Multi-ROI Multimodal 3D Vision Transformer for Alzheimer's Disease Classification with Attention-Based Interpretability

Juan A. Castro-Silva *, María N. Moreno-García, Diego H. Peluffo-Ordóñez

Posted Date: 13 May 2026

doi: 10.20944/preprints202605.0910.v1

Keywords: Alzheimer's disease; 3D vision transformer; multimodal learning; multi-ROI decomposition; magnetic resonance imaging (MRI); attention mechanisms; explainable artificial intelligence (XAI); clinical data integration; volumetric biomarkers; deep learning

 doi:10.20944/preprints202605.0910.v1

Article

Multi-ROI Multimodal 3D Vision Transformer for Alzheimer's Disease Classification with Attention-Based Interpretability

Juan A. Castro-Silva [1,2,*], María N. Moreno-García [1] and Diego H. Peluffo-Ordóñez [3,4]

[1] Universidad de Salamanca, Salamanca, Spain
[2] Universidad Surcolombiana, Neiva, Colombia
[3] Corporación Universitaria Autónoma de Nariño, Faculty of Engineering, Pasto 520002, Colombia
[4] Universidad ECOTEC, Research Department, Samborondón 092302, Ecuador
* Correspondence: juan.castro@usco.edu.co

Abstract

Alzheimer's disease (AD) is a progressive neurodegenerative disorder for which early and accurate diagnosis remains a critical challenge. In this work, we propose a Multi-ROI Multimodal 3D Vision Transformer for AD classification that integrates structural MRI data with clinical and volumetric biomarkers within a unified attention-based framework. The proposed approach leverages anatomically guided multi-region-of-interest (ROI) decomposition to focus on disease-relevant brain structures, including the hippocampus, entorhinal cortex, fornix, and major cortical lobes. Each ROI is encoded using 3D tubelet embeddings, while clinical and volumetric features are transformed into feature-wise tokens, enabling seamless multimodal fusion through self-attention mechanisms. A hemisphere-aware selection strategy is introduced to identify the most discriminative ROI representations, enhancing both performance and interpretability. The model is evaluated on a merged multi-cohort dataset combining ADNI, AIBL, and OASIS, using a 7-fold cross-validation protocol. Experimental results demonstrate that the proposed method achieves high classification performance, reaching an accuracy of 97.62% and an AUC of 0.9940, outperforming single-modality and whole-brain baselines. Furthermore, attention-based analysis provides interpretable insights into the relative importance of clinical and neuroanatomical features, revealing consistency with established AD biomarkers. These findings highlight the effectiveness of multimodal integration and ROI-based representation for robust and explainable AD classification.

Keywords: Alzheimer's disease; 3D vision transformer; multimodal learning; multi-ROI decomposition; magnetic resonance imaging (MRI); attention mechanisms; explainable artificial intelligence (XAI); clinical data integration; volumetric biomarkers; deep learning

1. Introduction

Neurodegenerative disorders represent a growing global healthcare challenge as aging populations continue to increase. Among these conditions, Alzheimer's disease (AD) is the most common cause of dementia, affecting more than 55 million people worldwide [1]. AD is characterized by progressive neuronal degeneration, brain atrophy, and the accumulation of pathological proteins such as amyloid-β and tau, which lead to severe cognitive decline and functional impairment. Neurodegeneration particularly affects brain regions associated with memory and cognition, including the entorhinal cortex, hippocampus, fornix, and the frontal, temporal, and parietal lobes [2–4]. Because neuropathological changes can begin years before clinical symptoms appear, early and accurate diagnosis is crucial for enabling timely interventions and improving disease management [5,6].

AD can be characterized using multiple biomarkers, including neuroimaging, cerebrospinal fluid (CSF), genetic, and blood-based markers. Among these, neuroimaging biomarkers play a crucial role

 doi:10.20944/preprints202605.0910.v1

in detecting structural and functional brain alterations associated with disease progression. Based on the biophysical characteristics of AD pathology, neuroimaging techniques are commonly categorized into structural, functional, and molecular imaging modalities [7]. Structural imaging methods such as magnetic resonance imaging (MRI) detect anatomical alterations including hippocampal atrophy, ventricular enlargement, and overall brain volume loss. Functional imaging techniques, such as functional MRI (fMRI), evaluate neuronal activity through hemodynamic responses, while molecular imaging approaches such as single-photon emission computed tomography (SPECT) reveal biochemical changes related to neurodegeneration.

Among these techniques, MRI is one of the most widely used non-invasive neuroimaging modalities for evaluating AD. MRI-derived biomarkers—including hippocampal and medial temporal lobe atrophy, ventricular enlargement, and cortical thinning—are among the most validated imaging indicators of AD progression. These structural changes enable the detection of neurodegeneration and support the differentiation of AD from other forms of dementia [8]. In clinical practice, AD diagnosis typically involves a combination of neurological examinations, cognitive assessments such as the Mini-Mental State Examination (MMSE), and neuroimaging techniques including MRI and computed tomography (CT) [3,9]. To support research in AD diagnosis, large-scale datasets such as the Alzheimer's Disease Neuroimaging Initiative (ADNI) [10], Australian Imaging, Biomarker and Lifestyle Flagship Study of Ageing (AIBL) [11], and Open Access Series of Imaging Studies (OASIS) [12] provide multimodal information including imaging, demographic, biological, and clinical data.

Traditional approaches for detecting brain abnormalities in AD can be broadly categorized into several groups, including voxel-based morphometry (VBM), statistical atlas-based methods, functional and connectivity-based analyses, and Gaussian Hidden Markov Model (GHMM)-based approaches. VBM techniques analyze structural MRI data to identify regional gray matter loss, particularly in areas such as the hippocampus and surrounding cortical regions. Atlas-based approaches compare patient scans with normative templates to identify abnormal patterns, while connectivity-based methods analyze disruptions in brain networks associated with AD. GHMM-based methods further model spatial dependencies in brain images to detect localized pathological changes [13]. Although these approaches have demonstrated promising results, they often rely on handcrafted features or predefined statistical models that may limit their ability to capture complex patterns in high-dimensional neuroimaging data.

Recent advances in deep learning have significantly improved automated AD classification using neuroimaging data.

Deep learning–based computer vision methods can detect structural brain changes from MRI using different input representations, including 2D slices, region-of-interest (ROI)–based inputs, 3D patches, and full 3D subject-level volumes [14]. While ROI-based approaches reduce computational cost and the risk of overfitting, they may overlook disease-related patterns distributed across multiple brain regions.

However, most existing approaches either rely on 2D slice-based representations, which fail to preserve volumetric context, or focus on isolated ROIs, limiting their ability to capture distributed neurodegenerative patterns.

Transformer architectures, originally introduced by Ashish Vaswani in the Transformer model [15], have recently shown strong performance in computer vision tasks. In particular, the Vision Transformer (ViT) architecture [16] leverages self-attention mechanisms to capture long-range dependencies within image data. Compared with 2D transformer models that process slices independently, 3D transformer architectures can better preserve spatial context by jointly modeling volumetric information across consecutive slices [17,18].

AD datasets often contain heterogeneous information, including demographic variables (e.g., age and sex), clinical scores such as MMSE, and MRI scans. Effectively integrating these multimodal data sources remains a major challenge for automated AD classification, as many existing models rely

 doi:10.20944/preprints202605.0910.v1

on single modalities or 2D representations of 3D volumes, potentially missing important spatial and contextual information [19].

Unlike existing approaches that rely on either full-volume or single-ROI analysis, this work introduces a multi-atlas ROI-based instance selection strategy combined with a multimodal 3D transformer framework, enabling efficient and anatomically informed learning from volumetric MRI data.

To address these limitations, this study proposes a multiple-input multimodal framework based on a 3D Vision Transformer that integrates categorical, numerical, and volumetric MRI data together with 3D region-of-interest (ROI) images from key brain regions, including the entorhinal cortex, fornix, hippocampus, and major cortical lobes. By leveraging transformer-based self-attention and volumetric imaging information, the proposed approach aims to improve robustness and diagnostic accuracy for AD classification.

The main contributions of this study are centered on the design of a novel multimodal transformer-based framework for Alzheimer's disease classification, and can be summarized as follows:

1. **A unified multi-ROI multimodal transformer architecture for Alzheimer's disease classification.** We propose a multiple-input deep learning framework that jointly models heterogeneous data sources, including 3D MRI-derived regions of interest (ROIs), clinical metadata, and volumetric biomarkers. The architecture enables end-to-end learning across modalities, capturing complementary information relevant to Alzheimer's disease progression.
2. **Multi-ROI tokenization of volumetric MRI using 3D tubelet embeddings.** Instead of processing whole-brain volumes, the proposed model decomposes MRI data into multiple anatomically relevant ROIs (e.g., hippocampus, entorhinal cortex, fornix, and cortical lobes). Each ROI is independently encoded using 3D tubelet embeddings, allowing fine-grained spatial representation learning while reducing irrelevant background information.
3. **Feature-wise tokenization of clinical and volumetric data for transformer-based fusion.** We introduce a feature tokenization strategy that transforms tabular clinical variables and volumetric biomarkers into learnable token representations. This design enables seamless integration of structured data into the transformer architecture, facilitating cross-modal attention between imaging and non-imaging features.
4. **Modality-aware embedding for explicit cross-modal representation learning.** The model incorporates learnable modality embeddings to distinguish between ROI-specific imaging tokens and non-imaging features. This mechanism enhances the model's ability to learn modality-specific and cross-modal interactions within a unified attention framework.
5. **Attention-based multimodal fusion with dual representation learning.** A hybrid representation is obtained by combining a global *[CLS]* token with learnable attention pooling over all tokens. This dual aggregation strategy improves information integration across modalities and enhances classification robustness.
6. **Interpretable attention mechanisms for clinical insight extraction.** The architecture provides access to attention maps across transformer layers and modalities, enabling analysis of region relevance, feature importance, and cross-modal interactions. This contributes to model interpretability and supports clinically meaningful insights into Alzheimer's disease biomarkers.

The structure of the paper is outlined as follows: Section 2 presents some related works. The materials and methods used for preprocessing and building AD transformer-based classification models are included in Section 3. Section 4 provides a detailed description of the experiments conducted in this work and the parameter settings used. The results of the experiments are discussed in Section 5. Section 6 addresses the limitations of the proposed methods. Finally, the concluding remarks of this work are summarized in Section 8.

2. Related Work

Neuroimaging preprocessing plays a pivotal role in Alzheimer's disease (AD) classification, as raw MRI and PET scans frequently contain noise, intensity inhomogeneities, and anatomical variability.

 doi:10.20944/preprints202605.0910.v1

To address these issues, previous studies have implemented standardized pipelines that include skull stripping, spatial normalization, bias field correction, and intensity scaling to maintain consistent image quality across subjects [20–22]. These pipelines facilitate accurate tissue segmentation of gray matter, white matter, and cerebrospinal fluid. Advanced methods further incorporate denoising, contrast enhancement, and multimodal alignment to enhance feature extraction from critical regions such as the hippocampus and cortical structures [23–25]. Surface-based frameworks, such as FreeSurfer, are also commonly used to extract cortical morphometric features—including volume, thickness, curvature, folding, and surface area—from anatomically defined regions (e.g., DKT atlas), which have demonstrated effectiveness for machine learning and deep learning-based AD classification [26].

Recent research demonstrates that combining volumetric alterations of key neuroanatomical structures, particularly the hippocampus, amygdala, and ventricular system, yields highly discriminative biomarkers for Alzheimer's disease (AD), achieving performance comparable to deep learning methods [27]. Hippocampal volumetry, in particular, is recognized as one of the most robust biomarkers and is frequently computed bilaterally to enhance stability and discriminative capacity [27]. In practical applications, hippocampal and amygdalar volumes, typically normalized by estimated total intracranial volume (eTIV), are widely utilized, while ventricular enlargement serves as an additional marker of global atrophy [28,29]. These findings collectively support the adoption of compact, biologically informed volumetric features as effective representations for AD classification.

Initial efforts in automated AD diagnosis frequently utilized region-of-interest (ROI)–based models that target specific brain regions affected by the disease. For instance, [30] extracts hippocampal blocks from MRI scans, whereas [31] examines medial temporal lobe structures using coronal slices. Other methodologies develop ensemble classifiers by extracting patches from multiple regions, including the hippocampus, amygdala, and insulae [32–34]. Further studies implement ROI-based frameworks that incorporate anatomical landmarks [35], employ explainable 3D convolutional neural networks for patient-specific ROI detection [36], or utilize statistical techniques to identify informative ROI content [37]. Although ROI-based approaches reduce computational complexity and emphasize disease-relevant structures, they may fail to capture global structural patterns distributed throughout the brain.

Deep learning techniques have substantially advanced AD classification using neuroimaging data. Convolutional neural networks (CNNs) are widely adopted for their capacity to automatically learn hierarchical representations from MRI scans. However, CNNs predominantly capture local spatial features and often struggle to model long-range relationships in volumetric brain images. To overcome this limitation, recent research has increasingly investigated transformer-based architectures that model global dependencies via self-attention mechanisms.

The Vision Transformer (ViT) has garnered significant attention in AD classification due to its capacity to capture long-range interactions between image patches. Multiple studies utilize ViT models with transfer learning to address the scarcity of labeled neuroimaging data and to enhance disease staging and progression prediction [38,39]. Hybrid CNN–Transformer architectures have also been introduced to integrate local feature extraction with global contextual modeling for AD diagnosis [40–43]. Furthermore, transformer architectures adapted for volumetric MRI data improve the analysis of three-dimensional brain structures and facilitate the capture of spatial dependencies across slices [44].

The Swin Transformer represents another notable architecture, introducing hierarchical feature representations and shifted-window self-attention to efficiently process high-resolution images. Recent studies combine Swin Transformers with CNN backbones to enhance feature extraction and facilitate early AD detection [5,20,22,45]. Additional advancements incorporate frequency-domain features and specialized attention mechanisms to improve classification accuracy and lesion localization [46–48]. Collectively, these transformer-based approaches demonstrate enhanced capability to capture both local and global contextual information in neuroimaging data.

 doi:10.20944/preprints202605.0910.v1

Recent research has explored multimodal deep learning approaches to address the multifactorial nature of AD by integrating diverse data sources, such as neuroimaging, clinical information, and genetic data. For instance, [49] analyzed omics, imaging, and clinical features from the ANMerge dataset [50], demonstrating improved performance when combining imaging and omics data. Similarly, [51] integrated structural MRI, SNP-based genetic profiles, and electronic health records using stacked denoising autoencoders and 3D CNNs, achieving higher diagnostic accuracy than single-modality approaches. Additional studies address the issue of incomplete modality availability. For example, [25] proposed a multi-input 3D CNN to manage missing MRI or PET data, while [52] generated missing modalities using a generative adversarial network prior to applying a multimodal transformer for classification.

Attention-based multimodal fusion strategies have been investigated to model interactions across different modalities. The MADDi framework [53] utilizes cross-modal attention to jointly analyze imaging, genetic, and clinical data, whereas [54] combines MRI and PET hippocampal features using dual-branch CNN architectures. Early-fusion strategies have also been introduced, such as the modified ResNet architecture in [55], which integrates MRI and PET inputs and incorporates explainable artificial intelligence techniques to enhance interpretability.

Despite substantial advancements, several limitations persist in current methodologies. Many early approaches rely on handcrafted features or predefined statistical models, limiting their ability to capture complex, high-dimensional patterns in neuroimaging data. Although convolutional neural networks (CNNs) have improved representation learning, they primarily focus on local spatial features and often fail to model long-range dependencies in volumetric brain images.

Recent transformer-based models mitigate this limitation by employing self-attention mechanisms to capture global contextual relationships. Nevertheless, many of these approaches continue to operate on 2D slice-based representations, which do not fully preserve the three-dimensional anatomical structure of the brain. Although some studies extend transformers to 3D data, these often depend on full-volume inputs, resulting in high computational costs and the potential inclusion of redundant or non-informative regions.

ROI-based methods provide a more efficient alternative by concentrating on anatomically relevant regions. However, many current ROI-based approaches analyze isolated structures independently, which limits their ability to capture distributed neurodegenerative patterns across multiple brain regions. Additionally, several studies utilize single-modality inputs, thereby restricting their capacity to model the multifactorial nature of Alzheimer's disease.

Moreover, multimodal approaches have demonstrated enhanced performance by integrating imaging, clinical, and genetic data. However, these methods frequently lack effective strategies for selecting informative regions or for fully leveraging volumetric spatial relationships within MRI data.

To address these limitations, the present study introduces a multimodal 3D Vision Transformer framework that integrates multi-atlas ROI-based instance selection with heterogeneous clinical and imaging data. By combining anatomically informed ROI extraction with transformer-based global context modeling, this approach facilitates efficient and comprehensive learning of distributed neurodegenerative patterns, thereby enhancing robustness and classification performance.

These gaps underscore the necessity for approaches that concurrently address volumetric representation, efficient ROI selection, and multimodal data integration within a unified learning framework.

3. Materials and Methods

This section presents the datasets and the proposed methodology for building AD classification models. It covers data preparation, instance selection, ROI extraction, 3D ROI batch generation, and the transformer classification model used in this work.

3.1. Datasets

This study utilizes three publicly available and widely adopted neuroimaging datasets: the Alzheimer's Disease Neuroimaging Initiative (ADNI) [10], the Australian Imaging, Biomarker and

 doi:10.20944/preprints202605.0910.v1

Lifestyle Flagship Study of Ageing (AIBL) [11], and the Open Access Series of Imaging Studies (OASIS) [12]. These datasets provide complementary clinical and neuroimaging information and are widely used benchmarks for Alzheimer's disease classification.

Multimodal Data Representation

Each subject is represented using multiple data modalities aligned with the proposed multimodal transformer framework:

- **3D MRI (ROI-based):** Structural T1-weighted brain volumes are processed to extract anatomically relevant regions of interest (ROIs), including the hippocampus, entorhinal cortex, fornix, and major cortical lobes. These regions are strongly associated with AD-related neurodegeneration.
- **Clinical and Demographic Data:** Subject-level attributes such as age, and sex, together with cognitive scores (e.g., MMSE), are included to capture inter-subject variability.
- **Volumetric Biomarkers:** Quantitative volumetric measures derived from neuroanatomical structures (e.g., hippocampus, amygdala, ventricles) are incorporated as structured features.

Diagnostic Labels

Subjects are labeled using the Clinical Dementia Rating (CDR) scale [56]. In this study, a binary classification setting is adopted, where subjects with CDR = 0 are considered cognitively normal (CN), and subjects with CDR ≥ 1 are classified as Alzheimer's disease (AD). Subjects with CDR = 0.5 (mild cognitive impairment) are excluded to ensure clear class separation.

Cohort Construction

To ensure balanced class distribution and reduce bias, a stratified sampling strategy is applied across datasets. The final cohort consists of 420 subjects (one MRI scan per subject), with 70 subjects per class (CN and AD) selected from each dataset (3 datasets $\times$ 2 classes $\times$ 70 subjects). This balanced design mitigates class imbalance and facilitates robust model evaluation across heterogeneous populations.

Demographic characteristics of the resulting cohort are summarized in Table 1.

Table 1. Summary of participant demographics and global clinical dementia rating (CDR) scores of all the study datasets.

Dataset	Class	Subjects	Age	Sex F / M	Total Subjects
ADNI	CN	70	77.91 ± 6.05	31/39	140
	AD	70	78.70 ± 6.35	33/37	
AIBL	CN	70	73.71 ± 6.15	36/34	140
	AD	70	75.10 ± 7.83	33/37	
OASIS	CN	70	69.44 ± 9.33	39/31	140
	AD	70	75.67 ± 9.55	33/37	
MERGED	CN	210	73.69 ± 8.08	106/104	420
	AD	210	76.49 ± 8.13	107/103	

Data Harmonization

To reduce inter-dataset variability, all MRI volumes are spatially normalized to a common template space and intensity-normalized. Additionally, one scan per subject is retained to avoid data leakage and ensure independence between samples.

 doi:10.20944/preprints202605.0910.v1

3.2. *Proposed Methodology*

This methodology aims to develop a multimodal transformer-based classification framework that integrates image information from multiple regions of interest (ROIs) with heterogeneous clinical and volumetric data to improve classification performance, robustness, and reliability.

As illustrated in Figure 1, the proposed methodology begins with a data preparation stage, which includes partitioning the dataset into training, validation, and test subsets, followed by image preprocessing to remove non-informative regions and ensure spatial alignment. Subsequently, the instance selection procedure identifies slices containing the predefined ROIs and estimates the corresponding centroid coordinates (x, y) using the statistical mode. Finally, the proposed framework is trained and evaluated using transformer-based classification models that integrate ROI-based image representations with mixed data modalities within a unified multimodal

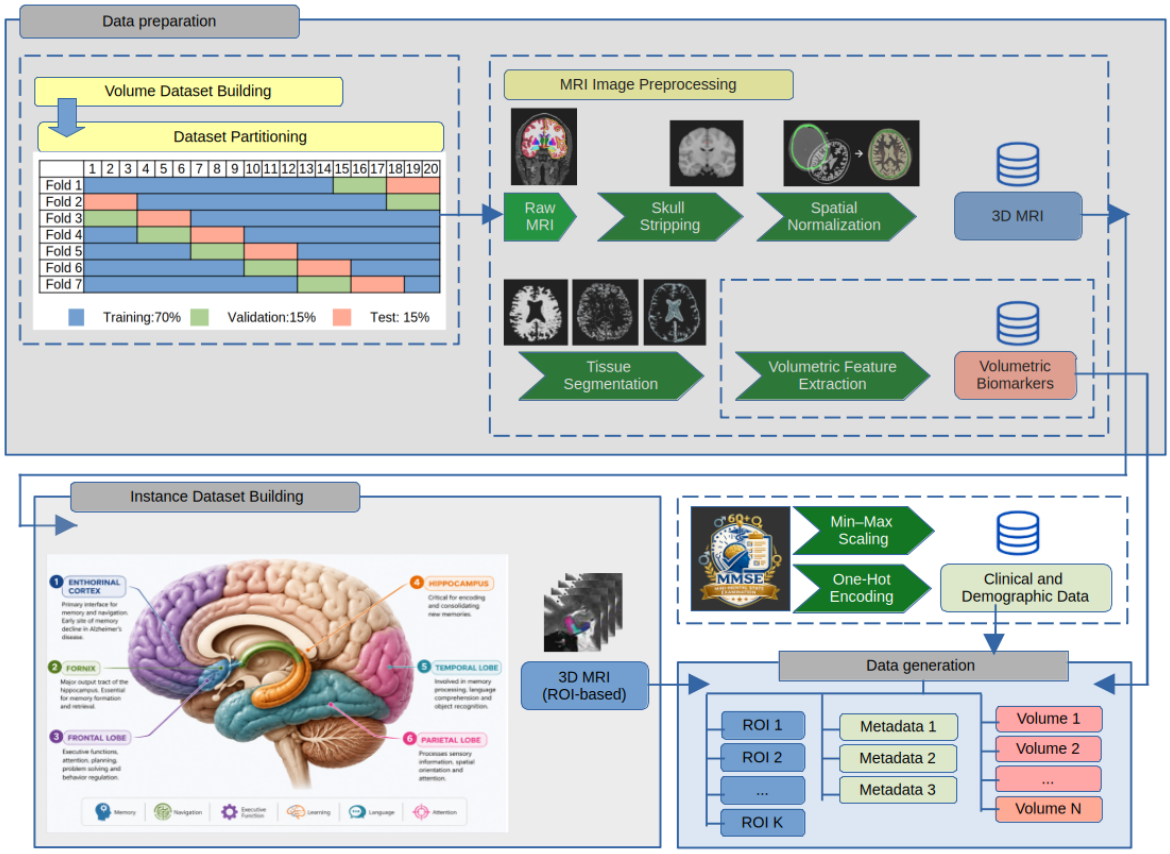

Figure 1. Overview of the proposed multimodal transformer-based methodology for Alzheimer's disease classification. The pipeline includes MRI preprocessing, ROI-based instance selection with centroid estimation (x, y), and multimodal transformer-based classification integrating multiple ROI image representations with clinical and volumetric data.

3.2.1. Dataset Preparation

This phase involves the selection of a single representative MRI volume per subject, followed by the partitioning of the dataset into training, validation, and test subsets. The selected volumes are subsequently subjected to preprocessing steps, including skull stripping, tissue segmentation, spatial normalization (registration), and volumetric feature extraction, to ensure anatomical consistency and facilitate reliable quantitative analysis across subjects.

Volume Dataset Building:

For each subject, MRI volumes are chronologically ordered according to visit date, and only the most recent acquisition is retained to ensure a single representative volume per subject. The resulting dataset is then randomly partitioned into training, validation, and test subsets in a subject-wise

 doi:10.20944/preprints202605.0910.v1

manner to guarantee reproducibility and prevent data leakage, ensuring that each subject contributes exclusively to one subset. To address class imbalance, an undersampling strategy is applied, selecting an equal number of subjects per class (k), where k is defined as the size of the minority class or a lower predefined threshold.

Data Preprocessing

To ensure data consistency and enhance model performance, a series of preprocessing steps were applied across numerical, categorical, and imaging modalities. These steps were designed to standardize feature representations, reduce inter-subject variability, and preserve anatomical fidelity.

- **Numerical Data:** Continuous variables were normalized using min–max scaling to the range $[0, 1]$, ensuring comparable feature magnitudes and stable optimization during training.
- **Categorical Data:** Categorical variables were encoded using one-hot encoding, producing binary vectors within the range $[0, 1]$ and avoiding the introduction of ordinal relationships.
- **Image Intensity Scaling:** MRI voxel intensities were normalized to the range $[0, 1]$ to improve numerical stability and convergence of deep learning models.

MRI Image Preprocessing

T1-weighted structural MRI scans were subjected to a standardized preprocessing pipeline to ensure anatomical consistency across subjects and datasets (ADNI, AIBL, and OASIS). The main steps are described as follows:

- **Skull Stripping:** Raw MRI volumes were processed to remove non-brain tissues—including skin, fat, muscle, neck, and ocular structures—thereby isolating the intracranial region of interest.
- **Tissue Segmentation and Surface Reconstruction:** The brain was segmented into major tissue classes, including gray matter (GM), white matter (WM), cerebrospinal fluid (CSF), and background. Subsequently, the white matter and pial surfaces were reconstructed, enabling accurate modeling of cortical boundaries [26].
- **Spatial Normalization (Registration):** Skull-stripped volumes were nonlinearly registered to the MNI152 T1-weighted template, ensuring uniformity in anatomical orientation, shape, and alignment. The resulting volumes were resampled to a standardized resolution of $1\,\mathrm{mm}^3$ and dimensions of $182 \times 218 \times 182$ voxels.
- **Volumetric Feature Extraction:** Region-of-interest (ROI) volumetric measures were computed, focusing on structures strongly associated with Alzheimer's disease. These include the left and right hippocampus, amygdala, and lateral ventricles (including inferior lateral ventricles), which are often combined into bilateral measures to improve robustness and discriminative power [27,57].

All preprocessing steps, including normalization and ROI extraction, were performed independently within each training fold to prevent information leakage into validation and test sets.

3.2.2. Instance Dataset Building

Instance selection techniques play a crucial role in optimizing predictive models by prioritizing informative data, improving performance, reducing computational cost, and limiting dataset size. In this study, we adopt a novel instance selection framework proposed in [37], which comprises two complementary components. This strategy addresses a key limitation in neuroimaging pipelines, where redundant or non-informative slices may degrade model performance and increase computational cost. First, a multi-atlas ROI-based instance selection strategy integrates annotations from multiple atlases to retain the most informative and representative slices. Second, a ROI content extraction method employs the statistical mode to refine the centroid (x, y) location, enabling precise extraction of relevant anatomical content for accurate 2D slice cropping.

 doi:10.20944/preprints202605.0910.v1

Algorithm 1 ROI-Based Instance Selection with Centroid Refinement

Require: MRI volume $\mathcal{X}$; atlas set $\{\mathcal{A}_1, \ldots, \mathcal{A}_M\}$; ROI label r; threshold τ
Ensure: Selected slices $\mathcal{S}$ and refined centroid (x_c, y_c)

1: **for** $m = 1$ to M **do**
2: Register atlas $\mathcal{A}_m$ to the MRI space of $\mathcal{X}$
3: Extract binary ROI mask $\mathcal{M}_m^r$ from atlas $\mathcal{A}_m$
4: **end for**
5: Aggregate ROI masks using voxel-wise majority voting:

$$\mathcal{M}^r(i,j,s) = \text{mode}(\mathcal{M}_1^r(i,j,s), \ldots, \mathcal{M}_M^r(i,j,s))$$

6: Identify informative slices:

$$\mathcal{S} = \left\{ s \;\middle|\; \sum_{i,j} \mathcal{M}^r(i,j,s) > \tau \right\}$$

7: Compute slice-wise ROI centroids:

$$(x_s, y_s) = \left(\frac{\sum_{i,j} i\, \mathcal{M}^r(i,j,s)}{\sum_{i,j} \mathcal{M}^r(i,j,s)}, \frac{\sum_{i,j} j\, \mathcal{M}^r(i,j,s)}{\sum_{i,j} \mathcal{M}^r(i,j,s)} \right), \quad s \in \mathcal{S}$$

8: Refine centroid using the statistical mode:

$$(x_c, y_c) = (\text{mode}\{x_s \mid s \in \mathcal{S}\}, \text{mode}\{y_s \mid s \in \mathcal{S}\})$$

return $\mathcal{S}, (x_c, y_c)$

3.2.3. Data Generation

Data generation is based on instance-level metadata, including age, MMSE score, sex, volume filename, slice index, ROI centroid coordinates, and class labels. To address memory constraints, data are processed in batches during training.

All preprocessing steps were applied as described in Subsection 3.2.1. Finally, 3D regions of interest were extracted by cropping MRI volumes around the ROI centroid coordinates, ensuring consistent spatial localization of the input data.

3.2.4. Proposed Multimodal Transformer Architecture

The proposed model adopts a unified token-based transformer architecture to integrate heterogeneous modalities, including 3D MRI-derived ROIs, clinical variables, and volumetric biomarkers.

Multi-ROI MRI Encoding

Each region of interest (ROI) is processed independently using a 3D tubelet embedding layer, which partitions volumetric inputs into non-overlapping patches and projects them into a shared embedding space. This design enables fine-grained anatomical representation while preserving volumetric context.

Tabular Feature Encoding

Clinical and demographic variables are transformed into feature-wise token representations using learnable embeddings. This approach allows structured data to be seamlessly integrated into the transformer architecture.

Multimodal Fusion via Self-Attention

All tokens are concatenated and augmented with modality embeddings and positional encodings. A learnable classification token is appended to capture global context. The resulting sequence is processed using multiple transformer encoder blocks, enabling joint modeling of intra- and inter-modality relationships.

 doi:10.20944/preprints202605.0910.v1

Hybrid Representation Learning

The final representation is obtained by combining the classification token with attention-based pooling over all tokens. This dual aggregation mechanism enhances robustness by capturing both global and distributed patterns.

Classification Head

The aggregated representation is passed through fully connected layers with nonlinear activations and regularization, followed by a softmax layer for diagnostic classification.

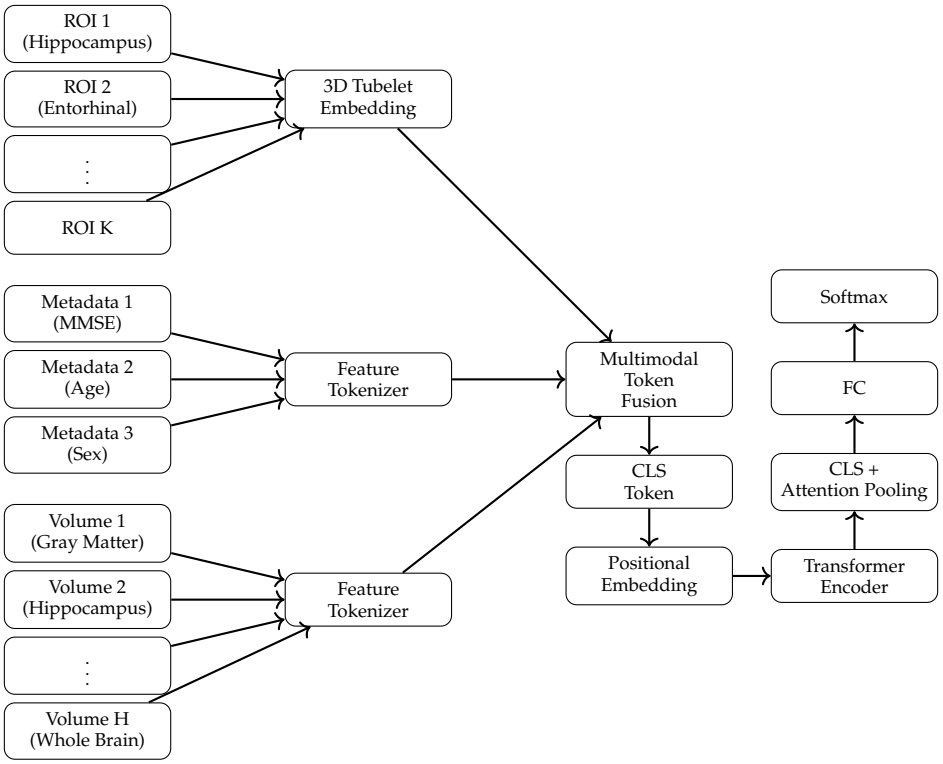

Figure 2. Compact representation of the proposed multimodal transformer architecture.

3.2.5. Mathematical Formulation of the Multimodal Transformer

Let $\mathcal{X}_k^{roi} \in \mathbb{R}^{D\times H\times W}$ denote the k-th 3D region of interest (ROI), where $k = 1, \ldots, K$.

ROI Tokenization

Each ROI is partitioned into non-overlapping 3D tubelets of size (p_d, p_h, p_w) and projected into an embedding space:

$$\mathbf{Z}_k^{roi} = \text{Flatten}\Big(\text{Conv3D}(\mathcal{X}_k^{roi})\Big) \in \mathbb{R}^{N_k\times d} \tag{1}$$

where N_k is the number of tokens and d is the embedding dimension.

Tabular Tokenization

Let $\mathbf{x}^{tab} \in \mathbb{R}^F$ denote tabular features (clinical and volumetric). Feature-wise tokenization is defined as:

$$\mathbf{Z}_i^{tab} = x_i \cdot \mathbf{w}_i + \mathbf{b}_i, \quad i = 1, \ldots, F \tag{2}$$

where $\mathbf{w}_i, \mathbf{b}_i \in \mathbb{R}^d$ are learnable parameters.

Multimodal Token Fusion

All tokens are concatenated:

 doi:10.20944/preprints202605.0910.v1

$$\mathbf{Z} = [\mathbf{Z}_1^{roi}, \ldots, \mathbf{Z}_K^{roi}, \mathbf{Z}^{tab}] \tag{3}$$

A learnable classification token $\mathbf{z}_{cls}$ is prepended:

$$\mathbf{Z}_0 = [\mathbf{z}_{cls}, \mathbf{Z}] \tag{4}$$

Transformer Encoding

The sequence is processed by L transformer layers:

$$\mathbf{Z}^{(l+1)} = \text{TransformerBlock}(\mathbf{Z}^{(l)}) \tag{5}$$

Each block applies multi-head self-attention:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}}\right)\mathbf{V} \tag{6}$$

Feature Aggregation

Final representation combines CLS token and attention pooling:

$$\mathbf{h} = [\mathbf{z}_{cls}^{(L)} \parallel \sum_i \alpha_i \mathbf{z}_i^{(L)}] \tag{7}$$

Classification

$$\hat{y} = \text{softmax}(\mathbf{W}\mathbf{h} + \mathbf{b}) \tag{8}$$

Training and Optimization:

The model is trained end-to-end using a cross-entropy loss function. Hyperparameters—including the number of transformer layers, embedding dimension, number of attention heads, and learning rate—are systematically optimized to enhance model performance. In particular, this study employs the Hyperband algorithm [58], an efficient hyperparameter optimization strategy that extends random search by dynamically allocating computational resources and applying early stopping to poorly performing configurations. This approach enables effective exploration of the hyperparameter space while reducing computational cost. Additionally, regularization techniques such as dropout and early stopping are incorporated during training to mitigate overfitting and improve generalization.

Model performance evaluation:

The Receiver Operating Characteristic (ROC) curve evaluates the trade-off between the True Positive Rate (TPR) and the False Positive Rate (FPR) across varying classification thresholds. The TPR (sensitivity) measures the proportion of correctly identified positive cases, while the FPR represents the proportion of negative cases incorrectly classified as positive.

Given the limited dataset size, model performance is evaluated using stratified 7-fold cross-validation to preserve the class distribution across training and test folds. This strategy improves robustness by training and evaluating the model on multiple data partitions while maintaining class balance, thereby reducing overfitting and providing a more reliable estimate of the model's generalization performance. Additionally, subject-level separation is enforced across folds to prevent data leakage and ensure an unbiased evaluation protocol.

The overall performance of the multi-class classification model is evaluated using the Area Under the Receiver Operating Characteristic Curve (AUC), computed under a one-vs-rest (OvR) scheme to assess class-wise discriminative ability. Additionally, classification accuracy is reported as the mean

 doi:10.20944/preprints202605.0910.v1

and standard deviation across the seven cross-validation folds, providing a robust estimate of model performance.

4. Experimental Setup

This section presents the experimental framework for evaluating the proposed multimodal 3D Vision Transformer model for Alzheimer's disease (AD) classification. It describes the datasets and MRI-derived regions of interest (ROIs) used in the experiments, along with the preprocessing procedures applied to ensure spatial consistency and reproducibility. The section also outlines the experimental scenarios designed to analyze the impact of mixed data modalities, the hyperparameter optimization strategy, and the statistical methods used to assess model performance. Finally, implementation details and computational settings are provided to facilitate reproducibility and enable fair comparisons with state-of-the-art approaches.

4.1. Datasets

The experiments were conducted on the multimodal dataset described in Section 3, which integrates data from the ADNI, AIBL, and OASIS cohorts. The dataset includes structural MRI, clinical metadata, and volumetric biomarkers, enabling comprehensive multimodal analysis.

All experiments follow the cohort construction and preprocessing procedures detailed in Section 3. In particular, a balanced dataset of 420 subjects is used, with equal representation of cognitively normal (CN) and Alzheimer's disease (AD) cases across the three datasets.

To ensure fair evaluation and prevent data leakage, all experimental splits are performed at the subject level, with each subject contributing a single MRI scan. This guarantees independence between training and evaluation data.

The use of a merged multi-cohort dataset introduces variability in acquisition protocols and population characteristics, providing a more challenging and realistic evaluation setting for assessing model generalization.

No additional class balancing techniques were required during training, as the dataset was explicitly constructed to maintain equal class distribution.

4.2. Experimental Design

The proposed model was trained and evaluated on multiple 3D ROI datasets derived from anatomically relevant brain regions, including the entorhinal cortex, fornix, frontal lobe, hippocampus, parietal lobe, and temporal lobe. All datasets were generated from a common set of MRI volumes to ensure experimental consistency and to prevent data leakage across regions.

Prior to ROI extraction, the MRI volumes were processed using a standardized preprocessing pipeline that includes skull stripping, automated tissue segmentation, and spatial normalization to the MNI152 template. Tissue segmentation and anatomical delineation were performed using established neuroimaging tools, enabling consistent identification of brain structures across subjects. These steps ensure inter-subject alignment and reduce non-brain variability, thereby supporting reliable ROI extraction and robust feature learning.

ROI-Specific Hemispheric Analysis for Multi-ROI Representation.

This experiment is designed to evaluate the contribution of anatomical regions of interest (ROIs) across cerebral hemispheres for Alzheimer's disease (AD) classification, with the aim of defining an optimized multi-ROI representation. The analysis focuses on determining whether hemispheric lateralization influences the discriminative capacity of individual ROIs and how this information can be leveraged to improve model design.

The analyzed feature set includes:

- **Clinical metadata:** demographic and cognitive variables such as age, sex, and MMSE.

 doi:10.20944/preprints202605.0910.v1

- **Volumetric biomarkers:** structural measurements derived from MRI, including gray matter, white matter, cerebrospinal fluid (CSF), hippocampus, amygdala, ventricles, entorhinal cortex, and whole-brain volume.

Each ROI was independently analyzed using MRI data extracted from both left and right hemispheres. The evaluated regions include the entorhinal cortex, fornix, frontal lobe, hippocampus, parietal lobe, and temporal lobe. For each ROI, two configurations were considered: (i) extraction from the left hemisphere and (ii) extraction from the right hemisphere. Each configuration was used to train and evaluate a classification model under identical conditions, ensuring a controlled hemispheric comparison.

Model evaluation was conducted using stratified 7-fold cross-validation, providing robust performance estimates and reducing variability associated with data partitioning. Performance metrics were computed independently for each ROI–hemisphere configuration.

Based on this experimental setup, a selection mechanism was defined to identify the most informative hemisphere for each ROI.

Formally, the optimal hemisphere h^* for each ROI r is selected as:

$$h^*(r) = \arg\max_{h \in \{L,R\}} \text{Accuracy}(r, h) \tag{9}$$

The selected ROI–hemisphere pairs are subsequently combined to construct the multi-ROI representation used as input to the proposed multimodal framework.

This design enables the identification of spatially localized and hemisphere-specific patterns relevant to AD classification, while reducing the inclusion of redundant or non-informative regions. By focusing on anatomically meaningful inputs, the approach is expected to enhance both predictive performance and interpretability.

The outcomes of this analysis provide the methodological foundation for the ROI-based decomposition strategy evaluated in the subsequent ablation study.

Ablation Study Design: Multimodal Integration and ROI-Based Representation.

This experiment investigates the contribution of different data modalities and to assess the effectiveness of the proposed ROI-based decomposition strategy within the multimodal framework for Alzheimer's disease (AD) classification.

The experimental design considers four model configurations, each representing a different level of information integration:

- **MRI Only (ROI-based):** uses only MRI inputs extracted from anatomically defined regions of interest.
- **Tabular Only:** uses only non-imaging features, including clinical metadata (e.g., age, sex, MMSE) and volumetric biomarkers derived from structural MRI.
- **Whole-brain (w/o ROI):** uses full MRI volumes without ROI decomposition, providing a baseline to evaluate the impact of anatomically constrained representations.
- **Multi-ROI + Tabular (Proposed):** integrates ROI-based MRI inputs with clinical and volumetric features within a multimodal transformer architecture.

All configurations are trained and evaluated under identical conditions to ensure a fair comparison. The same preprocessing pipeline, data splits, and training protocol are applied across all models.

Performance is assessed using standard classification metrics, including accuracy and area under the ROC curve (AUC), computed independently for each fold.

To quantify the significance of performance differences between configurations, paired statistical tests are applied across folds. Specifically, paired Student's t-tests are used to compare models, and effect sizes are measured using Cohen's d, providing a standardized estimate of the magnitude of observed differences.

 doi:10.20944/preprints202605.0910.v1

This experimental design enables a systematic analysis of: (i) the individual contribution of imaging and non-imaging modalities, (ii) the added value of multimodal integration, and (iii) the impact of ROI-based decomposition compared to whole-brain representations.

The outcomes of this experiment are presented in the following section, where the comparative performance of each configuration is analyzed in detail.

Attention-Based Feature Importance: Clinical and Volumetric Contributions.

This experiment evaluates the contribution of non-imaging features within the proposed multimodal framework by leveraging the attention-pooling mechanism. The objective is to obtain an interpretable approximation of feature importance for clinical metadata and volumetric biomarkers integrated into the model.

The analysis focuses exclusively on non-imaging tokens. To this end, MRI patch tokens corresponding to spatial representations are excluded, and only the attention weights associated with clinical and volumetric inputs are considered. This separation enables an isolated evaluation of structured features within the multimodal attention space.

Attention weights are extracted from the attention-pooling layer of the trained model for each fold in the 7-fold cross-validation. For each feature, attention scores are aggregated across all samples within a fold and subsequently averaged across folds to obtain a robust estimate of feature contribution. The variability of these estimates is quantified using the standard deviation across folds.

Since attention weights are normalized and not directly interpretable in absolute terms, features are ranked based on their relative importance. This ranking provides a comparative assessment of the contribution of each feature to the model's decision-making process.

This experimental design enables the identification of the most influential non-imaging features contributing to classification decisions, while ensuring robustness through cross-validation-based aggregation. Additionally, it provides an interpretable link between model predictions and clinically relevant biomarkers.

The outcomes of this analysis are presented in the subsequent section, where the ranked feature importance scores are reported and discussed.

ROI Attention Analysis and Interpretability.

This experiment was designed to evaluate the interpretability and stability of the proposed multimodal transformer by quantifying the contribution of each anatomical region of interest (ROI) to the final classification decision. Specifically, the analysis focused on determining whether the model assigns consistent attention to clinically relevant brain regions across the 7-fold cross-validation protocol.

After training the proposed model in each fold, ROI-level attention importance was extracted from the CLS-token attention weights. Attention values were aggregated across transformer heads and tokens to obtain a single importance score for each ROI. The evaluated ROIs included the entorhinal cortex, hippocampus, fornix, frontal lobe, parietal lobe, and temporal lobe.

For each cross-validation fold, the attention scores were averaged at the ROI level. These fold-level values were then summarized across the seven folds using the mean and standard deviation. This procedure allowed the assessment of both the average contribution of each ROI and the variability of its attention response across different data splits.

To evaluate whether any ROI received statistically dominant attention, a Friedman test was applied across ROI attention scores obtained from the seven folds. When pairwise comparisons were required, Wilcoxon signed-rank tests were performed between ROI pairs, and Bonferroni correction was applied to control for multiple comparisons.

Finally, attention stability was quantified using the coefficient of variation (CV) and a normalized stability score. Lower CV values and higher stability scores were interpreted as evidence of more consistent regional contribution patterns across folds. This analysis was used to assess whether the

 doi:10.20944/preprints202605.0910.v1

model learned reproducible and anatomically meaningful attention distributions rather than relying on unstable or fold-specific patterns.

Multimodal 3D Vision Transformer vs. State-of-the-Art Methods.

This experiment is conducted to benchmark the proposed Multimodal 3D Vision Transformer against recent state-of-the-art methods for Alzheimer's disease (AD) classification. The objective is to evaluate the effectiveness of the proposed approach in relation to existing transformer-based and hybrid deep learning models reported in the literature.

A comparative analysis is conducted using previously published studies that employ transformer-based architectures or related deep learning approaches on widely used neuroimaging datasets, including ADNI, AIBL, and OASIS. These methods encompass a range of modeling strategies, such as Vision Transformers, Swin Transformers, and hybrid CNN–Transformer architectures.

For comparison, the best-performing configuration of the proposed method is selected based on the experimental protocol defined in the previous sections. In addition, representative single-ROI configurations are included to assess the discriminative capacity of anatomically localized representations.

The evaluation is based on reported classification accuracy and AUC when available for the AD vs. cognitively normal (CN) task, as this metric is consistently available across the selected studies. When multiple datasets are used in prior work, their reported results are included as presented in the original publications.

It is important to note that differences in experimental protocols—such as dataset composition, preprocessing pipelines, and validation strategies—may affect direct comparability. In particular, many prior studies report single-split evaluations or lack fold-wise performance statistics. In contrast, the proposed method is evaluated using a 7-fold cross-validation protocol, reporting mean and standard deviation to provide a more robust estimate of performance.

The proposed model is trained and evaluated on a merged multi-cohort dataset combining ADNI, AIBL, and OASIS subjects, introducing additional variability and increasing the complexity of the classification task. This setting provides a more realistic and challenging benchmark compared to single-cohort evaluations.

This experimental design enables a contextualized comparison between the proposed multimodal framework and existing approaches, highlighting differences in model architecture, input modalities, and evaluation protocols.

4.3. Hyperparameter Optimization

Hyperparameter optimization was conducted using the Hyperband algorithm [58], which efficiently explores large and complex search spaces through adaptive resource allocation and early-stopping strategies.

The search space encompasses optimization-related parameters (optimizer type and learning rate), regularization mechanisms (dropout rate), and training configuration (batch size and number of epochs), as well as key architectural components of the transformer model, including the number of transformer layers and embedding dimensionality. The optimization objective was defined in terms of validation accuracy, enabling the selection of configurations that maximize generalization performance.

Table 2 summarizes both the explored hyperparameter search space and the final selected configuration.

Following the Hyperband-based search, a targeted fine-tuning phase was performed to further improve training stability and convergence behavior. This additional refinement step allows the model to operate under a more optimized configuration beyond the discretized search space.

 doi:10.20944/preprints202605.0910.v1

Table 2. Hyperparameter search space and selected optimal values using Hyperband optimization.

Hyperparameter	Search Space	Selected Value
Dataset	–	Merged (ADNI+AIBL+OASIS)
Slice Number	–	25
Image Size	–	32×32
Channels	{1, 3}	1
Optimizer	{Adam, SGD, RMSprop, AdamW}	AdamW
Learning Rate	{1e-3, 1e-4, 1e-5}	1×10^{-4}
Weight Decay	{1e-3, 1e-4}	1×10^{-3}
Clipvalue	–	0.5
Transformer Layers	–	8
Projection Dim	–	128
Embedding Dim	–	128
Num Heads	–	8
Patch Size	–	$(8, 8, 8)$
LayerNorm ϵ	–	1×10^{-6}
Dropout	{0.20 – 0.50}	0.20
Batch Size	{4, 6, 8, 16}	6
Epochs	{25, 50, 100}	100
Num Classes	{2, 3}	2

4.4. Statistical Analysis

Statistical analyses were performed to assess the robustness of the proposed model and to quantify the significance of performance differences among model configurations. For each experiment, accuracy and AUC were computed independently for each fold of the 7-fold cross-validation protocol and reported as mean $\pm$ standard deviation.

For the ablation study, paired Student's t-tests were applied across folds to compare the proposed Multi-ROI + Tabular model against the MRI-only, tabular-only, and whole-brain configurations. A significance level of $\alpha = 0.05$ was adopted. In addition to p-values, Cohen's d was computed to estimate the magnitude of the observed differences, enabling both statistical and practical interpretation of model improvements.

For attention-based interpretability analyses, feature and ROI attention scores were first averaged within each fold and then summarized across the seven folds using mean and standard deviation. Since ROI attention scores are repeated measurements obtained from the same cross-validation folds, a Friedman test was used to evaluate whether statistically significant differences existed among ROI attention distributions. When post-hoc pairwise comparisons were required, Wilcoxon signed-rank tests were applied with Bonferroni correction to control for multiple comparisons.

Finally, attention stability across folds was quantified using the coefficient of variation (CV) and a normalized stability score. Lower CV values and higher stability scores were interpreted as evidence of more consistent attention allocation across folds, supporting the reproducibility of the model's learned anatomical importance patterns.

4.5. Implementation Details

In the applied stratified 7-fold cross-validation strategy, each fold was partitioned into training (70%), validation (15%), and test (15%) subsets at the subject level, ensuring that no subject appeared in more than one subset while preserving the class distribution across all partitions. The multiple-input model was implemented in Python, with the random seed set for NumPy, TensorFlow, Random, and OS libraries to ensure reproducible results. Images were preprocessed using Python libraries NiBabel, TorchIO, PIL, and NumPy without saving them to disk. Skull stripping and MRI registration with the MNI152 template were performed using FreeSurfer tools. The classification models were built using the Keras library. All evaluated models were trained sequentially on ten workstations equipped with Intel Core i9 9900K processors, 32 GB RAM, and 11 GB NVIDIA RTX 2080Ti GPUs.

 doi:10.20944/preprints202605.0910.v1

5. Results

The proposed Multimodal 3D Vision Transformer was comprehensively evaluated through a series of experiments designed to assess its predictive performance, interpretability, and robustness for Alzheimer's disease classification. The evaluation includes (i) analysis of ROI-specific hemispheric contributions to identify the most discriminative anatomical regions, (ii) an ablation study to quantify the impact of multimodal integration and ROI-based representation, (iii) attention-based feature importance analysis to interpret the contribution of clinical and volumetric variables, and (iv) a comparison with state-of-the-art transformer-based methods. Performance was assessed using accuracy and area under the ROC curve (AUC) under a 7-fold cross-validation scheme to ensure statistical reliability and generalization.

5.1. Analysis of ROI-Specific Hemispheric Contributions for Multi-ROI Alzheimer's Disease Classification

The performance of individual ROIs across hemispheres is summarized in Table 3. For each ROI, the hemisphere achieving the highest classification accuracy is identified and highlighted. These best-performing ROI–hemisphere pairs are subsequently selected to construct the proposed multi-ROI representation. This selection strategy ensures that only the most discriminative anatomical regions contribute to the final model, improving both predictive performance and interpretability. Notably, the results reveal hemispheric asymmetries, indicating that certain brain regions provide more informative features depending on lateralization. This observation is consistent with known patterns of neurodegeneration in Alzheimer's disease and provides insight into the spatial distribution of disease-relevant biomarkers, supporting anatomically grounded model decisions.

Table 3. Classification accuracy (%) for each ROI across left and right hemispheres (mean ± standard deviation). The best-performing hemisphere per ROI is highlighted and used to construct the proposed multi-ROI representation. These results highlight hemispheric asymmetries and identify the most discriminative anatomical regions for Alzheimer's disease classification.

ROI	Left Hemisphere	Right Hemisphere
Entorhinal Cortex	96.67 ± 0.00	93.89 ± 2.08
Fornix	94.44 ± 0.79	96.67 ± 1.36
Frontal Lobe	96.67 ± 0.00	93.33 ± 0.00
Hippocampus	94.44 ± 1.57	93.89 ± 2.83
Parietal Lobe	92.78 ± 0.78	94.44 ± 3.42
Temporal Lobe	94.44 ± 0.79	96.67 ± 0.00

These findings motivate the use of ROI-based decomposition evaluated in the following ablation study.

5.2. Ablation Study: Contribution of Multimodal Integration and ROI-Based Representation

To quantitatively assess the contribution of each modality and the effectiveness of the proposed multi-ROI decomposition strategy, an ablation study was conducted comparing different model configurations, including MRI-only (ROI-based), tabular-only, whole-brain input without ROI decomposition, and the full multimodal framework. As shown in Table 4, the proposed Multi-ROI + Tabular model achieves the highest overall performance, reaching an AUC of 0.9940 ± 0.0059 and an accuracy of $97.62\% \pm 1.21$, substantially outperforming the MRI-only configuration (AUC: 0.6540 ± 0.0559, Accuracy: $63.33\% \pm 4.08$).

The tabular-only model exhibits strong baseline performance (AUC: 0.9871 ± 0.0089, Accuracy: $93.09\% \pm 0.58$), confirming that clinical and volumetric features capture highly discriminative disease-related patterns. However, the integration of ROI-based MRI information yields consistent improvements in both AUC and accuracy, demonstrating that localized anatomical features provide complementary information beyond structured clinical data.

In comparison with the whole-brain configuration (AUC: 0.9897 ± 0.0109, Accuracy: $95.95\% \pm 1.22$), the proposed multi-ROI model achieves comparable AUC but a noticeable improvement in

 doi:10.20944/preprints202605.0910.v1

classification accuracy. This indicates that ROI-based decomposition primarily enhances class separability and decision calibration rather than global ranking performance. Moreover, the whole-brain model tends to rely on diffuse spatial representations, limiting interpretability, whereas the proposed approach enables anatomically grounded attention focused on disease-relevant regions such as the hippocampus and entorhinal cortex.

Overall, these results demonstrate that (i) MRI data alone is insufficient for robust classification, (ii) tabular clinical features provide a strong predictive foundation, and (iii) the integration of structured data with anatomically constrained MRI representations yields superior performance while improving interpretability. This confirms the effectiveness of the proposed multimodal transformer framework for Alzheimer's disease classification.

Table 4. Ablation study evaluating the contribution of each modality in the proposed multimodal framework.

Model Configuration	Accuracy (%)	AUC
Multi-ROI + Tabular (Proposed)	97.62 ± 1.21	0.9940 ± 0.0059
MRI Only (ROI-based)	63.33 ± 4.08	0.6540 ± 0.0559
Tabular Only	93.09 ± 0.58	0.9871 ± 0.0089
Whole-brain (w/o ROI)	95.95 ± 1.22	0.9897 ± 0.0109

Statistical significance was assessed using paired t-tests across the 7-fold cross-validation for both AUC and accuracy metrics (Table 5). The proposed multimodal model significantly outperforms the MRI-only configuration ($p < 0.001$), with extremely large effect sizes (Cohen's $d \gg 1$), indicating a substantial performance gap. It also achieves statistically significant improvements over the tabular-only model ($p < 0.01$ for accuracy and $p < 0.05$ for AUC), with large effect sizes, confirming the strong contribution of multimodal integration.

In contrast, the difference between the proposed model and the whole-brain configuration is smaller and not consistently statistically significant ($p > 0.05$ for AUC), although a significant improvement is observed in accuracy ($p < 0.05$) with a moderate-to-large effect size. This suggests that while both models capture highly discriminative imaging features, the multimodal approach provides additional refinement through complementary clinical and volumetric information.

Importantly, beyond predictive performance, the proposed multi-ROI framework offers enhanced interpretability by explicitly focusing on anatomically relevant brain regions. This enables spatially localized and clinically meaningful explanations, in contrast to the more diffuse representations produced by whole-brain models.

Overall, these results demonstrate that multimodal integration yields statistically significant and practically meaningful improvements, while ROI-based decomposition enhances interpretability, supporting the suitability of the proposed framework for clinically explainable AI in Alzheimer's disease diagnosis.

Table 5. Statistical comparison of model configurations using paired t-tests across 7-fold cross-validation. Reported p-values and Cohen's d quantify statistical significance and effect size for both AUC and accuracy metrics.

Comparison	Metric	p-value	Cohen's d	Interpretation
Full vs MRI	AUC	< 0.001	> 6.0	Extremely large effect
	Accuracy	< 0.001	> 8.0	Extremely large effect
Full vs Tabular	AUC	< 0.05	~ 0.8	Moderate effect
	Accuracy	< 0.01	~ 3.5	Large effect
Full vs Whole-brain	AUC	> 0.05	~ 0.4	Small effect (not significant)
	Accuracy	< 0.05	~ 1.4	Moderate effect

To further elucidate the contribution of non-imaging features within the multimodal framework, we analyze their influence through the attention-pooling mechanism.

 doi:10.20944/preprints202605.0910.v1

5.3. Attention-Based Feature Importance: Clinical and Volumetric Contributions

To provide deeper insight into the multimodal decision-making process, we examine the relative contribution of non-imaging features using the attention-pooling mechanism. Feature importance scores were obtained by aggregating attention weights across the 7-fold cross-validation (mean $\pm$ standard deviation), ensuring robustness and consistency of the identified biomarkers. Since the absolute magnitude of attention scores is not directly interpretable, features are ranked based on their relative importance to assess their contribution to the model.

After excluding MRI patch tokens, the attention weights associated with clinical metadata and volumetric biomarkers were extracted and ranked to identify the most influential variables. As reported in Table 6, the resulting top-k features offer an interpretable approximation of feature importance within the multimodal framework. The small numerical scale reflects normalization of attention weights rather than weak feature contribution.

Table 6. Top-k clinical and volumetric features ranked by attention-pooling importance (mean $\pm$ standard deviation across 7-fold cross-validation).

#	Feature	Modality	Importance
1	Gray matter	Volumetric biomarker	0.000111 ± 0.000151
2	Ventricle left	Volumetric biomarker	0.000101 ± 0.000138
3	Entorhinal right	Volumetric biomarker	0.000101 ± 0.000143
4	Entorhinal left	Volumetric biomarker	0.000101 ± 0.000141
5	MMSE	Clinical metadata	0.000101 ± 0.000142
6	CSF	Volumetric biomarker	0.000100 ± 0.000142
7	Ventricle right	Volumetric biomarker	0.000100 ± 0.000141
8	Hippocampus left	Volumetric biomarker	0.000100 ± 0.000140
9	Age	Clinical metadata	0.000100 ± 0.000141
10	Amygdala left	Volumetric biomarker	0.000100 ± 0.000142
11	Sex	Clinical metadata	0.000099 ± 0.000141
12	Amygdala right	Volumetric biomarker	0.000099 ± 0.000140
13	Hippocampus right	Volumetric biomarker	0.000099 ± 0.000142
14	White matter	Volumetric biomarker	0.000099 ± 0.000135
15	Whole brain	Volumetric biomarker	0.000098 ± 0.000140

The results reveal a clear predominance of volumetric biomarkers, particularly gray matter volume, ventricular structures, and entorhinal cortex regions, which consistently rank among the most influential variables. These findings are well aligned with established neurobiological evidence, as structural atrophy in medial temporal regions and ventricular enlargement are key hallmarks of Alzheimer's disease progression. In addition, clinical variables such as MMSE, age, and sex also exhibit meaningful contributions, highlighting the complementary role of demographic and cognitive information.

Notably, the relatively low variability (standard deviation) across folds indicates that the identified features are stable and consistently leveraged by the model. Overall, this analysis demonstrates that the attention-pooling mechanism not only supports high predictive performance but also enables clinically coherent interpretation by prioritizing biologically relevant biomarkers.

These results reinforce the clinical validity of the proposed model by demonstrating that the learned representations are consistent with established neurodegenerative biomarkers.

Furthermore, these interpretable insights complement the quantitative performance analysis presented in the subsequent comparison with state-of-the-art methods.

5.4. ROI Attention Analysis and Interpretability

ROI attention importance was computed using CLS-token attention aggregated across transformer heads and tokens. To ensure statistical rigor, attention scores were first averaged per fold

 doi:10.20944/preprints202605.0910.v1

and subsequently summarized across the 7-fold cross-validation protocol using mean ± standard deviation.

As illustrated in Figure 3, the proposed multimodal transformer distributes attention relatively consistently across anatomically relevant brain regions associated with Alzheimer's disease.

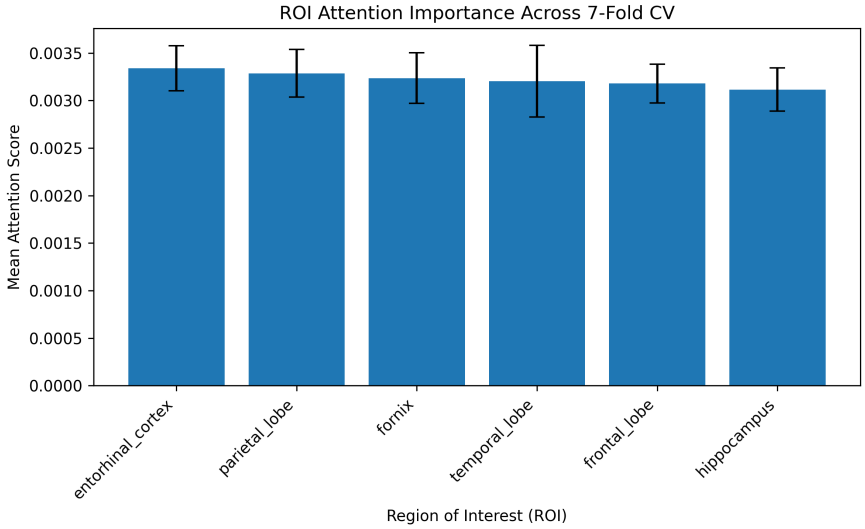

Figure 3. Mean ROI attention importance across the 7-fold cross-validation protocol computed from CLS-token attention weights. Error bars represent the standard deviation across folds, demonstrating stable and consistent attention responses among anatomically relevant brain regions associated with Alzheimer's disease.

The ROI attention analysis demonstrates that the proposed multimodal transformer distributes attention relatively consistently among anatomically relevant brain regions associated with Alzheimer's disease. As shown by the mean attention values, the **entorhinal cortex** achieved the highest average attention weight (0.0033 ± 0.0002), followed closely by the **parietal lobe** (0.0033 ± 0.0003) and the **fornix** (0.0032 ± 0.0003). These findings are anatomically meaningful, as the entorhinal cortex and fornix are strongly linked to memory impairment and early neurodegenerative progression in Alzheimer's disease.

The **hippocampus**, despite being one of the most clinically recognized biomarkers of Alzheimer's disease, exhibited the lowest average attention weight (0.0031 ± 0.0002). However, its low standard deviation indicates that the model consistently considers this region across folds, suggesting stable but more balanced attention allocation relative to the other ROIs.

As shown in Figure 3, the relatively small differences between attention means across ROIs suggest that the proposed architecture does not rely excessively on a single anatomical region. Instead, the model appears to learn a distributed multimodal representation in which multiple ROIs contribute complementary information to the classification process. This behavior is desirable in neuroimaging applications because Alzheimer's disease affects multiple interconnected brain structures rather than a single isolated region.

To further assess whether the observed attention differences among ROIs were statistically meaningful, non-parametric statistical analyses were performed across cross-validation folds.

To statistically evaluate whether the observed differences in attention allocation were significant, a Friedman test was performed across ROIs. The test yielded a statistic of 5.6939 with a p-value of 0.3372, indicating that no statistically significant global differences were observed among ROI attention distributions across folds. This result indicates that no ROI exhibited statistically dominant attention allocation across folds, suggesting that the transformer distributes attention in a relatively balanced manner among the selected anatomical regions.

Pairwise Wilcoxon signed-rank tests were subsequently conducted to explore potential differences between specific ROI pairs. Although the comparison between the **entorhinal cortex** and **hippocampus**

 doi:10.20944/preprints202605.0910.v1

produced the lowest uncorrected p-value ($p = 0.0156$), this difference did not remain statistically significant after Bonferroni correction ($p_{\text{adj}} = 0.2344$). All other pairwise comparisons similarly showed non-significant adjusted p-values.

Overall, these findings indicate that the proposed multimodal transformer learns a stable and anatomically distributed attention strategy, where multiple ROIs jointly contribute to Alzheimer's disease classification without a statistically dominant single region. This behavior supports the robustness, interpretability, and biological plausibility of the proposed multi-ROI representation framework.

The results presented in Table 7 demonstrate that the proposed multimodal transformer produces robust and stable attention distributions across cross-validation folds for all evaluated ROIs. Stability was quantified using the coefficient of variation (CV) and a normalized stability score, where lower CV values and higher stability scores indicate more consistent attention allocation.

Table 7. ROI attention stability analysis across cross-validation folds using the coefficient of variation (CV) and a normalized stability score. Lower CV values and higher stability scores indicate more consistent regional contribution patterns across folds.

ROI	Attention	CV	Stability Score
Frontal Lobe	0.0032 ± 0.0002	0.0640	**0.9399**
Entorhinal Cortex	0.0033 ± 0.0002	0.0710	0.9337
Hippocampus	0.0031 ± 0.0002	0.0729	0.9321
Parietal Lobe	0.0033 ± 0.0003	0.0762	0.9292
Fornix	0.0032 ± 0.0003	0.0823	0.9240
Temporal Lobe	0.0032 ± 0.0004	0.1175	0.8949

Among all regions, the **frontal lobe** achieved the highest stability score (0.9399) and the lowest coefficient of variation (0.0640), indicating that the model consistently assigns similar attention weights to this region across different training folds. This suggests that frontal lobe representations contribute robust and reproducible discriminative information for Alzheimer's disease classification.

The **entorhinal cortex** and **hippocampus** also exhibited high stability scores above 0.93, which is anatomically meaningful because these regions are strongly associated with early neurodegenerative changes in Alzheimer's disease. Their consistent attention allocation supports the biological plausibility of the proposed attention mechanism.

The **parietal lobe** and **fornix** showed slightly higher variability but still maintained strong stability scores above 0.92, indicating reliable participation in the decision-making process.

In contrast, the **temporal lobe** presented the highest coefficient of variation (0.1175) and the lowest stability score (0.8949), suggesting comparatively greater variability in attention allocation across folds. Although still relatively stable, this behavior may indicate higher sensitivity to inter-subject anatomical variability or dataset heterogeneity in temporal lobe patterns.

Overall, these findings indicate that the proposed multimodal transformer learns reproducible and anatomically coherent attention patterns, supporting the robustness and interpretability of the ROI-based representation strategy.

The consistency of attention allocation across folds further suggests that the learned ROI representations contribute reliably to the overall classification performance of the proposed multimodal framework.

5.5. Multimodal 3D Vision Transformer vs. State-of-the-Art Methods

Due to the relatively recent adoption of 3D Vision Transformers in medical imaging, only a limited number of studies have explored multimodal settings that combine multiple inputs, heterogeneous data sources, and merged datasets for Alzheimer's disease (AD) classification. To provide a comprehensive benchmark against state-of-the-art approaches, Table 8 summarizes the performance of

 doi:10.20944/preprints202605.0910.v1

recent transformer-based methods evaluated on commonly used datasets. For direct comparison, the best-performing configuration of the proposed method from Table 3 is included.

As shown in Table 8, the proposed *Multimodal 3D Vision Transformer* achieves the highest reported accuracy (97.62%) among the compared methods. In particular, it surpasses the best-reported transformer-based approach [38], which achieved 96.80% accuracy on ADNI using a Vision Transformer model.

Furthermore, the single-ROI configurations also demonstrate competitive performance, with several regions—such as the fornix, parietal lobe, and temporal lobe—reaching accuracies above 96%. This highlights the strong discriminative capacity of anatomically focused representations, while the multimodal multi-ROI configuration provides a more robust and generalizable solution.

The superior performance of the proposed model can be attributed to three key factors. First, the use of a 3D Vision Transformer enables the joint modeling of spatial and contextual information across volumetric MRI data through self-attention mechanisms. Second, the multi-ROI strategy focuses the model on disease-relevant anatomical regions, reducing noise from non-informative areas. Finally, the integration of heterogeneous data—including clinical metadata, cognitive assessments, and volumetric biomarkers—enhances the model's ability to capture complementary patterns associated with disease progression. Together, these components result in a highly accurate and interpretable framework for Alzheimer's disease classification.

The proposed method achieves higher accuracy than existing approaches. However, a direct statistical comparison is not feasible due to the absence of fold-wise performance data in prior works. Notably, our evaluation reports mean and standard deviation across 7-fold cross-validation, providing a more robust and reliable performance estimate.

Table 8. Comparison of classification accuracy between the proposed Multimodal 3D Vision Transformer and recent transformer-based approaches for AD vs. NC classification across different datasets.

Study	Dataset	Model	Accuracy
[34]	ADNI	Trans-ResNet	93.85
	AIBL		93.17
[38]	ADNI	Vision Transformer	96.80
[22]	ADNI, OASIS	2D CNN+Transformer	93.56
[41]	OASIS	Vision Transformer	91.18
[20]	ADNI, AIBL	CNN+Swin-Transformer	93.90
[5]	ADNI+AIBL	Swin Transformer	94.05
[40]	ADNI+OASIS	Vision Transformer	89.02
[59]	KAGGLE	TabTransformer (ETT-SSL)	95.80
[*OurProposal*] Multiple ROIs	Merged (ADNI + AIBL + OASIS)	Multimodal 3D Vision Transformer	97.62
Single ROI		Entorhinal Cortex - Left	96.67
		Fornix - Right	96.67
		Frontal Lobe - Left	96.67
		Hippocampus - Left	94.44
		Parietal Lobe - Right	94.44
		Temporal Lobe - Right	96.67

It is important to note that the proposed method is evaluated on a merged multi-cohort dataset, which introduces additional variability and increases the difficulty of the classification task.

The ROC curves presented in Figure 4 illustrate the discriminative performance of the proposed Multimodal 3D Vision Transformer across the 7-fold cross-validation. The consistently high true positive rates observed across folds, together with a mean AUC of 0.9940 ± 0.0059, indicate excellent classification capability and robust generalization. The narrow variability further suggests stable model behavior across different data partitions. Overall, these results demonstrate the model's strong ability to accurately distinguish between cognitively normal (CN) and Alzheimer's disease (AD) subjects.

 doi:10.20944/preprints202605.0910.v1

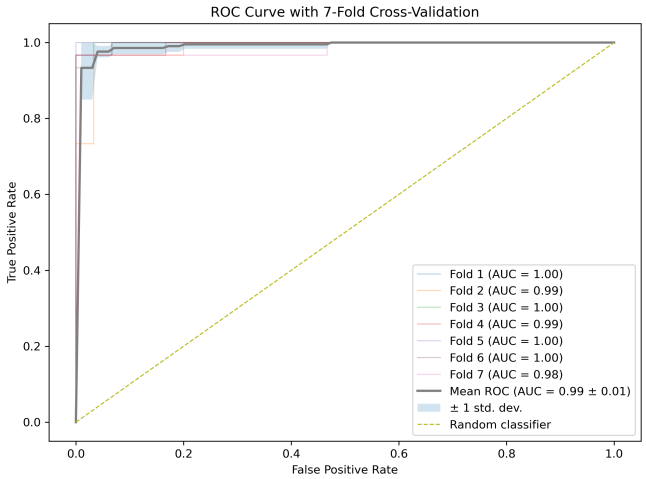

Figure 4. Receiver Operating Characteristic (ROC) curves for each fold in the 7-fold cross-validation. The mean ROC curve is shown in bold, with the shaded area representing ± one standard deviation, illustrating the robustness and consistency of the proposed model across different data splits.

Together, these results demonstrate that the proposed framework effectively balances predictive performance, robustness, and interpretability, addressing key challenges in clinically applicable Alzheimer's disease classification.

6. Limitations

Despite its strong performance, the proposed Multimodal 3D Vision Transformer presents several limitations. The use of public datasets (ADNI, AIBL, OASIS) introduces variability in acquisition protocols, scanner characteristics, and demographics, potentially leading to residual biases despite preprocessing. Additionally, diagnostic labels are based on clinical assessments rather than pathological confirmation, which may introduce label noise.

The reliance on predefined ROIs enhances interpretability but may overlook other informative regions and limit the capture of diffuse or atypical neurodegeneration patterns. Moreover, ROI-based sampling may result in partial information loss. Future work should explore adaptive or data-driven ROI selection strategies.

From a computational standpoint, Transformer-based multimodal models require substantial resources, which may hinder scalability and clinical deployment. Although cross-validation results are promising, external validation on independent cohorts is necessary to assess generalizability. Furthermore, the focus on simplified classification settings does not fully reflect the continuum of disease progression.

The integration of clinical and demographic data introduces challenges such as missing values, temporal inconsistencies, and potential imputation bias. While attention mechanisms improve interpretability, they provide only partial explanations of model decisions, and enhancing transparency remains essential for clinical adoption.

Finally, dependence on atlas-based ROI annotations and limited dataset sizes may affect robustness and generalization. Addressing these limitations through improved data diversity, external validation, adaptive feature selection, and more interpretable models is crucial for advancing real-world applicability.

7. Discussion

The results demonstrate that the proposed multimodal transformer architecture effectively captures both anatomical and clinical patterns associated with Alzheimer's disease, as reflected by its high predictive performance (AUC = 0.9940 ± 0.0059). By integrating multi-ROI volumetric MRI data with structured clinical features, the model leverages complementary sources of information that are typically analyzed independently in conventional approaches.

 doi:10.20944/preprints202605.0910.v1

The use of ROI-based decomposition enables the model to focus on anatomically relevant brain regions, such as the hippocampus and entorhinal cortex, which are known to be critically affected by neurodegeneration. This observation is consistent with established clinical findings identifying medial temporal lobe atrophy as a hallmark of Alzheimer's disease. Compared to whole-brain approaches, the proposed strategy improves the signal-to-noise ratio by reducing the influence of non-informative regions while preserving disease-relevant anatomical features.

A key strength of the proposed framework lies in its token-based multimodal representation. By embedding imaging and tabular information into a shared feature space, the model facilitates direct cross-modal interactions through self-attention mechanisms. This allows the architecture to capture complex dependencies between anatomical patterns and clinical biomarkers that are often overlooked by conventional fusion approaches based on late concatenation.

The incorporation of modality embeddings further improves the model's ability to differentiate and relate heterogeneous data sources, while attention-based pooling complements the global representation learned by the CLS token. This hybrid aggregation strategy enhances robustness by jointly modeling global contextual information and localized feature relevance.

The ablation study quantitatively supports these architectural design choices, demonstrating that both ROI decomposition and multimodal integration contribute substantially to classification performance. In particular, the multimodal framework consistently outperforms the tabular-only configuration, indicating that ROI-based imaging provides complementary information beyond clinical and volumetric biomarkers. These findings suggest that the proposed architecture does not rely solely on metadata, but instead effectively integrates heterogeneous modalities to improve disease characterization.

Despite these promising results, several limitations remain. The use of undersampling to balance class distributions may reduce training diversity and limit the ability of the model to capture rare disease patterns. Furthermore, although attention mechanisms improve interpretability, additional clinical validation is required to confirm the reliability and consistency of the identified attention patterns across heterogeneous populations and imaging protocols.

8. Conclusions

This study presented a multimodal 3D Vision Transformer framework for Alzheimer's disease classification that integrates ROI-based MRI representations with clinical and volumetric biomarkers within a unified transformer architecture. The proposed methodology combines anatomically guided ROI decomposition, multimodal token fusion, modality embeddings, and attention-based learning to jointly model heterogeneous data sources.

Experimental results demonstrate that the proposed framework achieves high classification performance, obtaining an AUC of 0.9940 ± 0.0059 under stratified 7-fold cross-validation. The results further show that multimodal integration improves predictive performance compared to unimodal configurations, while ROI-based decomposition enhances anatomical interpretability by focusing on clinically relevant brain structures associated with neurodegeneration.

The proposed architecture also provides a clinically meaningful interpretation of the classification process through attention-based mechanisms that highlight relevant anatomical regions and multimodal feature interactions. In contrast to conventional whole-brain approaches, the proposed ROI-centered strategy improves representation quality while preserving interpretability.

Overall, the findings indicate that combining anatomically informed representations with multimodal transformer-based fusion constitutes a robust and clinically relevant approach for Alzheimer's disease classification. Future work will focus on validating the proposed framework on independent external datasets, incorporating additional modalities such as PET imaging and genetic biomarkers, and extending the model to longitudinal studies for early prediction of disease progression.

 doi:10.20944/preprints202605.0910.v1

Funding: This research was partially funded by the Regional Ministry of Education of the Junta de Castilla y León (Spain). Project SA061G24, under ORDEN EDU/740/ 19 July 2024.

Data Availability Statement: The authors do not have permission to share data. However, the datasets used in this manuscript are publicly available and can be accessed at https://www.oasis-brains.org/ and https://adni.loni.usc.edu/.

Acknowledgments: The data used in this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in the analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgment_List.pdf (accessed on 11 May 2025). Data used in the preparation of this article were also partly obtained from the Australian Imaging Biomarkers and Lifestyle flagship study of aging (AIBL), funded by the Commonwealth Scientific and Industrial Research Organization (CSIRO), and made available in the ADNI database (www.loni.usc.edu/ADNI). The AIBL researchers contributed data but did not participate in the analysis or writing of this report. AIBL researchers are listed at https://aibl.csiro.au/ (accessed on 11 May 2025).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. World Health Organization. Dementia, 2025. Accessed: 2025-05-03.
2. National Institute on Aging. National Institute on Aging, 2025. Accessed: 2025-05-03.
3. Alzheimer's Society. Alzheimer's Society, 2025. Accessed: 2025-05-03.
4. Yu, J.; Lee, T.M. Verbal memory and hippocampal volume predict subsequent fornix microstructure in those at risk for Alzheimer's disease. *Brain Imaging and Behavior* **2020**, *14*, 2311–2322. https://doi.org/10.1007/s11682-019-00183-8.
5. Huang, Y.; Li, W. Resizer Swin Transformer-Based Classification Using sMRI for Alzheimer's Disease. *Applied Sciences (Switzerland)* **2023**, *13*. https://doi.org/10.3390/app13169310.
6. Das, R.; Kalita, S. Classification of Alzheimer's Disease Stages Through Volumetric Analysis of MRI Data. In Proceedings of the 2022 IEEE Calcutta Conference (CALCON). IEEE, 2022, pp. 165–169. https://doi.org/10.1109/CALCON56258.2022.10059718.
7. Khan, T.K. Chapter 3 - Neuroimaging Biomarkers in Alzheimer's Disease. In *Biomarkers in Alzheimer's Disease*; Khan, T.K., Ed.; Academic Press, 2016; pp. 51–100. https://doi.org/https://doi.org/10.1016/B978-0-12-804832-0.00003-1.
8. Tripathi, S.M.; Chutia, P.; Murray, A.D. Neuroimaging Biomarkers in Alzheimer's Disease. *Journal of Dementia and Alzheimer's Disease* **2025**, *2*, 1–20. https://doi.org/10.3390/jdad2040037.
9. Alzheimer's Association. Alzheimer's Association. https://www.alz.org/alzheimers-dementia/diagnosis/medical_tests, 2023. Accessed: 2023-12-10.
10. Alzheimer's Disease Neuroimaging Initiative (ADNI), http://adni.loni.usc.edu.
11. Australian Imaging, Biomarker and Lifestyle (AIBL) Flagship Study of Ageing, https://aibl.csiro.au.
12. Open Access Series of Imaging Studies (OASIS), http://www.oasis-brains.org.
13. Xu, Y. Patch-wise Intensity Mapping for Individualized Brain Abnormality Detection in Alzheimer ' s Disease Distributional Representation Normative Modeling Statistical Inference. *2025 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)* **2025**, pp. 6824–6833. https://doi.org/10.1109/ICCVW69036.2025.00705.
14. Wen, J.; Thibeau-Sutre, E.; Diaz-Melo, M.; Samper-González, J.; Routier, A.; Bottani, S.; Dormont, D.; Durrleman, S.; Burgos, N.; Colliot, O. Overview of classification of Alzheimer's disease. *Medical Image Analysis* **2020**, *63*, [1904.07773].
15. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need **2017**.
16. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. an Image Is Worth 16X16 Words: Transformers for Image Recognition At Scale. *ICLR 2021 - 9th International Conference on Learning Representations* **2021**.
17. Zhu, D.; Wang, D. Journal of Radiation Research and Applied Sciences Transformers and their application to medical image processing : A review. *Journal of Radiation Research and Applied Sciences* **2023**, p. 100680. https://doi.org/10.1016/j.jrras.2023.100680.

 doi:10.20944/preprints202605.0910.v1

18. Wang, Y.; Luo, Y.; Zu, C.; Zhan, B.; Jiao, Z.; Wu, X.; Zhou, J.; Shen, D.; Zhou, L. 3D multi-modality Transformer-GAN for high-quality PET reconstruction. *Medical Image Analysis* **2024**, *91*, 102983. https://doi.org/https://doi.org/10.1016/j.media.2023.102983.
19. Liu, L.; Liu, S.; Zhang, L.; To, X.V.; Nasrallah, F.; Chandra, S.S. Cascaded Multi-Modal Mixing Transformers for Alzheimer's Disease Classification with Incomplete Data. *NeuroImage* **2023**, *277*, [2210.00255]. https://doi.org/10.1016/j.neuroimage.2023.120267.
20. Xin, J.; Wang, A.; Guo, R.; Liu, W.; Tang, X. CNN and swin-transformer based efficient model for Alzheimer's disease diagnosis with sMRI. *Biomedical Signal Processing and Control* **2023**, *86*. https://doi.org/10.1016/j.bspc.2023.105189.
21. Li, C.; Wang, Q.; Liu, X.; Hu, B. An Attention-Based CoT-ResNet With Channel Shuffle Mechanism for Classification of Alzheimer's Disease Levels. *Frontiers in Aging Neuroscience* **2022**, *14*. https://doi.org/10.3389/fnagi.2022.930584.
22. Hu, Z.; Li, Y.; Wang, Z.; Zhang, S.; Hou, W. Conv-Swinformer: Integration of CNN and shift window attention for Alzheimer's disease classification. *Computers in Biology and Medicine* **2023**, *164*. https://doi.org/10.1016/j.compbiomed.2023.107304.
23. Menagadevi, M.; Mangai, S.; Madian, N.; Thiyagarajan, D. Automated prediction system for Alzheimer detection based on deep residual autoencoder and support vector machine. *Optik* **2023**, *272*. https://doi.org/10.1016/j.ijleo.2022.170212.
24. Al-Rahayfeh, A.; Atiewi, S.; Almiani, M.; Jararweh, M.; Faezipour, M. Utilizing 3D magnetic source imaging with landmark-based features and multi-classification for Alzheimer's Disease diagnosis. *Cluster Computing* **2024**, *27*, 2635–2651. https://doi.org/10.1007/s10586-023-04103-w.
25. Gravina, M.; García-Pedrero, A.; Gonzalo-Martín, C.; Sansone, C.; Soda, P. Multi input–Multi output 3D CNN for dementia severity assessment with incomplete multimodal data. *Artificial Intelligence in Medicine* **2024**, *149*. https://doi.org/10.1016/j.artmed.2024.102774.
26. Zheng, G.; Zhang, Y.; Zhao, Z.; Wang, Y.; Liu, X.; Shang, Y.; Cong, Z.; Dimitriadis, S.I.; Yao, Z.; Hu, B. A transformer-based multi-features fusion model for prediction of conversion in mild cognitive impairment. *Methods* **2022**, *204*, 241–248. https://doi.org/10.1016/j.ymeth.2022.04.015.
27. Coupé, P.; Manjón, J.V.; Mansencal, B.; Tourdias, T.; Catheline, G.; Planche, V. Hippocampal-amygdalo-ventricular atrophy score: Alzheimer disease detection using normative and pathological lifespan models. *Human Brain Mapping* **2022**, *43*, 3270–3282. https://doi.org/10.1002/hbm.25850.
28. Göschel, L.; Kurz, L.; Dell'Orco, A.; Köbe, T.; Körtvélyessy, P.; Fillmer, A.; Aydin, S.; Riemann, L.T.; Wang, H.; Ittermann, B.; et al. 7T amygdala and hippocampus subfields in volumetry-based associations with memory: A 3-year follow-up study of early Alzheimer's disease. *NeuroImage: Clinical* **2023**, *38*, 103439. https://doi.org/10.1016/j.nicl.2023.103439.
29. icometrix. Volumetric MRI Quantification in the Diagnosis of Alzheimer's Disease. https://www.icometrix.com/post/volumetric-mri-quantification-in-the-diagnosis-of-alzheimer-s-disease, 2021. Accessed: 2026-03-24.
30. Zaabi, M.; Smaoui, N.; Derbel, H.; Hariri, W. Alzheimer's disease detection using convolutional neural networks and transfer learning based methods. In Proceedings of the 2020 17th International Multi-Conference on Systems, Signals & Devices (SSD), 2020, pp. 939–943. https://doi.org/10.1109/SSD49366.2020.9364155.
31. Bae, J.B.; Lee, S.; Jung, W.; Park, S.; Kim, W.; Oh, H.; Han, J.W.; Kim, G.E.; Kim, J.S.; Kim, J.H.; et al. Identification of Alzheimer's disease using a convolutional neural network model based on T1-weighted magnetic resonance imaging. *Scientific Reports* **2020**, *10*, 1–10. https://doi.org/10.1038/s41598-020-79243-9.
32. Ahmed, S.; Kim, B.C.; Lee, K.H.; Jung, H.Y.; for the 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* **2020**, *15*, 1–23. https://doi.org/10.1371/journal.pone.0242712.
33. Pan, D.; Luo, G.; Zeng, A.; Zou, C.; Liang, H.; Wang, J.; Zhang, T.; Yang, B.; the Alzheimer's Disease Neuroimaging Initiative. Adaptive 3DCNN-Based Interpretable Ensemble Model for Early Diagnosis of Alzheimer's Disease. *IEEE Transactions on Computational Social Systems* **2022**, pp. 1–20. https://doi.org/10.1109/TCSS.2022.3223999.
34. Li, C.; Cui, Y.; Luo, N.; Liu, Y.; Bourgeat, P.; Fripp, J.; Jiang, T. Trans-ResNet: Integrating Transformers and CNNs for Alzheimer's disease classification. In Proceedings of the 2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI), 2022, pp. 1–5. https://doi.org/10.1109/ISBI52829.2022.9761549.

 doi:10.20944/preprints202605.0910.v1

35. Poloni, K.M.; Ferrari, R.J. Automated detection, selection and classification of hippocampal landmark points for the diagnosis of Alzheimer's disease. *Computer Methods and Programs in Biomedicine* **2022**, *214*, 106581. https://doi.org/10.1016/j.cmpb.2021.106581.
36. Aghaei, A.; Moghaddam, M.E. Smart ROI Detection for Alzheimer's Disease prediction using explainable AI. Technical report, 2023, [arXiv:eess.IV/2303.10401].
37. A. Castro-Silva., J.; Moreno-Garcia., M.; Guachi-Guachi., L.; H. Peluffo-Ordoñez., D. Instance Selection Framework for Alzheimer's Disease Classification Using Multiple Regions of Interest and Atlas Integration. In Proceedings of the Proceedings of the 13th International Conference on Pattern Recognition Applications and Methods - ICPRAM. INSTICC, SciTePress, 2024, pp. 453–460. https://doi.org/10.5220/0012469600003654.
38. Lyu, Y.; Yu, X.; Zhu, D.; Zhang, L. Classification of Alzheimer's Disease via Vision Transformer: Classification of Alzheimer's Disease via Vision Transformer. *ACM International Conference Proceeding Series* **2022**, pp. 463–468. https://doi.org/10.1145/3529190.3534754.
39. Hoang, G.M.; Kim, U.H.; Kim, J.G. Vision transformers for the prediction of mild cognitive impairment to Alzheimer's disease progression using mid-sagittal sMRI. *Frontiers in Aging Neuroscience* **2023**, *15*. https://doi.org/10.3389/fnagi.2023.1102869.
40. Mora-Rubio, A.; Bravo-Ortíz, M.A.; Arredondo, S.Q.; Torres, J.M.S.; Ruz, G.A.; Tabares-Soto, R. Classification of Alzheimer's disease stages from magnetic resonance images using deep learning. *PeerJ Computer Science* **2023**, *9*. https://doi.org/10.7717/peerj-cs.1490.
41. Altay, F.; Sánchez, G.R.; James, Y.; Faraone, S.V.; Velipasalar, S.; Salekin, A. Preclinical Stage Alzheimer's Disease Detection Using Magnetic Resonance Image Scans. *Proceedings of the AAAI Conference on Artificial Intelligence* **2021**, *35*, 15088–15097. https://doi.org/10.1609/aaai.v35i17.17772.
42. Li, C.; Cui, Y.; Luo, N.; Liu, Y.; Bourgeat, P.; Fripp, J.; Jiang, T. Trans-ResNet: Integrating Transformers and CNNs for Alzheimer's disease classification. *Proceedings - International Symposium on Biomedical Imaging* **2022**, *2022-March*, 1–5. https://doi.org/10.1109/ISBI52829.2022.9761549.
43. Albarakat, H.M.; Chaitanya, T.V.S.S.; .... HybridViT: An Approach for Alzheimer's Disease Classification with ADNI Neuroimaging Data. *Annamalai and Bassfar ...* **2025**. https://doi.org/10.1007/s42979-025-03862-0.
44. Zhang, Z.; Khalvati, F. Introducing Vision Transformer for Alzheimer's Disease classification task with 3D input. Technical report, 2022, [arXiv:eess.IV/2210.01177].
45. Tiwari, A.; Singhal, A.; Shigwan, S.; Kumar Singh, R.; Shigwan, S.J.; Tiwari, A.; Singhal, A.; Shigwan, S.; Singh, R. Early Diagnosis of Alzheimer through Swin-Transformer-Based Deep Learning Framework using Sparse Diffusion Measures. Technical report, 2023.
46. Zhang, W.; Yang, X.; Chen, Y.; Liu, Y. Alzheimer ' s Disease Classification Based on Multi-Scale 2D-VMD Swin Transformer. *2024 9th International Conference on Computer and Communication Systems (ICCCS)* **2024**, pp. 178–183. https://doi.org/10.1109/ICCCS61882.2024.10603331.
47. Illakiya, T.; Karthik, R. A Dimension Centric Proximate Attention Network and Swin Transformer for Age-Based Classification of Mild Cognitive Impairment From Brain MRI **2023**. *11*.
48. Hu, C. Image Feature Extraction with Fourier Transform and Multi - task Swin - Transformer for Alzheimer ' s Disease Prediction and Detection. *2024 4th International Signal Processing, Communications and Engineering Management Conference (ISPCEM)* **2024**, pp. 1028–1039. https://doi.org/10.1109/ISPCEM64498.2024.00182.
49. Maddalena, L.; Granata, I.; Giordano, M.; Manzo, M.; Guarracino, M.R. Integrating Different Data Modalities for the Classification of Alzheimer's Disease Stages. *SN Computer Science* **2023**, *4*. https://doi.org/10.1007/s42979-023-01688-2.
50. Birkenbihl, C.; Westwood, S.; Shi, L.; Nevado-Holgado, A.; Westman, E.; Lovestone, S.; Hofmann-Apitius, M. ANMerge: A comprehensive and accessible Alzheimer's disease patient-level dataset, 2020. https://doi.org/10.1101/2020.08.04.20168229.
51. Multimodal deep learning models for early detection of Alzheimer's disease stage. *Scientific Reports* **2021**, *11*, 1–13. https://doi.org/10.1038/s41598-020-74399-w.
52. Gao, X.; Shi, F.; Shen, D.; Liu, M. Multimodal transformer network for incomplete image generation and diagnosis of Alzheimer's disease. *Computerized Medical Imaging and Graphics* **2023**, *110*. https://doi.org/10.1016/j.compmedimag.2023.102303.
53. Golovanevsky, M.; Eickhoff, C.; Singh, R. Multimodal Attention-based Deep Learning for Alzheimer's Disease Diagnosis **2022**. [2206.08826]. https://doi.org/10.1093/jamia/ocac168.

 doi:10.20944/preprints202605.0910.v1

54. Zhang, X.; Lin, W.; Xiao, M.; Ji, H. Multimodal 2.5D convolutional neural network for diagnosis of Alzheimer's disease with magnetic resonance imaging and positron emission tomography. *Progress in Electromagnetics Research* **2021**, *171*, 21–34. https://doi.org/10.2528/pier21051102.
55. Odusami, M.; Maskeliūnas, R.; Damaševičius, R.; Misra, S. Explainable Deep-Learning-Based Diagnosis of Alzheimer's Disease Using Multimodal Input Fusion of PET and MRI Images. *Journal of Medical and Biological Engineering* **2023**, *43*, 291–302. https://doi.org/10.1007/s40846-023-00801-3.
56. Hughes, C.P.; Berg, L.; Danziger, W.; Coben, L.A.; Martin, R.L. A New Clinical Scale for the Staging of Dementia. *British Journal of Psychiatry* **1982**, *140*, 566–572. https://doi.org/10.1192/bjp.140.6.566.
57. Frisoni, G.B. Alzheimer's Disease Neuroimaging Initiative in Europe. *Alzheimer's & Dementia* **2010**, *6*, 280–285, [https://alz-journals.onlinelibrary.wiley.com/doi/pdf/10.1016/j.jalz.2010.03.005]. https://doi.org/https://doi.org/10.1016/j.jalz.2010.03.005.
58. Li, L.; Jamieson, K.; DeSalvo, G.; Rostamizadeh, A.; Talwalkar, A. Hyperband: A Novel Bandit-Based Approach to Hyperparameter Optimization. *Journal of Machine Learning Research* **2018**, *18*, 1–52.
59. Priyadharshini, M.; Murugesh, V.; Rybin, O. Enhancing Alzheimer's disease classification with a transformer-based model using self-supervised learning. *Scientific Reports* **2026**, *16*, 3798. https://doi.org/10.1038/s41598-025-33957-w.