82 (↑ 5.4)	72.03 (↑ 1.00)	73.43 (↑ 2.73)	0.71 (↓ 0.18)
3D ResNet + ViT + Rel-SA	10	82.05 (↑ 4.05)	84.23 (↑ 8.86)	79.42 (↑ 5.56)	76.72 (↑ 7.57)	76.97 (↑ 6.29)	0.53 (↓ 0.11)
3D DenseNet + ViT + Rel-SA	10	**83.16** (↑ 3.3)	**85.66** (↑ 4.24)	**80.66** (↑ 3.89)	**76.89** (↑ 6.29)	**77.71** (↑ 5.19)	**0.52** (↓ 0.01)
3D DenseNet + ViT + Rel-SA	0.1	77.05	78.80	72.41	65.31	66.89	0.53
3D DenseNet + ViT + Rel-SA	1	82.16	84.82	**82.82**	73.83	76.46	**0.5**
3D DenseNet + ViT + Rel-SA	10	**83.16**	**85.66**	80.66	**76.89**	**77.71**	0.52
3D DenseNet + ViT + Rel-SA	50	81.69	83.85	77.47	76.29	76.83	0.54
3D DenseNet + ViT + Rel-SA	100	81.57	84.1	80.92	75.77	77.64	0.56

Table 2. Performance comparison with SOTA models on ADNI and AIBL datasets for binary classification between Alzheimer's Disease (AD) and Cognitive Normal (CN). Results are reported as mean±std for stratified 5-fold subject-wise splits. The "Setup" column indicates the type of input used by the model. The "Explanation" column denotes whether the model provides behaviour analysis with respect to clinically relevant explanations, as presented in their respective studies.

Dataset	**ADNI**		**AIBL**		**Setup**	**#Samples**	**Explanation**
Model	**Acc. (%)**	**AUC (%)**	**Acc. (%)**	**AUC (%)**		(ADNI, AIBL)	
MedicalNet [13] [33] (CNN)	88.89	88.80	82.76	79.07	3D Volume	4786, 817	×
ADDformer [37]	88.2	96.00	-	-	Single-slice	388, 0	×
Khatri et.al [36]	95.37	96.00	-	-	Single-slice	705, 0	partial
M3T [33]	93.21	96.34	93.27	92.58	Multi-slice, Multi-plane	4786, 817	partial
3D DenseNet + ViT	79.86±0.94	81.42±1.83	82.81±1.55	81.42±2.29	3D Volume	4178, 1097	×
3D DenseNet + ViT + Rel-SA	83.16±0.54	85.66±0.87	87.40±2.58	83.57±2.25	3D volume	4178, 1097	✓

isons help us assess classification performance and model explanations with established AD-related regions derived from clinical neuroscience literature, with our results in Table 2 indicating that this approach achieves performance comparable to state-of-the-art methods, while simultaneously being highly interpretable. We also note that there exist inconsistencies in the training and evaluation splits used in existing literature. Hence, we report our results on the largest dataset available to us.

5. Analysis

We present our results in Tables 1 and 2 for quantitative analysis. We find that adding Rel-SA empirically improves performance across all baselines, supporting our hypothesis that injecting prior knowledge of disease-relevant regions improves the model's performance. We additionally provide qualitative results indicating the effectiveness of Rel-SA in providing clinically aligned post-hoc explanations, thus improving the reliability over baseline models. Finally, we perform a Leave-One-Out analysis for a fine-grained comparison of the behaviour shown by models with and without Rel-SA by performing interventions on the input data to simulate atrophy in specific brain regions. Our results indicate that the models trained with Rel-SA strongly align with inferences from clinical theories regarding the effect of AD on various brain regions, while models without Rel-SA show significant deviations. This further showcases the ability of models trained with Rel-SA to provide clinically relevant interpretations.

5.1. Qualitative Results

In this section, we compare the differences between the vanilla ViT and the Rel-SA-augmented ViT in terms of how they allocate their attention under three different scenarios: (a) both models correctly classify an AD scan, (b) Rel-SA correctly identifies AD while the vanilla ViT misclassifies the same scan as CN, and (c) both models correctly classify a CN scan. For a better appreciation of the anatomical regions, classical visual cues for Alzheimer's disease are presented in Figure A6 of the appendix. We also visualize the effects of including Rel-SA by qualitatively comparing the attended regions in models with and without the module in Figure 4, with a more comprehensive visualization in the appendix Figure A8.

Case 1: GT = AD, both models predict AD: Examining sagittal slices (e.g., slice 90 in Figure 4) reveals that the Rel-SA model concentrates heavily on the parietal and mid-temporal lobes, regions well-documented for AD-related atrophy [44]. The vanilla ViT exhibits scant attention but no-

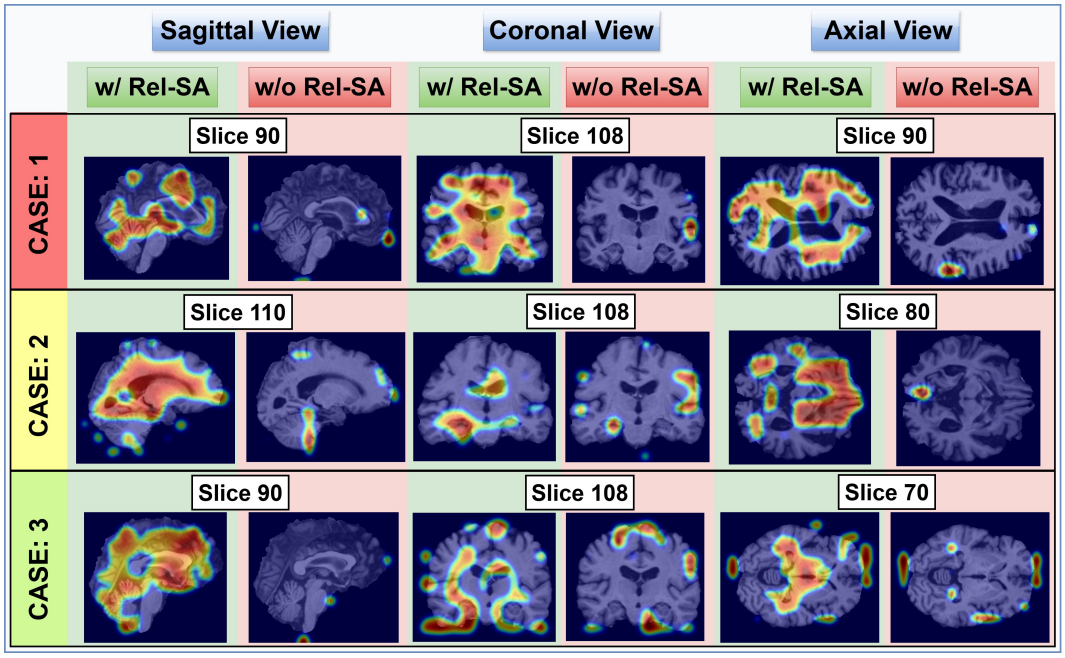

Figure 4. Case-wise Comparative Analysis (§5.1) of Attention Rollout visualizations between $3D$ image scans belonging to the same subject, when processed with Rel-SA augmented ViT and without Rel-SA (Vanilla ViT). Cases demonstrated are, Case 1: GT = AD, both models predict AD; Case 2: GT = AD, Rel-SA predicts AD while vanilla ViT predicts CN; and Case 3: GT = CN, both models predict CN.

tably overlaps with Rel-SA in the mid-temporal area, likely enabling it to arrive at the correct AD label. In coronal slice 108 and axial slice 90, Rel-SA highlights classic disease markers such as enlarged ventricles, hippocampal atrophy, and sulcal widening [21], while the vanilla ViT's attention is weaker yet still sufficient to classify the scan correctly. Thus, although both models converge on an AD prediction, Rel-SA provides a more clinically relevant distribution of attention aligned with known biomarkers.

Case 2: GT = AD, Rel-SA predicts AD while vanilla ViT predicts CN: In this instance, Rel-SA consistently attends to hallmark AD indicators—including the parietal cortex, medial temporal regions, and ventricular enlargement across multiple slices (e.g., slices 110, 108, 80). Conversely, the vanilla ViT's attention is scattered or muted in precisely these disease-relevant structures, causing it to overlook key morphological cues of AD. For example, in slice 108, Rel-SA clearly demarcates the ventricles and hippocampus, whereas the vanilla ViT yields only faint attention in those locations. This lack of focus culminates in a CN classification by the vanilla ViT, illustrating how Rel-SA's targeted attention mechanism can prevent crucial AD features from being missed.

Case 3: GT = CN, both models predict CN: When a healthy subject is correctly identified, Rel-SA's heatmaps show pronounced focus on "healthy" anatomical landmarks, such as the hippocampus, thalamus, cingulate gyrus, and overall cortical structure indicating a thorough check for the *absence* of AD pathology (e.g., undilated ventricles, lack of substantial cortical thinning [9]). In contrast, the vanilla ViT's attention maps are less consistently concentrated on these canonical healthy regions, although they still arrive at the correct CN decision. These differences imply that Rel-SA's attention mechanism provides a more methodical survey of known healthy markers, while vanilla ViT's scattered coverage, although still adequate for classification, does not provide relevant interpretations.

In summary, these three scenarios illustrate that Rel-SA systematically directs attention toward anatomically relevant sites in both diseased and healthy subjects, allowing for clinically consistent predictions. The vanilla ViT may sometimes succeed (Cases 1 and 3), but its less targeted approach risks missing key AD markers (Case 2) or neglecting standard healthy indicators, thus sacrificing interpretability.

5.2. Leave-One-Out Analysis

Rel-SA intuitively imbues information about the importance of certain brain regions in the prediction of AD to the attention matrix. To verify if the model effectively learns the contribution of specific brain regions in AD prediction, we perform a Leave-One-Out analysis, masking out a set of regions from the input scan before passing it to the model. For this analysis, we divide the brain into ten subregions based on clinical literature, neuropathological staging, and known AD-relevant atrophic patterns. We then create 10 binary masks (M1–M10), each mask occluding a specific set

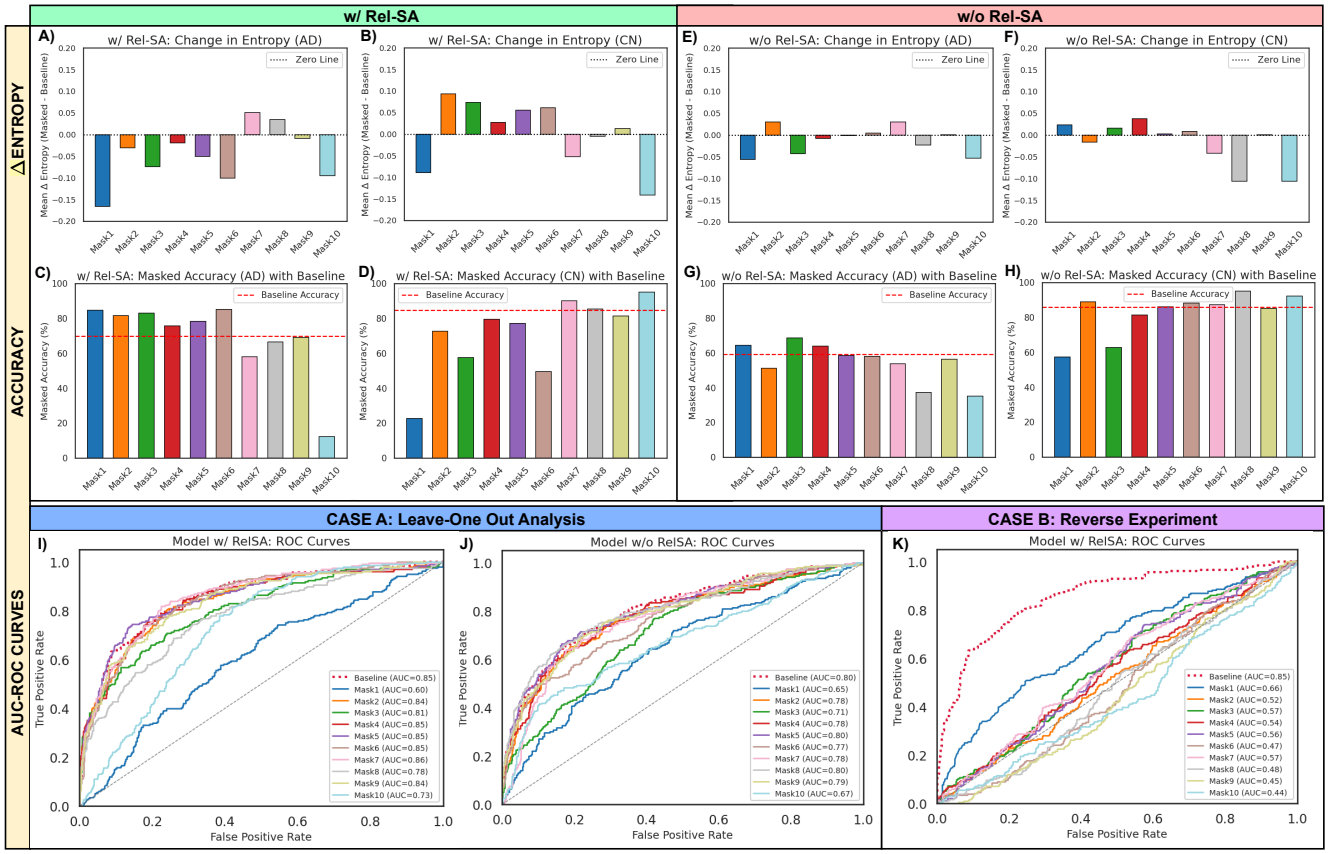

Figure 5. We compare models with and without Rel-SA using the Leave-One-Out analysis described in §5.2. We compare the change in entropies of model predictions for patients with AD (plots A and E) and the CN subjects (plots B and F). Plots C and G contain the accuracy of AD predictions with and without Rel-SA, respectively, while plots lots D and H provide the same analysis for CN subjects. Plots I and J show the AUC curves for models with and without Rel-SA, while the final plot K is a visualization of the AUC curve in the "Reverse" Leave-One-Out analysis (§A.3).

of regions. We discuss the regions covered in each mask in §A.2 of the Supplementary Material.

In contrast to standard computer vision applications, where occluding or masking out a part of an image hides certain visual cues, removing a region in sMRI can mimic "tissue loss". Although this does not perfectly reflect true atrophy, our masking experiments suggest that the model may interpret missing voxels as pathological degenerations. To systematically investigate this, we conduct a selective region removal analysis and assess its impact on model predictions. For this analysis, we evaluate changes in entropy of logit probabilities, class-wise accuracy variations, and ROC curve shifts following the removal of specific brain regions to quantify how the model adjusts its confidence and classification behavior in response to missing anatomical structures. The results are summarized in Figure 5.

First, **Mask 1** (covering regions typically affected in early AD like the hippocampus): reveals that removal of the regions contained in it drastically lowers the RelSA-augmented ViT overall AUC (e.g., from a baseline of 0.85 to about 0.60) and pushes the model towards high-confidence AD predictions as indicated by the predictions' entropies in Figure 5 (A,B,C,D,I). This behavior implies that the network views absent or severely altered voxels in these early-atrophy-prone areas as a strong AD cue. Correspondingly, the same masking impairs CN accuracy, reflecting the role of a preserved Mask 1 region in recognizing healthy structures. In contrast, the vanilla ViT exhibits a more *modest* performance decline when M1 is removed, and it does *not* always become highly confident about AD (Figure 5 (E,F,G,H,J)). This less pronounced effect suggests the vanilla ViT is *not as sensitive* to early atrophy signals in Mask1, aligning less closely with clinical expectations. These results indicate that the Rel-SA-augmented ViT better adheres to clinical research - which states that the regions within Mask1 show high levels of atrophy in patients with AD [4]. Evidently, not having Rel-SA makes it hard for the vanilla ViT model to derive this relationship between AD and the brain regions involved.

Masks 2–6: encompass mid-brain and sensorimotor cor-

tices commonly implicated in later AD stages. Occluding these regions yields moderate decreases in AUC (ranging from $\sim$0.73 to 0.83) and typically pushes the ViT having Rel-SA to classify subjects as AD. In particular, the model gains confidence in its AD predictions as quantified by entropy values in Figure 5 (A). The accuracy of AD classification increases while the accuracy of CN classification goes down (Figure 5 (C,D)). However, the effects of these regions are not as prominent as the regions covered in Mask 1. This pattern suggests both confounding effects (masking may appear as atrophy) and the model's ability to rely on alternative areas for discrimination. Additionally, CN classification accuracy reduces from removing these mid-brain voxels, possibly due to misclassification from CN to AD, signaling that the network looks for relatively intact mid-brain regions in healthy scans to predict CN. In contrast, vanilla ViT (Figure 5 (E,F,G,H,J)) exhibits inconsistent shifts in performance. For some masks (e.g., Mask 3, Mask 4), we see a modest increase in AD accuracy, suggesting the model sometimes treats masked mid-brain tissue as an AD cue. In other cases, accuracy and confidence show little or no change. This analysis highlights the importance of domain knowledge in identifying AD from regions that do not have significant visual atrophy in the early stages of AD.

Mask 7: Although ranked seventh in relevance, its removal causes a mild yet noticeable effect on both AD and CN classification. For the Rel-SA model, we observe a small drop in AD AUC (e.g., from $\sim$0.82 to $\sim$0.78), alongside a modest rise in CN accuracy as seen in Figure 5 (A,B,C,D). These shifts, while not as pronounced as higher-priority atrophy regions, suggest that Mask 7 might contribute relevant visual cues derived from data rather than acting as mere background. The vanilla ViT similarly exhibits some impact on Mask 7 removal, though its changes in accuracy and entropy are erratic without the guidance provided by Rel-SA. Thus, introducing Rel-SA "nudges" the model toward more consistent, clinically aligned patterns.

Masks 8–9: In contrast, removing Masks 8 and 9 (considered lower-priority regions for early AD) generally leaves AD detection performance near the baseline across the different metrics, indicating they offer fewer decisive disease cues than Mask 7. At the same time, the Rel-SA model often sees a slight boost in CN accuracy and confidence, implying that these regions do not strongly support AD identification but hold some information beneficial for distinguishing healthy scans. In contrast, in the case of the vanilla ViT, removing Mask 8 can lead to an increase in CN accuracy and confidence, but Mask 9 shows negligible correlation with classification outcomes, reflecting a more variable reliance on these sub-regions without Rel-SA's guidance.

Lastly, **Mask 10** (cerebellum): demonstrates a curious effect for both the Rel-SA and vanilla ViT models: although classic AD progression literature does not prioritize cerebellar changes early on, removing it can decrease performance for both AD and CN, accompanied by a small drop in AUC. This indicates that both models have unexpectedly leveraged cerebellar cues, possibly due to dense tissue signatures resembling cortical regions, proximity to medial temporal lobes, or spurious data artifacts.

6. Conclusion

This study explores how domain knowledge can influence a model's performance in predicting the presence of Alzheimer's Disease. Our results depict that integrating domain knowledge from Neuroscience into AI models can have a considerable impact on their intricate understanding of the sMRI data. The key lies in leveraging domain knowledge as the right inductive bias. The proposed method (Rel-SA), presents a simple, efficient, and highly effective method in this direction. It achieves a maximum accuracy improvement of approximately $\sim 4\%$ when integrated into a hybrid $3D$ ResNet-ViT architecture. Furthermore, the adaptability of Rel-SA is validated on the AIBL dataset, yielding a similar performance gain of $\sim 4.6\%$. Notably, this improvement is achieved with a negligible parametric overhead of just 24 additional parameters. In addition, interpretations derived from the model enhanced with Rel-SA align closely with established clinical literature on AD, as evidenced by findings from leave-one-out analysis. The model's performance is further supported by visual interpretations that are consistent with domain knowledge.

7. Discussion and Future work

We expect our study to achieve broader generalization across populations and AD subtypes with access to larger and more diverse labelled sMRI datasets. Minor residual skull fragments were observed in countable ADNI2 scans due to variability in skull stripping, which we addressed through careful manual review. Moreover, recent studies suggest that omitting skull-stripping may preserve informative signals relevant to prediction tasks [52], making this a valuable direction for future investigation. Spurious correlations involving cerebellar regions remain another area for investigation.

Looking ahead, we plan to extend our framework beyond binary classification to multi-stage and multi-disease settings such as Parkinson's and frontotemporal dementia using disease-specific atrophy masks [15, 49]. Longitudinal data will further enable sub-mask-driven tracking of early to late atrophy stages aligned with Braak's staging [9]. We also aim to integrate complementary modalities like PET, fMRI, and clinical test scores to enrich biomarker representations. Additionally, evaluating domain shift scenarios will enhance real-world generalization and clinical utility.

References

[1] Neurovault image 1401, 2024. Accessed: 2024-11-11. 2

[2] S. Ahmed, B.C. Kim, K.H. Lee, H.Y. Jung, and Alzheimer's Disease Neuroimaging Initiative. Ensemble of roi-based convolutional neural network classifiers for staging the alzheimer disease spectrum from magnetic resonance imaging. *PLoS One*, 15(12):e0242712, 2020. 2

[3] Alzheimer's Disease International. Dementia statistics, 2024. Accessed: 2024-11-11. 1

[4] L. G. Apostolova, A. E. Green, S. Babakchanian, K. S. Hwang, Y.-Y. Chou, A. W. Toga, and P. M. Thompson. Hippocampal atrophy and ventricular enlargement in normal aging, mild cognitive impairment and alzheimer's disease. *Alzheimer Disease and Associated Disorders*, 26(1): 17, 2012. 1, 3, 7

[5] Steven E. Arnold et al. The topographical and neuroanatomical distribution of neurofibrillary tangles and neuritic plaques in the cerebral cortex of patients with alzheimer's disease. *Cerebral Cortex*, 1(1):103–116, 1991. 1

[6] J. Barnes, R. I. Scahill, J. M. Schott, C. Frost, M. N. Rossor, and N. C. Fox. Does alzheimer's disease affect hippocampal asymmetry? evidence from a cross-sectional and longitudinal volumetric mri study. *Dementia and Geriatric Cognitive Disorders*, 19(5-6):338–344, 2005. 3, 1

[7] Randall J. Bateman et al. Clinical and biomarker changes in dominantly inherited alzheimer's disease. *New England Journal of Medicine*, 367(9):795–804, 2012. 1

[8] Y Blinkouskaya and J Weickenmeier. Brain shape changes associated with cerebral atrophy in healthy aging and alzheimer's disease. *Frontiers in Mechanical Engineering*, 7:705653, 2021. 5

[9] H. Braak and E. Braak. Neuropathological staging of alzheimer-related changes. *Acta Neuropathologica*, 82(4): 239–259, 1991. 1, 6, 8, 2

[10] Randy L. Buckner et al. Cortical hubs revealed by intrinsic functional connectivity: mapping, assessment of stability, and relation to alzheimer's disease. *Journal of Neuroscience*, 29(6):1860–1873, 2009. 1

[11] M.S. Byun, S.E. Kim, J. Park, D. Yi, Y.M. Choe, B.K. Sohn, H.J. Choi, H. Baek, J.Y. Han, J.I. Woo, et al. Heterogeneity of regional brain atrophy patterns associated with distinct progression rates in alzheimer's disease. *PLoS One*, 10(11): e0142756, 2015. 3, 1

[12] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020. 1

[13] Sihong Chen, Kai Ma, and Yefeng Zheng. Med3d: Transfer learning for 3d medical image analysis, 2019. 4, 5

[14] Z. Dai, H. Liu, Q. V. Le, and M. Tan. Coatnet: Marrying convolution and attention for all data sizes, 2021. 2

[15] A. Damulina et al. Cross-sectional and longitudinal assessment of brain iron level in alzheimer disease using 3-t mri. *Radiology*, 296(3):619–626, 2020. 3, 8, 1

[16] M.A. DeTure and D.W. Dickson. The neuropathological diagnosis of alzheimer's disease. *Molecular Neurodegeneration*, 14(1):32, 2019. 3, 1

[17] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, and S. Gelly et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2020. 1, 2

[18] R. Duara and W. Barker. Heterogeneity in alzheimer's disease diagnosis and progression rates: implications for therapeutic trials. *Neurotherapeutics*, 19(1):8–25, 2023. 3, 1

[19] A. Ebrahimi, S. Luo, and R. Chiong. Introducing transfer learning to 3d resnet-18 for alzheimer's disease detection on mri images. In *2020 35th International Conference on Image and Vision Computing New Zealand (IVCNZ)*, pages 1–6. IEEE, 2020. 2, 3, 1

[20] Jinwang Feng, Shao-Wu Zhang, Luonan Chen, and Chunman Zuo. Detection of alzheimer's disease using features of brain region-of-interest-based individual network constructed with the smri image. *Computerized Medical Imaging and Graphics*, 98:102057, 2022. 2

[21] L. Ferrarini, W. M. Palm, H. Olofsen, M. A. van Buchem, J. H. Reiber, and F. Admiraal-Behloul. Shape differences of the brain ventricles in alzheimer's disease. *Neuroimage*, 32 (3):1060–1069, 2006. 6

[22] B. Fischl, A. Van Der Kouwe, C. Destrieux, E. Halgren, F. Ségonne, D.H. Salat, E. Busa, L.J. Seidman, J. Goldstein, D. Kennedy, et al. Automatically parcellating the human cerebral cortex. *Cerebral Cortex*, 14(1):11–22, 2004. 3, 1

[23] H.M. Fonteijn, M. Modat, M.J. Clarkson, J. Barnes, M. Lehmann, N.Z. Hobbs, R.I. Scahill, S.J. Tabrizi, S. Ourselin, N.C. Fox, et al. An event-based model for disease progression and its application in familial alzheimer's disease and huntington's disease. *NeuroImage*, 60(3):1880–1889, 2012. 3, 1

[24] C. Fowler et al. Fifteen years of the australian imaging, biomarkers and lifestyle (aibl) study: Progress and observations from 2,359 older adults spanning the spectrum from cognitive normality to alzheimer's disease. *J Alzheimers Dis Rep*, 5(1):443–468, 2021. 4

[25] Y. Gao, M. Zhou, and D. N. Metaxas. Utnet: a hybrid transformer architecture for medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 61–71. Springer, 2021. 1

[26] A. Goyal and Y. Bengio. Inductive biases for deep learning of higher-level cognition. *Proceedings of the Royal Society A*, 478(2266):20210068, 2022. 1

[27] K. Han et al. A survey on vision transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1): 87–110, 2023. 1

[28] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 4

[29] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. Densely connected convolutional net-

works. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4700–4708, 2017. 4

[30] Qiu hui Chen, Qiang Fu, Hao Bai, and Yi Hong. Longformer: Longitudinal transformer for alzheimer's disease classification with structural mris. *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3563–3572, 2023. 1, 2

[31] C. R. Jack et al. Tracking pathophysiological processes in alzheimer's disease: an updated hypothetical model of dynamic biomarkers. *Lancet Neurology*, 12:207–216, 2013. 3

[32] C. R. Jr Jack et al. Tracking pathophysiological processes in alzheimer's disease: an updated hypothetical model of dynamic biomarkers. *Lancet Neurology*, 12(2):207–216, 2013. 3, 1

[33] J. Jang and D. Hwang. M3t: Three-dimensional medical image classifier using multi-plane and multi-slice transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20718–20729, 2022. 2, 4, 5

[34] K.A. Jellinger. Recent update on the heterogeneity of the alzheimer's disease spectrum. *Journal of Neural Transmission*, 129(1):1–24, 2022. 3, 1

[35] Clifford R. Jack Jr et al. The alzheimer's disease neuroimaging initiative (adni): Mri methods. *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 27(4):685–691, 2008. 4

[36] U. Khatri and G. R. Kwon. Diagnosis of alzheimer's disease via optimized lightweight convolution-attention and structural mri. *Comput Biol Med*, 171:108116, 2024. 4, 5

[37] Rafsanjany Kushol, Abbas Masoumzadeh, Dong Huo, Sanjay Kalra, and Yee-Hong Yang. Addformer: Alzheimer's disease detection from structural mri using fusion transformer. In *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, pages 1–5, 2022. 2, 4, 5

[38] C. Li et al. Trans-resnet: Integrating transformers and cnns for alzheimer's disease classification. In *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, pages 1–5, Kolkata, India, 2022. 2

[39] X. Li, D. Coyle, L. Maguire, D. R. Watson, and T. M. McGinnity. Gray matter concentration and effective connectivity changes in alzheimer's disease: a longitudinal structural mri study. *Neuroradiology*, 53:733–748, 2011. 3, 1

[40] G. M. McKhann et al. The diagnosis of dementia due to alzheimer's disease: recommendations from the national institute on aging-alzheimer's association workgroups on diagnostic guidelines for alzheimer's disease. *Alzheimer's & Dementia*, 7(3):263–269, 2011. 1, 3

[41] Sean M. Nestor et al. Ventricular enlargement as a possible measure of alzheimer's disease progression validated using the alzheimer's disease neuroimaging initiative database. *Brain*, 131:2443–2454, 2008. 3, 1

[42] C. Pennanen et al. Hippocampus and entorhinal cortex in mild cognitive impairment and early ad. *Neurobiology of Aging*, 25:303–310, 2004. 3, 1

[43] M. Perslev, E. B. Dam, A. Pai, and C. Igel. One network to segment them all: A general, lightweight system for accurate 3d medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 30–38. Springer, 2019. 2

[44] V. Planche, J. V. Manjon, B. Mansencal, E. Lanuza, T. Tourdias, G. Catheline, and P. Coupé. Structural progression of alzheimer's disease over decades: the mri staging scheme. *Brain Communications*, 4(3), 2022. 1, 3, 5, 2

[45] Adhish Prasoon, Kersten Petersen, Christian Igel, François Lauze, Erik Dam, and Mads Nielsen. Deep feature learning for knee cartilage segmentation using a triplanar convolutional neural network. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 246–253. Springer, 2013. 2

[46] A. Qiu, L. Xu, C. Liu, and Alzheimer's Disease Neuroimaging Initiative. Predicting diagnosis 4 years prior to alzheimer's disease incident. *NeuroImage: Clinical*, 34: 102993, 2022. 1

[47] Edmund T. Rolls, Chu-Chung Huang, Ching-Po Lin, Jianfeng Feng, and Marc Joliot. Automated anatomical labelling atlas 3. 2024. 2

[48] Holger R. Roth et al. A new 2.5 d representation for lymph node detection using random sets of deep convolutional neural network observations. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 520–527. Springer, 2014. 2

[49] William W. Seeley, Richard K. Crawford, Juan Zhou, Bruce L. Miller, and Michael D. Greicius. Neurodegenerative diseases target large-scale human brain networks. *Neuron*, 62(1):42–52, 2009. 8, 1

[50] Siddharth Shah, Hadeel Mansour, Tania Aguilar, and Brandon Lucke-Wold. Mesenchymal stem cell-derived exosomes as a neuroregeneration treatment for alzheimer's disease. *Biomedicines*, 12:2113, 2024. 5

[51] C. Tinauer, S. Heber, L. Pirpamer, et al. Interpretable brain disease classification and relevance-guided deep learning. *Scientific Reports*, 12:20254, 2022. 2

[52] Christian Tinauer, Maximilian Sackl, Rudolf Stollberger, Stefan Ropele, and Christian Langkammer. Pfungst and clever hans: Identifying the unintended cues in a widely used alzheimer's disease mri dataset using explainable deep learning, 2025. 8

[53] Di Wang et al. Heatmaps autoencoders robustly capture alzheimer's disease's brain alterations. *Glenn Biggs Institute for Neurodegenerative Disorders and Research Imaging Institute, University of Texas Health Science Center at San Antonio*, 2023. 2

[54] Zhuofan Xia, Xuran Pan, Shiji Song, Li Erran Li, and Gao Huang. Vision transformer with deformable attention. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1

[55] M.-Y. Xu, Z.-Q. Xu, and Y.-J. Wang. White matter "matters" in alzheimer's disease. *Neuroscience Bulletin*, 38:323–326, 2022. 1

[56] Q. Zhao et al. Ida-net: Inheritable deformable attention network of structural mri for alzheimer's disease diagnosis. *Biomedical Signal Processing and Control*, 84:104787, 2023. 1, 2, 3