International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(5s) 2026 pp. 690-709
https://doi.org/10.70917/ijcisim-2026-2751
Article

Structure-Aware Routing 3D Swin Transformer for Early-Stage Alzheimer's Disease Detection and Classification Using Structural MRI

Shruti VijayKumar Hegdekar[1]*, Prathibhavani P Maruthi[1], Venugopal Kuppanna Rajuk[2]

[1]Department of Computer Science and Engineering, University Visvesvaraya College of Engineering, Bengaluru, Karnataka, India
[2]Bangalore University, Bengaluru, Karnataka, India
***Corresponding author's Email:** hegdekarshruti@gmail.com

Abstract: Alzheimer's disease is a progressive neuroimaging disorder that causes severe cognitive and cellular decline in elderly people. Magnetic Resonance Imaging (MRI) is widely used as a non-invasive neuroimaging modality for evaluating brain atrophy patterns. However, existing models struggles to detect early stage structural changes due to rigid downsampling layers that blur critical tissue boundaries. To address this, SAR-Swin3D (Structure-Aware Routing 3D Swin Transformer) framework is introduced to effectively detect and classify the diseases stages. The process begins by feeding raw 3D T1-weighted brain volumes into patch embedding layer. The network then routes features through successive Swin3D blocks utilizing local 3D windowing methods to reduce whole-brain modeling complexity. A block-wise 3D SAR module replaces blind downsampling to extract continuous maps using coordinate attention. The proposed model introduces a specialized block-wise 3D-SAR module to extract continuous boundary maps. This mechanism maps subtle gray-matter tissue thresholds and prevents global, age-related brain shrinkage from complicating localized early stage disease pathology. Furthermore, local 3D windowing methods reduce whole brain modeling complexity from cubic to linear. The SAR-Swin3D model is evaluated on ADNI and OASIS datasets by performing accuracy of 98.47% for Cognitive Normal (CN) against Alzheimer's Disease (AD), 95.04% for stable Mild Cognitive Impairment (sMCI) against progressive (pMCI) and 99.81% for multi-class dementia staging.
Keywords: Alzheimer's, Coordinate attention, 3D Swin Transformer, Magnetic Resonance Imaging, Structure-Aware Routing.

1. Introduction

Alzheimer's disease is a terminal brain disorder that causes progressive cellular damage making the most dominant among elderly population [1]. Symptoms include memory loss, speech issues and behavioral shifts where diagnosis is depending on objective proof of mental decline stopping from memory loss [2]. Among different neuroimaging modalities, Magnetic Resonance Imaging (MRI) remains the primary diagnostic standard for evaluating Alzheimer's pathology [3]. Alzheimer's disease typically begins as Mild Cognitive Impairment (MCI) which is an intermediate state where 15% affected individuals transition to full each year and MCI serves as important predictive indicator for the disease [4]. While this disease remains irreversible, timely detection mitigating disease progression and preserving functional quality of life [5]. In clinical settings, expert radiologists perform visual evaluations of structural MRI scans by using standardized grading scales to measure regional atrophy within the medial temporal lobe, and posterior cortex [6]. However, the time and resource constraints of detailed neuropsychological evaluations make unfeasible for large-scale application in primary care reducing the value of screening [7]. In recent times, Deep Learning (DL) especially Convolutional Neural Network (CNN) have fundamentally advanced the field of medical image analysis [8]. CNN-based architectures have demonstrated high efficiency to neuroimaging by identifying subtle, pathological-driven structural variations that avoid manual visual inspection by clinical experts [9]. In particular, T1-weighted MRIs give better brain structure which is essential for measuring early stage biomarkers such as regional cortical thinning and hippocampal volume loss [10]. Alzheimer's is classified the progression of the disease into three phases such as mild, moderate and severe [11]. Traditional detection methods are labor-intensive and subjective as it

depends on skill and opinion of medical professionals where the final results vary between experts [12]. The exact mechanism behind Alzheimer remain unknown due to its complexity where experts have linked the disease to risk like smoking, high blood pressure and lack of education [13]. Recently researchers focus on two main causes to identify this disease such as amyloid beta

plaque hypothesis and the tau protein hypothesis [14]. Although Alzheimer's currently lacks a cure, developing treatments to slow its progression is highly effectively in early stages [15]. According to theWorld Health Organization (WHO), about 4% to 8% of adults aged 65 and older have dementia which is a type of Alzheimer [16]. CNN based architectures automatically learn layered feature from data and simplifies the classification process and drastically reduces the need for human oversight [17]. The Recurrent Neural Network (RNN) and Bi-Directional Gated Rectified Unit (Bi-GRU) models are used to track brain changes over time for better detection and classification [18]. The pre-trained EfficientNetB7 model uses a special technique to scale its depth, width and image resolution efficiently and equally [19]. Advanced diagnostic facilitates to generate massive volumes of medical imaging data, speed up the growth of large-scale hospital repositories [20]. Arfat Ahmad Khan *et al.* [21] developed Dual-3DM3-AD which used dual modalities, three dimensional integrated with Multimodal data fusion, Mixed-transformer segmentation and Multi-class staging for Alzheimer's Diagnosis. Mixed Transformer combined with a Furthered U-Net architecture executed pixel-level sematic segmentation to preserve white matter, grey matter and cerebrospinal fluid. Relevant structural and metabolic features were subsequently extracted from these segmented scans using parallel Resnet-51 encoders which were fused through Densely Connected Feature Aggregator Module. Multi-head attention fused data by reducing feature dimensionality allowing a final softmax layer to perform the multiclass diagnosis. However, Dual-3DM3-AD lack of interpretability where the complex architectures did not demonstrate underlying reason for stage specific decision. Santhosh Kumar Tripathy *et al*. [22] introduced Multiscale Feature Modeling Using Improved Spatial Attention Guided Depth Separable (I-SAB) CNN. The process started by feeding structural MRI scans into a 10-layer depth-wise separable CNN backbone to exploit spatial details with reduced computation. The feature maps from even-numbered layers were then channeled into parallel I-SAB. These blocks simultaneously extract max, average and min pooling properties to form more descriptive tensor. A specialized Feature Map Enhancement Module and skip connections process this tensor to generate enhanced spatial attention maps. These maps were undergoing elementwise multiplication by backbone features to produce spatially guided, multiscale vectors. Finally, these vectors were flattened and concatenated into scale-invariant features which a multilayer neural network classified into four stages using softmax output layer. However, the introduced model failed to classify the moderately demented class during domain adaptation. Fei Huang *et al.* [23] employed 3D-CNN enhanced Multiscale Progressive Vision Transformer (3D-CNN-MPVT) for early Alzheimer's disease staging. The process started by dividing high dimensional structural brain MRI scans into uniform, overlapping 3D patches. These blocks were fed into a pre-trained 3D Densenet121 feature extractor to map local regions into low-dimensional visual embedding's. This spatial embedding's were combined with trainable position tokens and passed directly into the MPVT. Within the MPVT, a vanilla transformer block used self-attention to process cross-patch relationships while an integrated inner CNN characterizes within patch structural atrophy. A specialized stitch operation progressively combined neighborhood spatial representations to optimize computational efficiency by reducing token numbers. However, the 3D-CNN-MPVT generalization capability remains highly dependent on large sample sizes, dropping when evaluating smaller external samples containing less severe cases. Huangjing Ni *et al*. [24] demonstrated Integer Ratio Based 3D Box-Counting Fractal Analysis (IRBCFA) to automatically identify individuals with Subjective Cognitive Decline (SCD). IRBCFA algorithm characterizes spatial complexity by dividing these irregular brain structures into flexible blocks using arbitrary box sizes derived from variable integer division ratios. The required box counts were dynamically calculated using fractional filling ratios to sensitively capture early disease induced structural changes. These regional fractal dimensions are processed through an iterative leave-one-out feature selection routine to filter out 75 consensus discriminative individuals with SCD from healthy aging. However, the performance of the 3D fractal dimension estimation is influenced by predominant reliance on edge blocks rather than main blocks during calculation. Shengchao Huang and Qun Dai [25] developed 3D Efficient and Essentialized Swin Transformer Network (E2STN) by feeding 3D structural MRI brains scans into an Efficient Swin Transformer (EST) to capture global structural pathology. A Focused Feature Enhancement Convolution Unit (FFE-CU) used convolutional operations and spatial attention to preserve fine-grained lesion details. These extracted global and local feature representations were combined through element-wise addition and passed to a Disease Risk Map generator (DRMg). Which visually maps voxel-level disease risks. Finally, an ROI-based classifier used Matthews correlation coefficient values to preserve top risk voxels for classification. However, the E2STN model completely separates disease risk localization and final diagnosis where diagnostic phase did not provide feedback loops to optimize the initial risk evaluation phase. Fuat Uyguroglu *et al.* [26] developed CNN-based Alzheimer's disease classification using fusion of multiple 3D angular

orientations by minimal preprocessing of raw structural brain MRI scans. The standardized 3D images were mathematically rotated into multiple unique angular orientations through spline interpolation. Multiple separate 3D Densenet121 architectures were then trained independently on different angular representations to generate standalone output probability boundaries. Finally, a multi-classifier collects these individual predictions using the sum rule to compute combine classification score. However, incorporating an expansive ensemble of multi-angle networks exponentially increase computational overheads. Nana Jia et al. [27] employed Multi-Modal Global-Local Fusion (MMGLF) framework for automated Alzheimer's disease classification. The process started by feeding raw 3D MRI brain scans and clinical tabular data into a Residual Network (ResNet) backbone and a custom text encoder respectively to extract high-level feature representations. A global fusion module then projects and concatenates the global modality vectors to preserve wide macro-structural contexts. An attention-based local module used the encoded tabular profiles to dynamically scale and weight fine-grained local blocks within the 3D MRI feature maps. Finally, these mixed global and local vectors were concatenated and pushed through a softmax output layer for multi-class staging. However, the MMGLF lacks initialization from large-scale 3D medical proxy tasks, forcing the 3D-CNN backbone to learn heavy geometric features entirely from scratch. Manish Kumar *et al.* [28] developed attentive DL with Randomized Vector Energy Least Square Twin Support Vector Machine (RV-ELSTSVM) by extracting 2D sagittal slices from preprocessed T1-weighted structural MRI scans. These slices were fed into a 10-layer Residual Network (ResNet) backbone integrated with Multi-Head Attention (MHA) mechanism to capture both local feature maps and global contextual dependencies. The extracted final features undergo a modular classification stage where it mapped onto a randomized feature space using random weighted network transformation. Finally, the Energy Least Square Twin Support Vector Machine used twin non-parallel hyperplanes with energy-based regularization to differentiate between disease stages. However, the RV-ELSTSVM dependent on extracting isolated 2D sagittal slices rather than directly processing 3D volumetric images restricts anatomical coverage and losing multi-planar spatial continuity across the whole brain.

1.1 Research Problem and Objective

Existing DL models face difficulties in early stage Alzheimer's detection, which uses blind downsampling layers. This blurs critical brain tissue boundaries and fails to preserve localized disease pathology from natural, age-related brain shrinkage. The main objective of this paper is to solve the above research problem by introducing a Structure-Aware Routing 3D Swin Transformer (SAR-Swin3D). This method replaces rigid downsampling with a 3D SAR module to preserve the delicate anatomical edges. This is achieved by integrating coordinate-aware attention and localized 3D windowing. The SAR-Swin3D model isolates subtle gray matter tissue loss while reducing 3D volumetric computational complexity from cubic to linear for accurate multi-class staging.

1.2 Contributions

The main contributions of SAR-Swin3D is as follows:

- A 3D Structure-Aware Routing (3D-SAR) module is introduced to preserve fine anatomical boundaries that are degraded by conventional patch-merging operations in hierarchical transforms.
- The boundary-guided coordinate-aware routing mechanism is developed to direct transformer attention toward localized disease-sensitive regions, so reducing the influence of normal age-related brain shrinkage.
- The proposed routing module is integrated into multiple stages of the hierarchical 3D Swin Transformer architecture, enabling simultaneous preservation of local structural information and global contextual representations.

1.3 Paper organization

The paper is organized as follows: Section 2 explains the proposed methodology in detail; Section 3 outlines the experimental setup and evaluation results of the proposed model, followed by a discussion; and Section 4 summarizes the paper and presents the conclusions.

2. Proposed Methodology

The proposed SAR-Swin3D is designed to process volumetric, high-resolution T1-weighted 3D structural MRI scans. The processing pipeline is divided into three main sequential phases: volumetric 3D data preprocessing, hierarchical 3D Swin transformer feature extraction, and 3D SR integration, as shown in Figure 1 represents the overall architecture.

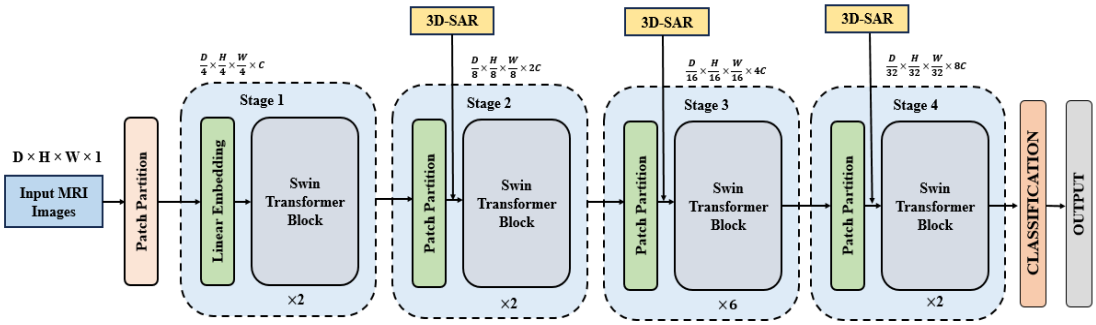

Figure 1: Overall Architecture of SAR-Swin3D for Alzheimer's Disease detection and classification

2.1 Dataset Description

Alzheimer's Disease Neuroimaging Initiative (ADNI) [29] dataset is accessed through the secure LONI Image and Data Archive (IDA). It is established as a comprehensive longitudinal multi-center cohort study, the main objective of the ADNI project is to evaluate, optimize and standardize neuroimaging, fluid and genetic biomarkers for multi-centers tracking and therapeutic clinical trials. Enrolled research participants undergo continuous tracking and are stratified into clinical stages with the diagnostic continuum based on standardized cognitive examinations. These stages are Cognitively Normal (CN), Mild Cognitive Impairment (MCI) which is demonstrated in Figure 2, which is subdivided into Early MCI and Late MCI in entry severity and Alzheimer's Disease (AD). For MCI researchers apply the labels as stable MCI (sMCI) and progressive MCI (pMCI) after tracking patients over several years to study the transformation by conversion to dementia. The repository provides highly dense, multimodal engineering environment. The corresponding data infrastructure integrates high-resolution 3D structural T1-weighted MRI volumes, multi-tracer metabolic PET grids, fluid bio specimen biomarker metrics and standardized clinical evaluation scores. Open Access Series of Imaging Studies [30] (OASIS-1) dataset is characterized as cross-sectional multi-modal unit comprising 416 subjects evaluated for clinical and structural brain morphology alterations associated with adult aging Alzheimer's disease. The dataset contains structural brain T1-weighted MRI scans captured from the original sagittal perspective. It is divided into two groups such as Dementia patients, who constitute 24.0% of the dataset (mean age 78.9 years) where 41.00% male) and non-demented control individuals who are remaining that 76.0% (mean age 54.226.2 years that is 37.7% male). Here Total number instances in dataset is 86,437 images where each class includes Mild Dementia 5,002, Moderate Dementia 488, Non-Dementia 67,692 and Very Mild Dementia (13,725). The samples based on classes are represented in Figure 3.

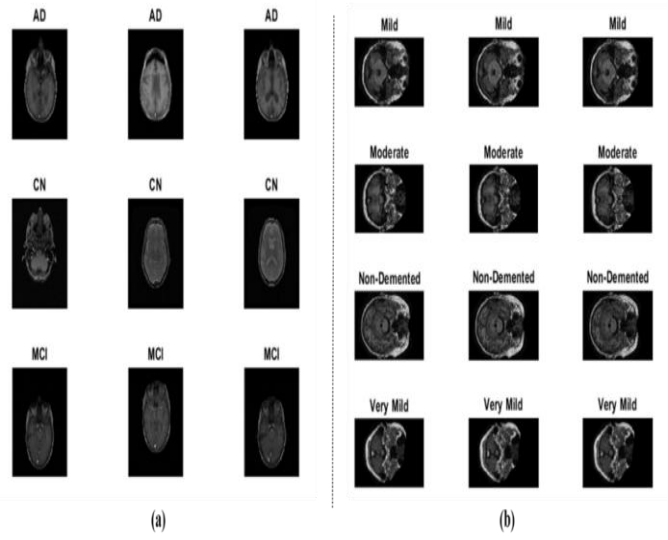

Figure 2: Sample Images of ADNI dataset for all three classes

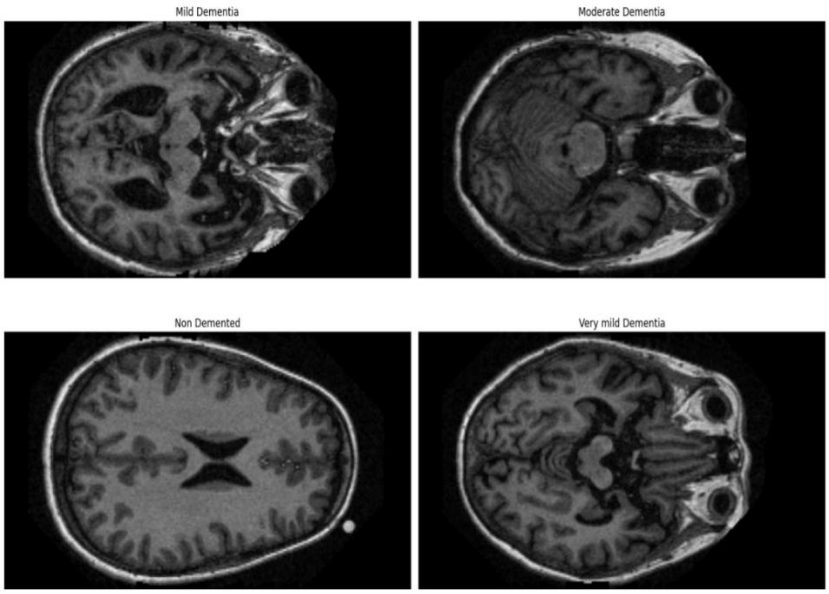

Figure 3: Sample images of OASIS-1 for each class

2.2 Volumetric 3D Data Preprocessing

To remove multi-center acquisition variations, Radio-Frequency (RF) field artifacts, and structural non-brain signals present within the ADNI and OASIS datasets, raw 3D structural MRI volumes undergo a three-stage preprocessing process.

2.2.1 3D Intensity Inhomogeneity Correction (N4 Bias Field)

Raw 3D structural MRIs naturally exhibit spatial intensity fluctuations caused by RF coil geometry and magnetic field variations, leading to artificial signal variance across uniform tissue types. The 3D n4 Bias Field Correction algorithm is applied to smooth out these non-pathological variations. This ensures that identical tissue types share uniform voxel values regardless of their location inside the scanner, preventing the downstream model from learning scanner-specific artifacts instead of true disease pathology. Let $I(x)$ represent the observed voxel intensity at a given 3D coordinate $x \in \Omega$ where, $\Omega \in R^3$ defines the spatial 3D volume image. The acquisition distortion is mathematically expressed in Equation (1).

$$I(x) = S(x) \cdot B(x) + N(x) \quad (1)$$

Here, $I(x) \in R$ is the corrupted voxel inetensity measures at the 3D spatial coordinate x, $S(x) \in R$ is the true, uncorrupted anatomically correct tissue signal. $B(x) \in R$ denotes the smoothly varying, low-frequency 3D spatial bias field multiplier, and $N(x) \in R$ is the independent additive white Gaussian noise. The N4 optimization routine iteratively minimizes the structural entropy of the log-transformed signal distribution to calculate an estimated bias field matrix $\hat{B}(x)$. The corrected 3D image volume is then preserved through element-wise division using Equation (2), where $I_{corr}(x) \in R$ represents the intensity-homogenized voxel value at the 3D coordinate x.

$$I_{corr}(x) = \frac{I(x)}{\hat{B}(x)} \quad (2)$$

2.2.2 3D Skull Stripping

Non-brain structures, including the skull, scalp, fat, eyes and Dural membranes contain heavy structural signals and high geometric variation. A DL based volumetric brain surface extractor is used to completely remove these non-brain elements and preserve the intracerebral volume. This process remove the irrelevant structural layers forces the network to concentrate of its parameters and spatial attention on internal brain anatomy, reducing false-positive features and improving processing speed. The 3D intracranial brain extraction is defined as in Equation (3), where $I_{brain}(x) \in R$ is the isolated intracerebral brain voxel value at the 3D coordinate x $\odot$ represents the 3D element-wise broadcasting multiplication operator $M(x) \in \{0,1\}$ is the binary spatial mask coordinate value generated by the extraction model, which forces all non-brain anatomical voxels strictly to zero.

$$I_{brain}(x) = I_{corr}(x) \odot M(x) \quad (3)$$

2.2.3 Spatial Normalization and Voxel Intensity Scaling

Human brains naturally vary in absolute size, shape, and spatial head tilt during imaging. A 12-degree-of-freedom was executed for a similar transformation to register all skull-stripped brains into a shared MNI152 standard space template at a uniform $1mm^3$ isotropic resolution, followed by global min-max voxel scaling. This helps spatial registration align every brain to identical coordinate addresses, enabling the transformer to directly compare specific anatomical regions across different subjects. Voxel intensity scaling standardizes the input distribution between 0 and 1, which speeds up the model optimization and completely prevents gradient saturation during backpropagation. The global min-max intensity scaling is mathematically expressed as follows (Equation (4)): Here, $V_{norm}(x)$ represents the final spatial and intensity-normalized isotropic 3D MRI block voxel value at coordinate x, $V(x) \in R$ is the similarity registered voxel value at coordinate x and $\min(V)$ and max (V) denote the absolute minimum and maximum voxel intensities calculated across the entire 3D brain volume tensor.

$$V_{norm}(x) = \frac{V(x) - \min(V)}{\max(V) - \min(V)} \quad (4)$$

2.3 Hierarchical 3D Swin Transformer Feature Extractor

The normalized structural 3D brain volume is forwarded to a hierarchical 3D Swin transformer architecture to construct multi-scale volumetric tissue representations across four stages.

2.3.1 Stage 1: 3D Patch Partitioning and 3D Linear Embedding

Transformers are sequence-based models which unable to read native 3D images; therefore, 3D patch partitioning is used to divide the continuous 3D volume into local $4 \times 4 \times 4$ voxel cubes, which are then flattened and projected through a 3D linear embedding layer into a formal token sequence. This structural division preserves localized, fine-grained 3D spatial properties within each cube while reshaping massive volumetric image data into a light, mathematically flexible token matrix format required for multi-head self-attention (MHA). The 3D patch partition and linear embedding computation are expressed in Equation (5). Here, $X_1 \in R^{L_1 \times C}$ represents the initial token sequence generated for Stage 1, C denotes the projected base channel representation depth, and L_1 represents the absolute sequence length parameter of Stage 1, derived from the 3D spatial dimension, as shown in Equation (6), where D, H, W represent the spatial depth, height, and width parameters of the 3D volume image, respectively.

$$X_1 = LinearEmbedding_{3D}\big(PatchPartition_{3D}(V_{norm})\big) \quad (5)$$

$$L_1 = \frac{D}{4} \times \frac{H}{4} \times \frac{W}{4} \quad (6)$$

2.3.2 3D Window Multi-Head Self-Attention (3D-WMSA)

The computation of standard global self-attention across an entire 3D brain volume creates an unsustainable computational burden that scales cubically with the image size. The 3D-WMSA is used to split the token map into localized $P \times P \times P$ sub-windows, restricting attention calculations within these small isolated 3D spaces, while alternating with a 3D Shifted Window (3D-SWMSA) step. This lowers the computational complexity from cubic to strictly linear, making 3D whole-brain modeling more practical. Alternating with the shifted window method allows adjacent 3D sub-windows to share information across their boundaries, smoothly capturing the interconnected cross-regional brain networks. The localized 3D-WMSA block computation is defined as in Equation (7), where $Q, K, V \in R^{P^3 \times d}$ represent the Query, Key, and Value linear transformation matrices computed from the 3D window tokens X_ω. d denotes the internal attention-head dimension parameter, $B \in R^{P^3 \times P^3}$ represents the 3D learnable relative position bias matrix configured to maintain the 3D spatial coordinate context within the window. The window boundaries are displaced between blocks by the structural offset of $\left(\frac{P}{2}, \frac{P}{2}, \frac{P}{2}\right)$ voxels along the depth, height, and width axes to enable 3D-SWMSA cross-window communications.

$$Attention_{3D}(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d}} + B\right)V \quad (7)$$

2.3.3 Stage 2: 3D Downsampling and Channel Expansion

As features pass deeper into the model, the architecture changes from detecting tiny local textures to evaluating larger anatomical structures. The 3D patch merging layer is used to group clusters of $2 \times 2 \times 2$ neighboring tokens together, downsampling the spatial grid resolution by half while doubling the channel depth to 2C. This builds a multi-scale hierarchical representation framework. Stage 2 compresses the sequence length to maintain fast processing while expanding the feature depth to capture detailed local gray matter patterns. The Stage 2 3D patch merging step is expressed as in Equation (8), where $X_{m2} \in R^{L_2 \times C}$ represents the downsampled token sequence, and the Stage 2 spatial sequence length is scaled down, as shown in Equation (9).

$$X_{m2} = LinearReduction_{3D}\left(Concat_{2\times2\times2}(X_1)\right) \quad (8)$$

$$L_2 = \frac{D}{8} \times \frac{H}{8} \times \frac{W}{8} \quad (9)$$

2.3.4 Stage 3: Hierarchical 3D Volumetric Scaling

To accurately track intermediate structures, the model requires a broader receptive field. Stage 3 applies another round of 3D patch merging to further condense the spatial map and expand the channel depth to 4C. This allows the transformer blocks in stage 3 to observe larger regional fields of view, which is essential for detecting early volumetric tissue loss and shape changes in subcortical brain structures. The Stage 3 3D downsampling operation is modeled as in Equation (10), where $X_{m3} \in R^{L_3 \times 4C}$ is the Stage 3 token sequence, and its operating sequence length corresponds to Equation (11).

$$X_{m3} = LinearReduction_{3D}\left(Concat_{2\times2\times2}\left(X_{stage2_{out}}\right)\right) \quad (10)$$

$$L_3 = \frac{D}{16} \times \frac{H}{16} \times \frac{W}{16} \quad (11)$$

2.3.5 Stage 4: Global 3D Structural Feature Aggregation

The final stage is designed to capture macro structural changes across the whole brain. The final 3D patch merging layer shrinks the spatial sequence to its most compact form while expanding the channels to 8C. This extreme spatial reduction compiles all local and regional features into a dense, high-dimensional global context vector, allowing the final deep 3D Swin blocks to make highly accurate dimensional global context vector. The deepest 3D downsampling step is mathematically formulated as in Equation (12), where $X_{m4} \in R^{L_4 \times 8C}$ represents the deepest feature token sequence, and its condenses sequence length equals Equation (13).

$$X_{m4} = LinearReduction_{3D}\left(Concat_{2\times2\times2}\left(X_{stage3_{out}}\right)\right) \quad (12)$$

$$L_4 = \frac{D}{32} \times \frac{H}{32} \times \frac{W}{32} \quad (13)$$

2.4 3D Structure-Aware Routing (3D-SAR) Integration

Standard 3D patch merging depends on hard downsampling, which concentrates spatial groups that blur fine structural boundaries. This creates a critical overlapping boundary problem, where subtle diagnostic thresholds, such as separating CN from early MCI or Non-Demented from Very Mild Demented, are lost by normal age-related brain changes. Therefore, the insertion of a block-wise 3D-SAR module directly after patch merging in Stages 2,3, and 4 preserves these fine anatomical edges and prevents vanishing gradients. Instead of blindly downsampling the pixels, the 3D-SAR reconstructs the spatial tensor to extract a continuous boundary map. It then uses coordinate-aware routing to direct features with these critical tissue thresholds, ensuring that early pathological signs are never drowned out by normal aging or lost in deep network layers. Figure 4 shows the 3D-SAR module architecture and its process.

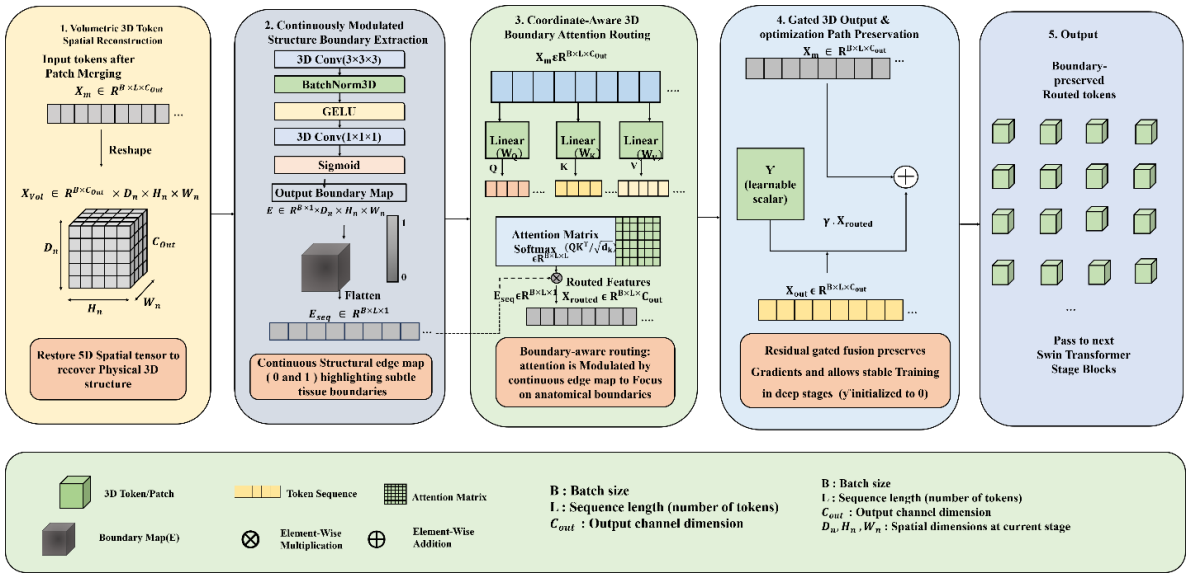

Figure 4: Architecture diagram of 3D-SAR module introduced for Swin Transformer

2.4.1 Volumetric 3D Token Spatial Reconstruction

Linearized token matrices lose their physical spatial coordinates; therefore, the flattening is temporarily reversed by reshaping the token sequence X_m back into its original 5D tensor configuration ($B \times C \times D \times H \times W$). This restores full geometric spatial continuity, allowing localized 3D convolutional filters to accurately view and evaluate the physical and anatomical edges and curves of the brain structure. The 3D spatial reconstruction is mathematically defined as in Equation (14), where $\mathcal{R}$ denotes the 3D spatial reconstruction operator, X_{vol} is the reconstructed 5D tensor structure containing physical 3D spatial continuity. B is the batch size, and D_n, H_n, W_n are the downsampled spatial parameters tracking the current active stage.

$$\mathcal{R}(X_m) = X_{vol} \in R^{B \times C_{out} \times D_n \times H_n \times W_n} \tag{14}$$

2.4.2 3D Continuously Modulated Structural Boundary Extraction

To preserve subtle tissue variations without depending on hard, artificial thresholds, the model route the spatial tensor through a specialized 3D convolutional gradient network that maps structural changes using a continuous scale between 0 and 1. This extracts a smooth, high-fidelity boundary map (E). It highlights early, subtle tissue thinning and boundary shifts, preventing these delicate pathological markers from being erased by the uniform downsampling layers. The continuous 3D boundary map extraction is defined in the Equation (15). Here $\Psi_{3\times3\times3}$ represents a 3D volumetric convolution layer with an isotropic $3 \times 3 \times 3$ receptive field designed to track localized spatial structural changes, $BatchNorm3D$ is the 3D Batch Normalization layer. GELU is the Gaussian ERROR Linear Unit activation function, $\Psi_{1\times1\times1}$ is a $1 \times 1 \times 1$ point-wise 3D convolutional layer compressing channel dimensions down to a single continuous map. σ is the sigmoid activation function restricting the final continuous structural tissue edge map representation to a localized density range $E \in R^{B\times1\times D_n\times H_n\times W_n}$. The resulting 3D structural edge map E is flattened back into a sequence representation $E_{seq} \in R^{B\times L\times1}$ to act as a token-wise gating multiplier.

$$E = \sigma\left(\Psi_{1\times1\times1}\left(GELU\left(BatchNorm3D\left(\Psi_{3\times3\times3}(X_{vol})\right)\right)\right)\right) \tag{15}$$

2.4.3 Coordinate-Aware 3D Boundary Attention Routing

General brain shrinkage occurs naturally as people age, which easily hide localized disease indicators. The map of Queries, Keys and Values into a coordinate-aware attention matrix, gating the final output with the continuous boundary map (E_{seq}). This focuses attention updates specifically along critical anatomical edges rather across uniform tissue zones, preventing normal, global aging variations from obscuring local, early stage Alzheimer's biomarkers. The coordinate-aware 3D routing computation is expressed as in Equation (16), (17) and (18). Here $W_Q, W_K \in R^{C_{out}\times\frac{C_{out}}{8}}$ are the linear projection parameter weights for query and key mappings, $W_V \in R^{C_{out}\times\frac{C_{out}}{8}}$ is the weight for

the value mapping, $A_R \in R^{B\times L\times L}$ is the structural routing matrix, $d_k = \frac{C_{out}}{8}$ is the scale factor, $X_{routed} \in R^{B\times L\times C_{out}}$ is the boundary-constrained routed feature vector and $\odot$ represents element-wise broadcasting multiplication.

$$Q = X_m W_Q, K = X_m W_K, V = X_m W_V \quad (16)$$

$$A_R = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right) \quad (17)$$

$$X_{routed} = (A_R \cdot V) \odot E_{seq} \quad (18)$$

2.4.4 Gated 3D Output and Optimization Path Preservation

The addition of custom modules to deep networks disrupts the gradient flow, leading to training instability and vanishing gradients. A learnable scaling factor (γ) is implemented by initializing it to 0 to control the residual shortcut connection. Initializing γ at 0 allows the model to stabilize its base training parameters early on. As training progresses, it dynamically opens the routing gates, enabling the gradients in the deep layers (stages 3 and 4). The gated residual combining step is mathematically formulated as follows (Equation (19)): Here, $X_{out} \in R^{B\times L\times C_{out}}$ is the final structural output tensor fed directly into the subsequent standard Stage N 3D Swin blocks, and $\gamma \in R^1$ is the learnable scalar residual parameter initialized to 0.

$$X_{out} = X_m + \gamma \cdot X_{routed} \quad (19)$$

2.4.1 Volumetric Token State Transitions

To trace the volumetric feature transformations across two consecutive 3D Swin Transformer blocks in Stage N, the network tracks a sequence of intermediate token states denoted as $z^{l-1}, \hat{z}^l, z^l, \hat{z}^{l+1}$ and z^{l+1}. As illustrated in the block-wise propagation path, the token states transition as follows in Equations (20), (21), (22), and (23):

$$\hat{z}^l = 3D - WMSA\left(LN(z^{l-1})\right) + z^{l-1} \quad (20)$$

$$z^l = MLP\left(LN(\hat{z}^l)\right) + \hat{z}^l \quad (21)$$

$$\hat{z}^{l+1} = 3D - SWMSA\left(LN(z^l)\right) + z^l \quad (22)$$

$$z^{l+1} = MLP\left(LN(\hat{z}^{l+1})\right) + \hat{z}^{l+1} \quad (23)$$

The initialization of the input token vector z^{l-1} depends on the active architectural stage, where in stage 1, it is initialized directly from the flattened output of the initial 3D linear embedding layer (X_1). To prevent subtle diagnostic thresholds from being blurred by standard downsampling, z^{l-1} is explicitly replaced by the boundary-enhanced output tensor X_{out} from the 3D-SAR module, as shown in Equation (24). Figure 5 shows the two successive swin transformer blocks updated with 3D W-MSA and 3D SW-MSA.

$$z^{l-1} = X_{out} \quad (24)$$

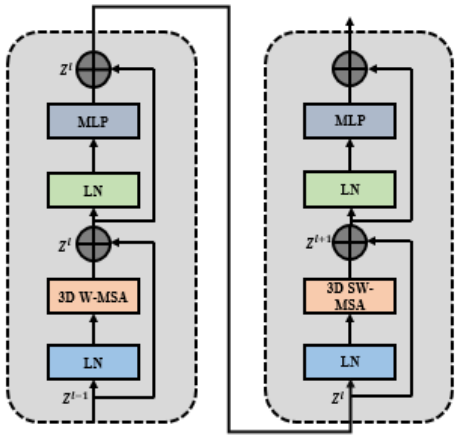

Figure 5: Two successive swin transformer with 3D W-MSA and 3D SW-MSA

By routing X_{out} into z^{l-1} at the start of stages 2,3 and 4, the localized 3D window extraction state $\hat{z}^{l}$ and cross-window shifted attention state $\hat{z}^{l+1}$ are heavily weighted with continuous spatial edge maps before any self-attention calculation begins. This direct connection forces successive transformer blocks to track subtle pathological tissue changes rather than global, non-pathological brain aging.

2.5 Global Feature Aggregation, Multi-Class Classification and Optimization Loss

Following block-wise propagation across successive 3D Swin transformer blocks, the deep volumetric feature representations from stage 4 undergo global reduction to perform final classification. The sequence of high-dimensional token features output from the final blocks of Stage 4, denoted as $z^{L_{out}} \in R^{L_4 \times 8C}$ is compressed into a compact global context vector through 3D Global Average Pooling (3D GAP). This operational workflow maps spatial dimensions to an invariant multi-class diagnostic representation, as shown in Equation (25).

$$X_{global} = 3D - GAP(z^{L_{out}}) \in R^{1 \times 8C} \tag{25}$$

This global context vector is subsequently routed through a classification head consisting of Layer Normalization (LN) and Multi-Layer Perceptron (MLP) architecture. The output is projected into a configuration space corresponding to the absolute number of target stage categories, such as CN, CMCI, pMCI, and AD. The class probability vector $\hat{y}$ is mathematically formalized using a multi-class Softmax layer.

$$\hat{y} = Softmax\left(MLP\left(LN\left(X_{global}\right)\right)\right) \tag{26}$$

To optimize structural edge detection with simultaneous cross-window disease biomarker tracking, the entire framework is trained end-to-end using a standard Categorical Cross-Entropy Loss function (L_{CCE}). For a given batch size N_b across K target stages, the optimization path minimizes the divergence between the true one-hot encoded diagnostic labels $y_{i,j}$ as shown in Equation (27).

$$\mathcal{L}_{total} = -\frac{1}{N_b}\sum_{i=1}^{N_b}\sum_{j=1}^{K} y_{i,j}\log\left(\hat{y}_{i,j}\right) \tag{27}$$

By minimizing $\mathcal{L}_{total}$, the error gradients propagate backward directly through the learnable scaling gates γ of the 3D-SAR modules. This updates the continuous boundary extraction maps to preserve highly localized, pathologically thin gray matter tissue layouts over large-scale non-pathological variations.

Algorithm: SAR-Swin3D Framework

Input: Preprocessed 3D structural MRI (V)

Output: Predicted Alzheimer's disease (Y)

1. Load 3D MRI volume (V)
2. Apply N4 bias-field correction
3. Perform skull stripping to remove non-brain tissues.
4. Normalize the MRI volume to the standard space and scale voxel intensities.
5. Partition the normalized volume into fixed-size 3D patches
6. Generate embedded feature tokens using the 3D patch embedding layer.
7. For each hierarchical Swin Transformer stage do

   Apply 3D-WMSA

   Apply 3D-SWMSA

   Perform patch merging and channel expansion.

   Insert the proposed 3D-SAR module:

   Reconstruct the spatial feature tensor.

   Extract continuous structural boundary maps.

Compute coordinate-aware attention weights.

Route and enhance feature representation using boundary-guided attention.

Fuse routed features through residual connections

8. End For.
9. Apply Global Average Pooling to obtain the global feature vector.
10. Pass the feature vector through Layer Normalization and the MLP classifier.
11. Compute class probabilities using the Softmax function.
12. During training, optimize the network using Cross-Entropy Loss and the AdamW optimizer.
13. Return the predicted class Y.

3. Results and Discussion

The SAR-Swin3D model implementation, training, and evaluation pipelines are executed within the MATLAB (R2025b) environment, using the Deep Learning Toolbox, Image processing toolbox and medical image toolbox. Computational tasks are accelerated using high-performance workstation equipped with an AMD Ryzen Threadipper 3960X 24-Core processor running at 3.8 GHz, paired with 128 GB of DDR4 RAM to facilitate the handling of volumetric 3D structural MRI datasets. Table 1 presents the hyperparameter and its configurations used in the proposed model. Network training and parallel tensor operations are offloaded to an NVIDIA RTX 4090 GPU with 24GB of dedicated VRAM, leveraging CUDA 12.4 and the NVDIA DL Discriminative Libraries (cuDNN). The workstation operated on a 64-bit Ubuntu 24.04 LTS Linux operating system, providing a stable, high-throughput environment for end-to-end training and k-fold cross-validation routines. The dataset is divided at the subject-level to prevent data leakage between training and testing sets. To ensure reproducibility of the experimental results, a fixed random seed is used for dataset dividing, weight initialization, and training procedures. The random seed value is set to 42 and all experiments are conducted using same seed.

Table 1: Training Optimization Hyperparameter Table

Hyperparameter	Configuration Value
Optimization algorithm	AdamW
Initial Learning rate policy	
Learning rate policy	Cosine Annealing
Training epochs	300 epochs
Mini-batch size	32 volumes
Weight decay	
Loss function	Cross-Entropy Loss
Validation frequency	Every 1 epoch
Total samples considered	ADNI: 7,200 volumes OASIS-86,437 volumes
Training set (80%) and Testing (20%)	ADNI: 5,760 and 1,440 OASIS-69,159 and 17,278

3.1 *Evaluation Metrics*

Accuracy represents the overall percentage of structural brain MRI scans that the model identified completely correctly, and it is formulated as in Equation (28), where TP is the true positive, TN is the true negative, FP is the false positive, and FN is the false negative.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100 \quad (28)$$

Specificity measures the model accuracy in identifying completely healthy or stable control individuals (True Negative Rate), which is calculated using Equation (29).

$$\text{Specificity} = \frac{TN}{TN + FP} \times 100$$

Sensitivity (Recall) measures the model accuracy in catching progressive disease states (True Positive Rate), which is calculated using Equation (30).

$$\text{Sensitivity} = \frac{TP}{TP + FN} \times 100 \quad (30)$$

Precision represents the probability that a patient actually has Alzheimer's pathology when the model flags their scan as positive. It is calculated as shown in Equation (31).

$$Precsion = \frac{TP}{TP + FP} \times 100 \quad (31)$$

F1-Score acts as a balanced performance index by calculating the harmonic mean between precision and recall, which is calculated using Equation (32).

$$\text{F1} - \text{Score} = 2 \times \frac{\text{Precision} \times \text{Sensitivity}}{\text{Precision} + \text{Sensitivity}} \times 100 \quad (32)$$

The area Under the ROC Curve measures the overall performance of the model across all decision thresholds. It evaluates how the proposed map boundaries prevent intermediate categories by separating normal age-related brain shrinkage from early stage diseases.

3.2 *Performance Evaluation*

The Figure 6,7 and 8 show the performance of the proposed model with baseline models such as 3D ResNet, Vanilla 3D Swin transformer, Transformer-ResNet, and MHAGuidednet for the ADNI and OASIS-1 datasets. The proposed method provides improved results for the CN vs. AD and sMCI vs. pMCI classes. The OASIS-1 datasets evaluate the multi-class of dementia, where the proposed results highlight the integration of the SAR block to the Swin transformer by providing efficient values.

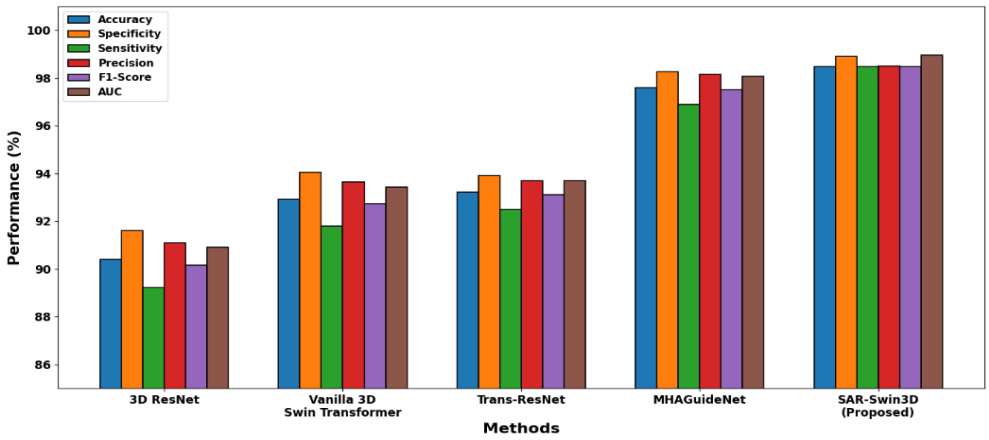

Figure 6: Performance evaluation of ADNI (CN vs AD) for SAR-Swin3D

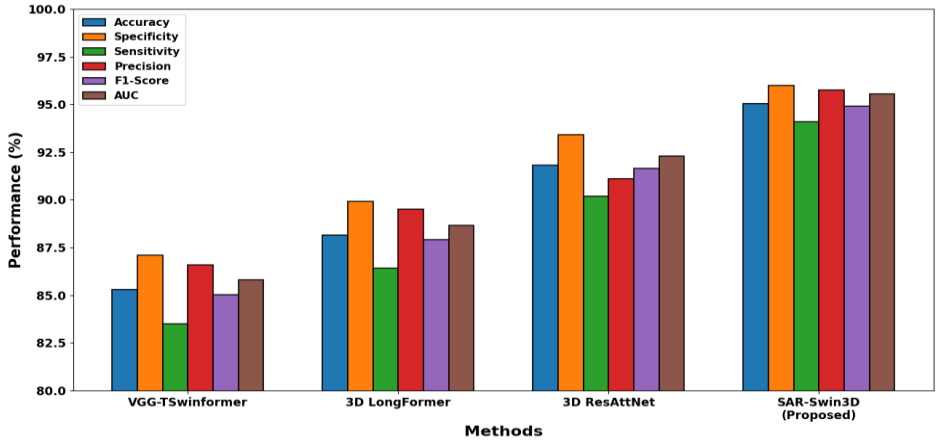

Figure 7: Performance evaluation of ADNI (sMCI vs pMCI)) for SAR-Swin3D

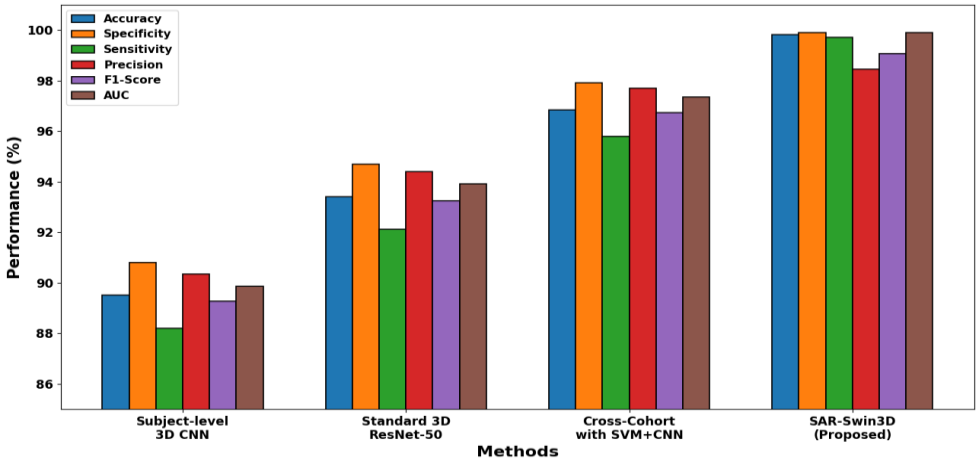

Figure 8: Performance evaluation of OASIS-1 for SAR-Swin3D

The training and validation performance curves across epochs for the evaluated on ADNI and OASIS-1 datasets is analyzed through curves. The Figure 9 shows the convergence rate and final classification accuracy. The Figure 10 show the training and validation loss curves, illustrating the minimization of the objective function. The proposed model exhibits faster convergence, higher asymptotic stability and minimal generalization gap.

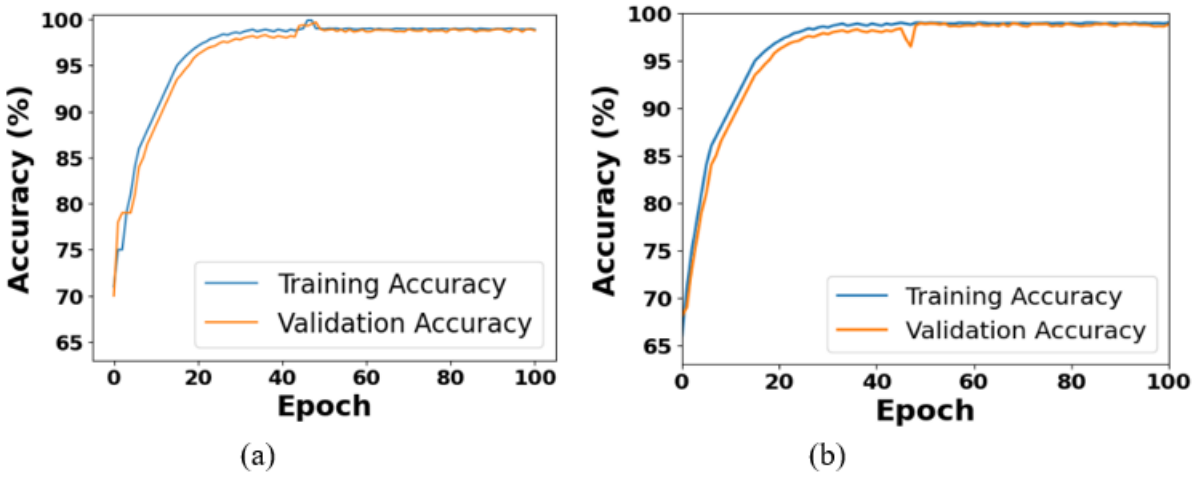

Figure 9: Accuracy curve of training and validation for (a) ADNI (b) OASIS-1 dataset

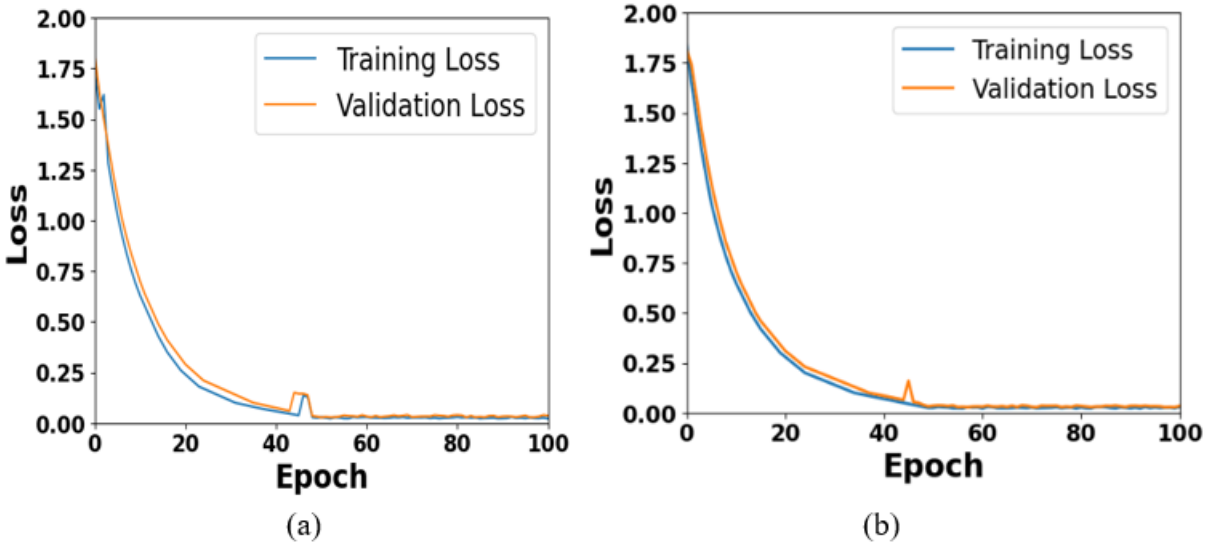

Figure 10: Loss Curve of training and validation for (a) ADNI (b) OASIS-1 dataset

Figure 11 presents the multi-class confusion matrices evaluating the SAR-Swin3D model across the ADNI and OASIS-1 datasets. The different diagonal distributions demonstrated high diagnostic precision and minimal off diagonal leakage. Figure 11 (a) classifies 466 CN, $&% MCI and $&& AD volumes by avoiding false-negative classification between the normal and AD groups. Figure 11 (b) shows the robust categorization across the four dementia stages, properly isolating 13,423 non-demented, 2,732 Very Mild, and all 98 Moderate Dementia scans. These results confirm that the 3D-SAR module preserves subtle early stage structural boundaries.

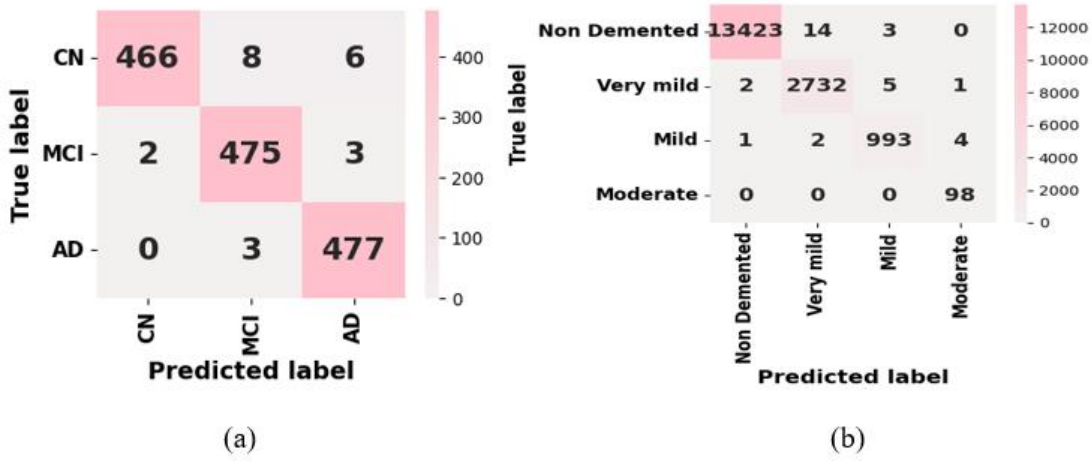

Figure 11: Confusion Matrix for (a) ADNI (b) OASIS datasets

Figure 12 presents the ROC curves and the corresponding AUC values for the proposed model across the ADNI and OASIS-1 validation datasets. These curves map the True Positive against the False Positive Rate to evaluate the overall discriminative strength of the model across all classification thresholds. As shown in Figure 12 (a), the model improves the high-fidelity class separation across the diagnostic range. The individual multi-class breakdown shows an AUC of 0.9844 for CN, 0.9891 for MCI, and 0.9922 for AD. The sharp rise of all three curves towards the upper-left quadrant indicates that the network maintained a high true-positive rate while minimizing false-positive errors across different pathological stages. In Figure 12 (b), the model illustrates robust generalization on a broader four-class multi-class dementia stage task. The model produced an AUC of 0.9990 for non-demented, 0.9980 for Very Mild, 0.9963 for mild, and 0.9990 for Moderate Dementia stages. These elevated AUC metrics prove that the integration of

coordinate-aware routing and block-wise 3D-SAR module prevents macro-structural, natural age-related brain shrinkage from clarifying highly localized, early prodromal variations.

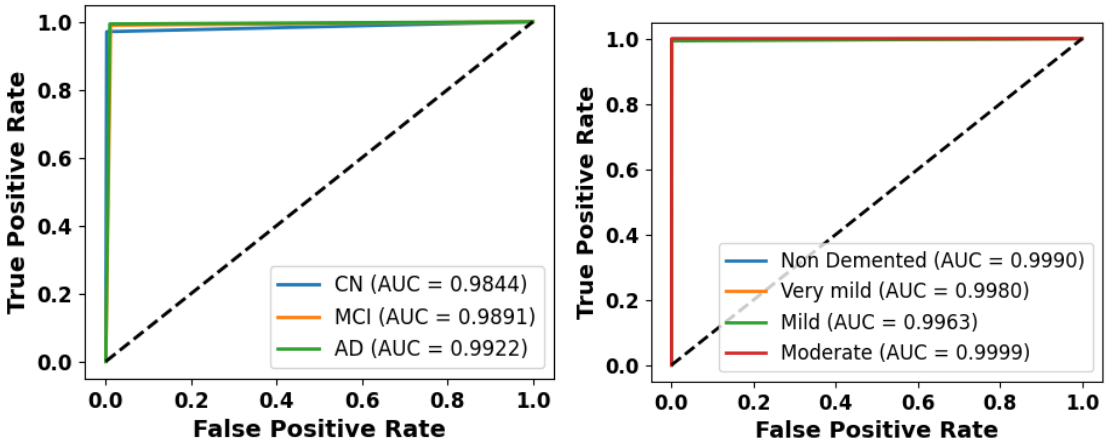

Figure 12: AUC-ROC Curve for proposed method for (a)ADNI (b) OASIS-1 dataset

3.3 *K-fold cross validation*

To ensure the reliability, stability, and generalizability of the proposed framework, a K-fold cross-validation strategy for the model is evaluated. Cross validation is important in medical image analysis to prevent data leakage and mitigate the risks of overfitting, especially given the high dimensionality 3D structural MRI volumes relative to sample sizes. The entire dataset is divided into K equal, non-overlapping subsets. To preserve the distribution of the dataset, a stratified partitioning method, ensuring that the ratio of disease stages remains consistent across all folds. During each iteration, K-1 folds serve as the training set, while the remaining single fold is held out as the independent validation set. This process is repeated K times, rotating the validation fold so that every sample is used for testing exactly once. The performance of the both the baseline architecture and the proposed method is tracked dynamically across each slice. Quantitative metrics are computed for each fold individually through Table 2. The choosing K=5 is the it allocates the maximum possible data to train the data3D Swin Transformer while keeping the massive 3D computational training time manageable. A stratified 5-fold subject-level cross-validation technique is used, ensuring that all scans from the same subject remained within a single fold.

Table 2: K-fold cross-validation of SAR-Swin3D model for both ADNI and OASIS-1

ADNI (CN vs AD)						
K values	Accuracy (%)	Specificity (%)	Sensitivity (%)	Precision (%)	F1-Score (%)	AUC (%)
K=2	95.82	96.10	95.54	96.08	95.81	96.40
K=3	97.10	97.45	96.75	97.43	97.09	97.60
K=4	98.02	98.40	97.62	98.38	98.00	98.50
K=5	**98.47**	**98.90**	**98.47**	**98.49**	**98.47**	**98.95**
K=7	98.52	98.95	98.10	98.94	98.52	99.02
ADNI (sMCI vs pMCI)						
K=2	92.10	93.02	91.15	92.85	91.99	92.50
K=3	93.65	95.40	92.78	94.30	93.53	94.10
K=4	94.52	95.40	93.62	95.20	94.40	95.00
K=5	**95.04**	**95.98**	**94.10**	**95.75**	**94.92**	**95.54**
K=7	95.21	96.12	94.28	95.90	95.08	95.72
OASIS-1						
K=2	98.12	98.30	97.92	98.30	98.11	98.45
K=3	99.20	99.35	99.05	99.35	99.20	99.40
K=4	99.62	99.75	99.50	99.75	99.62	99.78
K=5	**99.81**	**99.90**	**99.72**	**98.44**	**99.06**	**99.91**
K=7	99.85	99.92	99.78	99.92	99.85	99.94

3.4 *Computational Evaluation*

Table 3 evaluates the practical viability of the proposed method by measuring through the computational metrics. A primary challenge in processing high-resolution 3D structural MRI volumes using swin transformer is the abrupt demand on computational hardware and memory footprint. By adopting a 5-fold cross validation, the trade-off is optimizing between maximizing model training exposure and containing overall runtime. 3D-Transformer are prominent for crashing GPUs due to out of memory errors. By reporting the Peak GPU Memory Allocation, the result proves that the proposed model is highly optimized. It shows that the SAR-Swin3D model run on standard accessible GPU hardware.

Table 3: Computational Analysis of the proposed method for both datasets.

Architecture type	Baseline model	Dataset	Trainable parameters (M)	Computational Volume (FLOPs) (G)	Avg. Inference time (per scan) (ms)
Lightweight	EfficientNet –B0	ADNI	5.30	0.39	8.5
		OASIS-1	4.10	0.31	6.1
Standard CNN	ResNet-50	ADNI	25.60	4.12	15.2
		OASIS-1	23.50	3.80	11.4
Transformer baseline	ViT	ADNI	86.4	17.60	38.5
		OASIS-1	85.80	15.20	29.3
Proposed	**SAR-Swin3D**	ADNI	**34.2**	**11.45**	**28.4**
		OASIS-1	**31.85**	**9.12**	**21.8**

3.5 *Statistical Analysis*

The Table 4 verify that the improvements achieved by the SAR-Swin3D model are mathematically correct and not the result of random variations in the 5-fold cross validation, a statistical analysis is conducted. A paired parametric statistical test was applied to compare the classification outputs of the proposed architecture with those of the baseline architectures for statistical evaluation alone. Specifically, two-tailed paired t-tests were performed on the primary evaluation metrics. A p-value threshold of was established as the criterion for statistical significance. This statistical verification ensures that the coordinate-aware routing and structural boundary extraction mechanism introduces a stable and reliable advancement in identifying disease stages.

Table 4: Statistical Evaluation of the proposed method by conducting t-test

ADNI							
Methods	Accuracy (%)	Precision (%)	Sensitivity (%)	Specificity (%)	F1-Score (%)	CI (%)	p-value
EfficientNet –B0	88.42 ± 1.15	87.90±1.42	86.85±1.60	89.10 ± 1.10	87.37 ±1.35	[86.17,90.67]	0.245
ResNet-50	91.15 ± 0.95	90.65±1.12	89.70±1.25	92.15 ± 0.88	90.17±1.05	[89.29,93.01]	0.112
DenseNet-201	93.30 ± 0.78	92.88±0.92	93.20 ±1.05 2.10	94.05 ± 0.72	92.49±0.85	[91.77,94.83]	0.084
ViT	94.85 ± 0.82	94.50±0.98	93.65±1.10	95.55 ± 0.75	94.07±0.90	[93.24,96.46]	0.051

SAR-Swin3D (Proposed)	**98.47 ± 0.35**	**98.49 ±0.28**	**98.47±0.42**	**98.90 ± 0.31**	**98.47±0.33**	**[97.76,99.14]**	0.045
OASIS-1							
EfficientNet –B0	90.12 ± 1.08	89.65±1.30	88.90±1.25	90.85 ± 0.98	89.27±1.20	[88.00,92.24]	0.310
ResNet-50	93.45 ± 0.88	92.95±1.05	92.25±1.00	94.20 ± 0.80	92.60±0.95	[91.73,95.17]	0.185
DenseNet-201	95.18 ± 0.65	94.80±0.78	94.15±0.70	95.90 ± 0.60	94.47±0.72	[93.91,96.45]	0.092
ViT	96.60 ± 0.72	96.25±0.85	95.60±0.78	97.25 ± 0.68	95.92±0.80	[95.19,98.01]	0.063
SAR-Swin3D (Proposed)	**99.81 ± 0.12**	**98.44±0.08**	**99.72±0.09**	**99.06 ± 0.09**	**99.81±0.11**	**[99.57,99.9]**	**0.039**

3.6 *Ablation Study*

To preserve and verify the individual performance gains contributed by each core component of the proposed method, a comprehensive ablation study was conducted across two independent validation benchmarks. The structural complexity of the 3D neuroimaging data evaluates the architecture in a step-by-step configuration, allowing the evaluation of whether the developed choices directly resolve the targeted medical image processing struggles. The baseline configurations consist of a standard 3D Swin Transformer equipped with hierarchical patch merging. Then, we progressively integrate the architectural novelty. First, the 3D structural boundary extraction technique addresses the localized tissue degradation features, followed by the coordinate-aware attention routing module to anchor the global spatial relationship. Table 5 shows that the complete SAR-Swin3D pipeline combines both involvement simultaneously.

Table 5: Ablation analysis of the proposed method by increasing the configurations

Configurations	Dataset	Accuracy (%)	Precision (%)	Sensitivity (%)	F1-Score (%)
Baseline 3D Swin +Hierarchical Merging	ADNI	95.10	95.80	94.30	94.95
	OASIS-1	96.20	96.50	95.90	96.10
+ 3D Structural Boundary extraction	ADNI	96.75	97.10	96.05	96.50
	OASIS-1	97.95	98.10	97.60	97.80
+Coordinate Aware attention routing	ADNI	97.20	97.65	96.80	97.15
	OASIS-1	98.60	98.85	98.30	98.55
SAR-Swin3D (Full Proposed)	ADNI-1	**98.47**	**98.49**	**98.47**	**98.47**
	OASIS-1	**99.81**	**98.44**	**99.72**	**99.06**

3.7 *Comparison Evaluation*

To validate the efficiency and improvements of the proposed model, an extensive comparative analysis of existing methods was performed, as shown in Tables 6, 7, and 8. Evaluation was performed across three different classification tasks using the benchmark ADNI datasets for CN versus AD and sMCI versus pMCI classes. For OASIS-1, all four classes were considered for evaluation, where the proposed method handles both structural divergence and the highly subtle, localized tissue deformations characteristic of early stage prodromal progression. The proposed

method proves that it learns the intrinsic biological features of disease progression rather than just memorizing scanner-specific artifacts.

Table 6: Comparison of proposed with existing method for CN against AD classes

ADNI (CN vs AD)					
Methods	Accuracy (%)	Specificity (%)	Sensitivity (%)	Precision (%)	F1-Score (%)
Dual-3DM3-AD [21]	98.3	97.84	97.4	-	98.00
3D-CNN-MPVT [23]	91.4	93.7	89.6	-	-
E2STN [25]	94.1	98.3	88.4	-	92.7
CNN with multiple 3D angular orientations [26]	94.38	95.17	93.61	95.25	94.42
SAR-Swin3D (Proposed)	**98.47**	**98.90**	**98.47**	**98.49**	**98.47**

Table 7: Comparison of proposed with existing method for sMCI against pMCI classes

ADNI (sMCI vs pMCI)			
Methods	Accuracy (%)	Specificity (%)	Sensitivity (%)
3D-CNN-MPVT [23]	76.0	89.1	72.1
E2STN [25]	94.1	98.3	88.4
SAR-Swin3D (Proposed)	**95.04**	**95.98**	**94.10**

Table 8: Comparison of proposed with existing method for OASIS-1 dataset

OASIS-1				
Methods	Accuracy (%)	Precision (%)	Sensitivity (%)	F1 Score (%)
I-SAB CNN [22]	99.75	99.63	99.91	99.77
MMGLF [27]	86.77	92.42	-	87.33
SAR-Swin3D (Proposed)	**99.81**	**98.44**	**99.72**	**99.06**

3.8 *Discussion*

The proposed framework demonstrated exceptional performance in early stage Alzheimer's disease detection, significantly outperforming baseline and existing models. On the ADNI dataset for the CN versus AD class and sMCI versus pMCI classes, the SAR-Swin3D model improved the classification performance. This proves that the integration of the 3D SAR module successfully addressed the overlapping boundary problem. It effectively isolates subtle localized gray matter tissue loss while preventing natural global age-related brain shrinkage from obscuring key features. Furthermore, the model achieves an accuracy of 99.81% on the OASIS dataset with a lower error rate. The ablation study validated that combining the 3D structural boundary extraction with coordinate-aware attention routing improved individual performance, ensuring that the proposed framework captured intrinsic biological features rather than scanner artifacts.

4. Conclusion

The SAR-Swin3D model introduces a novel technique for the early detection and classification of Alzheimer's disease using volumetric 3D structural MRI scans. By replacing rigid downsampling with a block-wise 3D-SAR module, the architecture effectively resolves the overlapping boundary problem. It isolates subtle gray matter tissue thresholds that differentiate healthy aging from early cognitive decline. The experimental results validate the model performance by producing higher accuracies for both the ADNI and OASIS datasets. These results confirm that coordinate-aware routing successfully prevents macrostructural, age-related brain shrinkage from obscuring localized prodromal variations. Furthermore, the model addresses the computational barriers inherent to volumetric DL models. By employing the 3D-WMSA and 3D-SWMSA methods, the framework reduces the computational complexity of whole-brain 3D modeling from cubic to linear. Future work may focus on expanding the framework into a multimodal

by integrating high-resolution T1-weighted MRIs with multi-tracker metabolic PET grids and fluid specimen biomarkers.

References

1. Hu, Z., Wang, Y. and Xiao, L., 2025. Alzheimer's disease diagnosis by 3D-SEConvNeXt. Journal of Big Data, 12(1), p.15.
2. Rahman, A.U., Ali, S., Saqia, B., Halim, Z., Al-Khasawneh, M.A., AlHammadi, D.A., Khan, M.Z., Ullah, I. and Alharbi, M., 2025. Alzheimer's disease prediction using 3D-CNNs: Intelligent processing of neuroimaging data. SLAS technology, 32, p.100265.
3. Li, C., Gao, Z., Chen, X., Zheng, X., Zhang, X., Lin, C.Y. and Alzheimer's Disease Neuroimaging Initiative, 2025. Ensemble network using oblique coronal MRI for Alzheimer's disease diagnosis. Neuroimage, 310, p.121151.
4. Mmadumbu, A.C., Saeed, F., Ghaleb, F. and Qasem, S.N., 2025. Early detection of Alzheimer's disease using deep learning methods. Alzheimer's & Dementia, 21(5), p.e70175.
5. Huang, Y., Su, Y., Wang, X. and Yao, S., 2025. An adaptive learning framework for Alzheimer's disease diagnosis using structural Magnetic Resonance Imaging data analytics. Decision Analytics Journal, p.100667.
6. Nie, Y., Cui, Q., Li, W., Lü, Y. and Deng, T., 2024. MHAGuideNet: a 3D pre-trained guidance model for Alzheimer's Disease diagnosis using 2D multi-planar sMRI images. BMC Medical Imaging, 24(1), p.338.
7. Chua, J., Li, C., Antochi, F., Toma, E., Wong, D., Tan, B., Garhöfer, G., Hilal, S., Popa-Cherecheanu, A., Chen, C.L.H. and Schmetterer, L., 2025. Utilizing deep learning to predict Alzheimer's disease and mild cognitive impairment with optical coherence tomography. Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring, 17(1), p.e70041.
8. Turrisi, R. and Patané, G., 2026. Generating Synthetic MRI Scans for Improving Alzheimer's Disease Diagnosis. Medical Image Analysis, p.103947.
9. Komal, R., Dhavakumar, P., Rahul, K., Jaswanth, B. and Preeth, R., 2025. Hybrid deep learning framework for magnetic resonance imaging-based classification of Alzheimer's disease. Brain Network Disorders.
10. Jytzler, J.A. and Lysdahlgaard, S., 2024. Radiomics evaluation for the early detection of Alzheimer's dementia using T1-weighted MRI. Radiography, 30(5), pp.1427-1433.
11. I. Khan, P. U. Neetha, T. A. Azhikakathu, D. Somshekhar, Vijetha and Suman, Fusion-Based Deep Architecture Leveraging Convolutional Networks and Vision Transformers for Subtype Discrimination in NSCLC, 2025 IEEE 2nd International Conference on Green Industrial Electronics and Sustainable Technologies, Jamshedpur, India, 2025, pp. 1-6, doi: 10.1109/GIEST66547.2025.11387596.
12. Turrisi, R., Pati, S., Pioggia, G., Tartarisco, G. and Alzheimer's Disease Neuroimaging Initiative, 2025. Adapting to evolving MRI data: A transfer learning approach for Alzheimer's disease prediction. Neuroimage, 307, p.121016.
13. Kina, E., 2025. TLEABLCNN: Brain and Alzheimer's disease detection using attention based explainable deep learning and smote using imbalanced brain MRI. IEEE Access.
14. Kim, S.K., Duong, Q.A. and Gahm, J.K., 2024. Multimodal 3D deep learning for early diagnosis of Alzheimer's disease. IEEE Access, 12, pp.46278-46289.
15. Ramani, R., Ganesh, S.S., Rao, S.S. and Aggarwal, N., 2025. Integrated multi-modal 3D-CNN and RNN approach with transfer learning for early detection of Alzheimer's disease. Iranian Journal of Science and Technology, Transactions of Electrical Engineering, 49(1), pp.383-407.
16. Pradhan, N., Sagar, S. and Singh, A.S., 2024. Analysis of MRI image data for Alzheimer disease detection using deep learning techniques. Multimedia Tools and Applications, 83(6), pp.17729-17752.
17. Asaduzzaman, M., Alom, M.K. and Karim, M.E., 2025. ALZENET: deep learning-based early prediction of Alzheimer's disease through magnetic resonance imaging analysis. Telematics and Informatics Reports, 17, p.100189.
18. Kumari, R., Das, S. and Singh, R.K., 2024. Agglomeration of deep learning networks for classifying binary and multiclass classifications using 3D MRI images for early diagnosis of Alzheimer's disease: a feature-node approach. International Journal of System Assurance Engineering and Management, 15(3), pp.931-949.
19. Suchitra, S., Krishnasamy, L. and Poovaraghan, R.J., 2025. A deep learning-based early alzheimer's disease detection using magnetic resonance images. Multimedia tools and applications, 84(16), pp.16561-16582.
20. Sudheesh, K.V., Puttegowda, K., Naveenkumar, H.N., Chethan, K. and Mahadevaswamy, 2025. Convolution neural Network-Based alzheimer disease detection system using medical image retrieval approach with Multi-Class classification. SN Computer Science, 6(6), p.587.
21. Khan, A.A., Mahendran, R.K., Perumal, K. and Faheem, M., 2024. Dual-3DM 3 AD: mixed transformer based semantic segmentation and triplet pre-processing for early multi-class Alzheimer's diagnosis. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 32, pp.696-707.
22. Tripathy, S.K., Nayak, R.K., Gadupa, K.S., Mishra, R.D., Patel, A.K., Satapathy, S.K., Bhoi, A.K. and Barsocchi, P., 2024. Alzheimer's disease detection via multiscale feature modelling using improved spatial attention guided depth separable CNN. International Journal of Computational Intelligence Systems, 17(1), p.113.
23. Huang, F., Chen, N. and Qiu, A., 2025. 3D-CNN Enhanced Multiscale Progressive Vision Transformer for AD Diagnosis. IEEE Journal of Biomedical and Health Informatics.

24. Ni, H., Xue, J., Qin, J., Zhang, Y. and Alzheimer's Disease Neuroimaging Initiative (ADNI, 2024. Accurate identification of individuals with subjective cognitive decline using 3D regional fractal dimensions on structural magnetic resonance imaging. Computer Methods and Programs in Biomedicine, 254, p.108281.
25. Huang, S. and Dai, Q., 2025. A 3D efficient and essentialized swin transformer network for alzheimer's disease diagnosis. Applied Intelligence, 55(15), p.1003.
26. Uyguroğlu, F., Toygar, Ö. and Demirel, H., 2024. CNN-based Alzheimer's disease classification using fusion of multiple 3D angular orientations. Signal, Image and Video Processing, 18(3), pp.2743-2751.
27. Jia, N., Jia, T., Zhao, L., Ma, B. and Zhu, Z., 2024. Multi-modal global-and local-feature interaction with attention-based mechanism for diagnosis of Alzheimer's disease. Biomedical Signal Processing and Control, 95, p.106404.
28. Kumar, M., Kumar, B., Sharma, P., Sharma, R., Al-Dhaifallah, M. and Shakoor, A., 2025. Attentive deep learning with randomized vector energy least square twin support vector machine for Alzheimer's disease diagnosis. Computers and Electrical Engineering, 126, p.110412.
29. ADNI:https://adni.loni.usc.edu/data-samples/adni-data/neuroimaging/mri/, accessed on June 2026
30. OASIS-1:https://sites.wustl.edu/oasisbrains/home/oasis-1/, accessed on June 2026