# BREAKING THE ATLAS BARRIER: ADAPTIVE SUPERVOXELS FOR INTERPRETABLE ALZHEIMER'S DISEASE DIAGNOSIS ON PET

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Alzheimer's disease (AD) is a progressive neurodegenerative disorder and the leading cause of dementia worldwide. Positron emission tomography (PET) reveals metabolic and molecular abnormalities whose location, extent, and spatial pattern vary substantially across individuals. Existing learning-based methods nevertheless aggregate evidence over predefined slices, patches, or atlas regions that are shared by all subjects. These fixed units misalign with individual abnormalities: they fragment coherent abnormal patterns or mix them with preserved tissue, which limits diagnostic accuracy and makes region-level explanations unreliable when abnormal uptake spans or falls between anatomical boundaries. We propose ASNet, an end-to-end framework that learns the spatial units themselves for PET-based AD diagnosis. Its Adaptive Supervoxel Partitioning (ASP) module learns differentiable voxel-to-supervoxel associations from PET features and spatial coordinates and iteratively refines them for each subject without any atlas. Its Abnormality Evidence Classifier (AEC) measures how far each learned region deviates from a cognitively normal reference and uses this deviation to route the region to an abnormality expert or a preservation expert. The class logits are the exact sum of the resulting region-level contributions, so the explanation is the diagnostic computation itself rather than a post hoc approximation. Across ADNI and AIBL on FDG, amyloid, and tau PET, ASNet attains the highest accuracy in all eight settings and an AUC of up to 96.2%. Substituting fixed atlases for the learned partition lowers AUC by over 10 points, and visual analyses confirm subject-specific, region-level evidence from the learned supervoxels.

## 1 INTRODUCTION

Alzheimer's disease (AD) is a prevalent neurodegenerative disorder whose symptoms, including memory loss and language impairment, affect tens of millions worldwide. Medical imaging is central to computer-aided AD diagnosis because it exposes disease-related brain changes in vivo. Among imaging modalities, positron emission tomography (PET) uses radioactive tracers to image cerebral metabolism and molecular pathology, providing functional and molecular evidence beyond structural morphology (Laforce et al., 2014). AD-related PET abnormalities, however, may span several brain regions or emerge within small localized areas, and their location, extent, and severity vary substantially across individuals (Del Sole et al., 2008). A fixed anatomical parcellation can therefore split a coherent abnormal region or mix abnormal and preserved tissue within the same region. This mismatch limits diagnostic accuracy, because the model must reason over spatial units that do not reflect the underlying abnormality. More critically, it undermines interpretability. When abnormal uptake spans or falls between anatomical regions, region-level evidence no longer corresponds to the actual pathology, and the explanation becomes misleading rather than informative. Fig. 1 contrasts predefined spatial units with the adaptive supervoxels learned by our method.

Research on PET-based AD analysis has progressed from manually selected region-of-interest (ROI) statistics to whole-brain convolutional networks (Rogeau et al., 2024), multiscale feature learning (Lu et al., 2018), data-driven discovery of metabolic subtypes (Ryoo et al., 2024), and PET-driven disease-specific brain maps (Pham et al., 2026). Deep learning methods have further explored Transformers that capture long-range dependencies among 3D patches (Kunanbayev et al.,



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

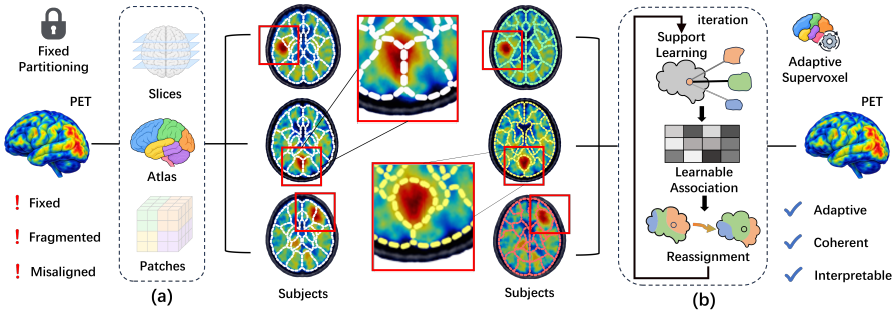



Figure 1: Comparison between fixed and adaptive spatial units. (a) Slices, atlas regions, and patches impose fixed units that may fragment or mismatch coherent PET abnormalities. (b) Support learning and reassignment update voxel-to-supervoxel associations, yielding supervoxels adapted to subject-specific PET patterns. Boxes highlight differences in abnormality coverage.

2024), graph neural networks that model relationships among brain regions (Sim et al., 2024), and hybrid CNN and Transformer models that operate on 2D slices (Bai et al., 2025). Although these approaches strengthen diagnostic representations, most still reason over atlas regions, fixed patches, or image slices. Such units cannot follow abnormalities whose location, extent, and shape vary across subjects, and they therefore inherit the accuracy and interpretability bottleneck described above. Supervoxels, which group adjacent voxels into contiguous 3D regions according to image content (Veksler et al., 2010), offer a natural alternative. Classical supervoxels, however, are computed by handcrafted or unsupervised criteria and remain fixed during learning, and differentiable grouping modules developed for natural images (Jampani et al., 2018; Locatello et al., 2020) are not designed to serve as the units of a diagnostic decision. These observations motivate a different view: the spatial unit of PET diagnosis should be learned for each subject under diagnostic supervision rather than prescribed in advance.

Building on this view, we propose ASNet, an end-to-end framework for interpretable PET-based AD diagnosis that learns its own spatial units. Given a PET volume, the Adaptive Supervoxel Partitioning (ASP) module learns soft voxel-to-supervoxel associations from image features and spatial coordinates and refines them iteratively through bidirectional exchange between voxels and regions, so that the partition follows each subject's uptake pattern without any atlas. The Abnormality Evidence Classifier (AEC) then links the learned regions to the diagnosis. AEC measures the deviation of each supervoxel from a cognitively normal (CN) reference and uses this deviation to route the region to an abnormality expert or a preservation expert, which separates abnormal from relatively preserved patterns even when abnormalities are subtle or spatially mixed. The selected expert outputs are summed into the class logits, so the logit margin decomposes exactly into region-level contributions, and the resulting contribution map is the diagnostic computation itself rather than a post hoc approximation. Because ASP and AEC are optimized jointly, diagnostic supervision shapes the partition itself, and the learned supervoxels become task-specific and subject-adaptive rather than generic image segments. Our contributions are summarized as follows.

- We identify fixed spatial units, including atlas regions, patches, and slices, as a fundamental bottleneck for both accuracy and interpretability in PET-based AD diagnosis, and we formulate the spatial partition as a learnable, subject-specific variable. ASNet realizes this idea through atlas-free supervoxel learning under diagnostic supervision.
- We introduce ASP, a differentiable 3D partitioning module that refines voxel-to-supervoxel associations through bidirectional voxel and region updates, and AEC, a reference-anchored classifier whose hard deviation gate decouples routing from evidence estimation and whose additive design makes region-level contributions faithful to the prediction by construction.
- Across FDG, amyloid, and tau PET, ASNet achieves the highest accuracy in all eight evaluation settings and ranks first or second on every metric, with an AUC of up to 96.2%. Ablations show that replacing the learned partition with standard atlases reduces AUC by more than 10 percentage points, and visual analyses show how the learned supervoxels and their contributions form subject-specific evidence.

## 2 Related Work

**Deep Learning Models for AD Diagnosis.** Existing methods differ mainly in the spatial unit over which they aggregate evidence. Grid-based models operate on the voxel lattice, patches, or slices: volumetric networks learn uptake patterns from PET (Rogeau et al., 2024) or capture global context by Fourier filtering (Zhang et al., 2022), multiple-instance frameworks pool patch and slice evidence (Zhu et al., 2021; Liu et al., 2023; Huang et al., 2025a), and Transformers aggregate multiplane or view-specific slices (Jang & Hwang, 2022; Wang et al., 2026). Region-based models parcellate the brain with a predefined atlas and reason over region graphs (Xiao et al., 2022; Liu et al., 2024), a formulation adopted by recent PET frameworks for ROI-graph classification (Kim et al., 2025), multimodal fusion (Hu et al., 2025), cross-modal distillation (Liu et al., 2025), and hierarchical anatomical features (Chen et al., 2025), and refined by subject-dependent prototypes and learned edges (Leng et al., 2023; Cui et al., 2025). A parallel line broadens the input instead, through incomplete multitracer learning (Ou et al., 2024), tau PET fusion (Huang et al., 2025b), longitudinal generative modeling (Gao et al., 2025), and whole-body PET (Li et al., 2025; Zhao et al., 2026). Both lines fix shared spatial units and only reweight their features. PET abnormalities vary in location, extent, and shape, often crossing anatomical boundaries, so fixed units fragment or dilute them, limiting subject-specific recognition.

**Interpretability in AD Diagnosis.** Interpretable AD diagnosis follows two strategies. Built-in designs derive evidence from the model's own computation: they gate atlas regions (Jiang & Miao, 2024), learn discriminative regional relations (Xiao et al., 2022), pool hub regions in atlas-defined networks (Jia et al., 2026), emit risk or counterfactual maps (Wu et al., 2024; Oh et al., 2023), localize pathology iteratively (Li et al., 2019), trace predictions to multi-scale patches (Ye et al., 2024), align attention with class logits (Xu et al., 2024), match prototypes on surface patches or slices (Bongratz et al., 2025; Diao et al., 2025), or expose expert routing (Jiang et al., 2026). Post hoc methods instead attribute a trained prediction using gradients (Selvaraju et al., 2017), Shapley values (Lundberg & Lee, 2017), or relevance propagation (Achtibat et al., 2024), and have been applied to visualize class-relevant locations (Mercaldo et al., 2024), compare saliency with clinically recognized disease regions (Tam et al., 2024), quantify overlap with atlas structures (Diao et al., 2025), and rank modality contributions (Huang et al., 2025b). Both strategies inherit the underlying units, anchoring explanations to atlas parcels, patches, or slices, while evidence from attention, localization, or attribution remains loosely coupled to class scores. This is acute in PET, where limited resolution and smooth uptake yield diffuse maps lacking clear regional extent, and summarizing them over atlas parcels reinstates boundaries that misalign with individual abnormalities. Faithful interpretation thus requires regional units following each subject's metabolic pattern and contributions explicitly aggregated into the prediction.

**Supervoxels and Learnable Spatial Grouping.** Supervoxels group adjacent, similar voxels into contiguous regions (Veksler et al., 2010) and have served as region-level units for tumor segmentation (Irving et al., 2014; Hansen et al., 2021) and deformable registration (Heinrich et al., 2016; Pei et al., 2018), but they follow handcrafted or unsupervised criteria and stay fixed during downstream learning. Differentiable grouping learns soft associations end to end (Jampani et al., 2018), as do slot attention (Locatello et al., 2020) and super token sampling (Huang et al., 2023), yet these modules target 2D natural images and treat the groups as intermediate representations for object discovery or efficiency. ASNet instead learns a 3D partition under diagnostic supervision and makes the learned regions the units over which the prediction decomposes, each carrying an explicit contribution anchored to a normative reference.

## 3 Method

As illustrated in Fig. 2, ASNet consists of an ASP module (Sec. 3.1) and an AEC (Sec. 3.2). Given a registered 3D PET volume, ASP initializes supervoxels from spatial coordinates and iteratively refines the voxel-to-supervoxel associations and features. AEC compares each learned supervoxel with a CN reference, routes it to one of two evidence experts, and sums the resulting regional contributions into the class logits. The staged optimization in Sec. 3.3 first establishes discriminative supervoxel representations and then learns the contribution-based classifier on top of them. Throughout, bold lowercase letters denote vectors, bold uppercase letters denote matrices, and the superscript $(t)$ indexes ASP iterations. Appendix B summarizes the notation.

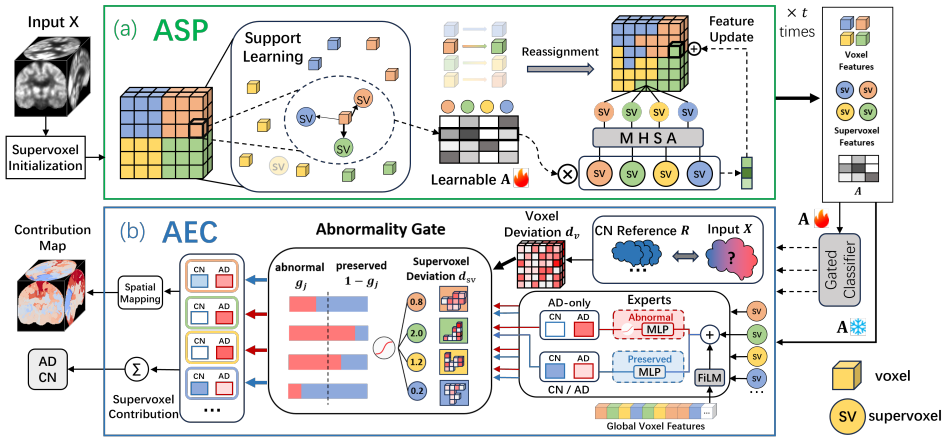



Figure 2: **Overview of ASNet.** (a) ASP initializes supervoxels and refines voxel/supervoxel features via support learning, reassignment, self-attention, and feedback. (b) AEC compares learned regions with CN references; an abnormality gate selects expert evidence to form regional contributions for diagnosis and a decision-linked map. The Gated Classifier is used only during training; flame/snowflake denote trainable/frozen ASP states during staged optimization.

### 3.1 Adaptive Supervoxel Partitioning (ASP)

**Initialization.** Let $\Omega$ denote the set of voxels inside the intracranial mask, indexed by $i$, and let $\mathbf{p}_i$ be the normalized 3D coordinate of voxel $i$. We apply MiniBatch K-Means to $\{\mathbf{p}_i\}_{i\in\Omega}$ to obtain $M$ clusters, whose centers serve as the initial supervoxel centers $\mathbf{c}_j^{(0)}$ for $j = 1, \ldots, M$. This shared spatial initialization provides anchors for association learning rather than a fixed set of anatomical regions. A 3D convolutional stem extracts voxel features $\mathbf{v}_i^{(0)}$, and their projections are aggregated within each cluster to obtain the initial supervoxel features $\mathbf{s}_j^{(0)}$.

**Support learning and reassignment.** At iteration $t$, each voxel $i$ is described by a feature $\mathbf{v}_i^{(t)}$ and a coordinate $\mathbf{p}_i$, and each supervoxel $j$ is described by a feature $\mathbf{s}_j^{(t)}$ and a center $\mathbf{c}_j^{(t)}$. For voxel $i$, we consider a candidate set $\mathcal{N}_i^{(t)}$ of spatially proximal supervoxels that may support its representation. The support score combines feature compatibility with spatial proximity, and normalizing it over the candidates yields the association weight

$$\ell_{ij}^{(t)} = \frac{\left(W_q \mathbf{v}_i^{(t)}\right)^\top \left(W_k \mathbf{s}_j^{(t)}\right)}{\sqrt{d_k}} - \lambda \left\| \mathbf{p}_i - \mathbf{c}_j^{(t)} \right\|_2^2, \quad a_{ij}^{(t)} = \frac{\exp\left(\ell_{ij}^{(t)}\right)}{\sum_{m\in\mathcal{N}_i^{(t)}} \exp\left(\ell_{im}^{(t)}\right)}, \quad j \in \mathcal{N}_i^{(t)}, \tag{1}$$

where $W_q$ and $W_k$ project voxel and supervoxel features into a shared $d_k$-dimensional space and $\lambda$ controls the strength of the spatial prior. We set $a_{ij}^{(t)} = 0$ for $j \notin \mathcal{N}_i^{(t)}$. Because Eq. 1 is recomputed at every iteration from the current features, the initial spatial assignment is progressively revised, and the partition follows subject-specific PET patterns rather than the initial clustering.

**Bidirectional update.** The weights form the voxel-to-supervoxel association matrix $\mathbf{A}^{(t)} \in \mathbb{R}^{|\Omega|\times M}$ with $[\mathbf{A}^{(t)}]_{ij} = a_{ij}^{(t)}$. This matrix mediates information exchange in both directions. In the voxel-to-supervoxel direction, it collects voxel features to update each regional state. In the supervoxel-to-voxel direction, it distributes contextualized regional states back to the voxels:

$$\mathbf{s}_j^{(t+1)} = \mathbf{s}_j^{(t)} + f_s\left(\frac{\sum_{i\in\Omega}[\mathbf{A}^{(t)}]_{ij} W_v \mathbf{v}_i^{(t)}}{\sum_{i\in\Omega}[\mathbf{A}^{(t)}]_{ij} + \epsilon}\right), \quad \mathbf{v}_i^{(t+1)} = \mathbf{v}_i^{(t)} + \sum_{j\in\mathcal{N}_i^{(t)}} [\mathbf{A}^{(t)}]_{ij} W_f \bar{\mathbf{s}}_j^{(t)}, \tag{2}$$

where $W_v$ and $W_f$ are learnable projections, $f_s(\cdot)$ is a learnable update function, $\epsilon$ is a small constant for numerical stability, and $\bar{\mathbf{s}}_j^{(t)}$ denotes the supervoxel features contextualized by multi-head self-attention over the regional states. The centers follow the same weighting, $\mathbf{c}_j^{(t+1)} =$

$\sum_{i\in\Omega} a_{ij}^{(t)}\mathbf{p}_i/(\sum_{i\in\Omega} a_{ij}^{(t)}+\epsilon)$. Since both directions share $\mathbf{A}^{(t)}$, the regional states, voxel features, and associations are refined jointly, and gradients from the diagnostic loss reach the partition through the soft weights. After $T$ iterations, we write $\mathbf{s}_j = \mathbf{s}_j^{(T)}$ and assign each voxel to its most strongly associated supervoxel, $\Omega_j = \{i \in \Omega : j = \arg\max_m a_{im}^{(T)}\}$. This hard assignment is used only for regional statistics and visualization.

### 3.2 Abnormality Evidence Classifier (AEC)

**Normative deviation.** AEC characterizes each supervoxel relative to a CN reference $\mathcal{R} = \{\mu_i^{\mathrm{CN}}, \sigma_i^{\mathrm{CN}}\}_{i\in\Omega}$, the voxel-wise mean and standard deviation estimated from the CN subjects in the training set. For a subject with intensity $x_i$, the voxel deviation and the supervoxel deviation are

$$d_i^{\mathrm{v}} = \frac{\left|x_i - \mu_i^{\mathrm{CN}}\right|}{\sigma_i^{\mathrm{CN}} + \epsilon}, \qquad d_j^{\mathrm{sv}} = \frac{1}{|\Omega_j|}\sum_{i\in\Omega_j} d_i^{\mathrm{v}}. \tag{3}$$

The supervoxel deviation defines a normative state that selects the evidence expert for each region.

**Abnormality gate.** The deviation $d_j^{\mathrm{sv}}$ is mapped to an abnormality score and a straight-through hard gate,

$$p_j = \mathrm{sigmoid}\left(\alpha\left(d_j^{\mathrm{sv}} - \tau\right)\right), \qquad g_j = p_j + \mathrm{sg}\left(\mathbb{I}\left[d_j^{\mathrm{sv}} > \tau\right] - p_j\right), \tag{4}$$

where $\tau$ is the reference-based abnormality threshold, $\alpha$ controls the sharpness of the transition, and $\mathrm{sg}(\cdot)$ denotes the stop-gradient operator. In the forward pass, $g_j \in \{0, 1\}$, and $g_j = 1$ exactly when $p_j > 0.5$. The gate thus selects the abnormal path ($g_j = 1$) or the preserved path ($g_j = 0$) on the basis of the normative deviation alone, while the evidence value is estimated from the learned supervoxel representation. Decoupling routing from evidence prevents the routing decision from being dominated by the learned features and keeps the interpretation anchored to the CN reference.

**Context-conditioned experts.** To condition evidence on the global brain state, we derive a whole-brain context vector $\mathbf{u}$ from the final voxel features and apply FiLM (Perez et al., 2018) to each supervoxel representation. The abnormal and preserved experts then produce their evidence:

$$\widetilde{\mathbf{s}}_j = \mathbf{s}_j \odot [1 + \tanh(\gamma(\mathbf{u}))] + \beta(\mathbf{u}), \qquad \mathbf{e}_j^A = [0,\ \mathrm{softplus}(f_A(\widetilde{\mathbf{s}}_j))], \qquad \mathbf{e}_j^P = f_P(\widetilde{\mathbf{s}}_j), \tag{5}$$

where $\gamma(\cdot)$ and $\beta(\cdot)$ produce the scale and shift, $f_A(\cdot)$ and $f_P(\cdot)$ are the abnormal and preserved expert networks, and the two entries of each evidence vector correspond to the (CN, AD) classes. The abnormal expert can only add non-negative evidence to the AD class, whereas the preserved expert produces evidence for both classes. This asymmetry encodes a monotonic prior: deviation from the CN reference can only increase the evidence for AD, whereas a relatively preserved region may support either class according to its learned representation.

**Additive contributions.** Let $\mathcal{M} \subseteq \{1, \ldots, M\}$ denote the set of valid supervoxels, namely those with $|\Omega_j| > 0$. The class-wise contribution of supervoxel $j$, the class logits, and the regional contribution score are

$$\mathbf{C}_j = \frac{g_j\mathbf{e}_j^A + (1-g_j)\mathbf{e}_j^P}{|\mathcal{M}|}, \qquad \mathbf{z} = \sum_{j\in\mathcal{M}} \mathbf{C}_j, \qquad r_j = C_{j,\mathrm{AD}} - C_{j,\mathrm{CN}}. \tag{6}$$

The prediction is obtained from $\mathbf{z}$, and by construction the logit margin decomposes exactly over the learned regions, $z_{\mathrm{AD}} - z_{\mathrm{CN}} = \sum_{j\in\mathcal{M}} r_j$. Mapping $r_j$ back to the voxels in $\Omega_j$ yields the region-level contribution map. This additive structure is related to neural additive models (Agarwal et al., 2021), which decompose a prediction over input features, whereas AEC decomposes it over learned, subject-specific regions. Unlike post hoc attribution, which estimates contributions after the fact and depends on a gradient path, a baseline, or a perturbation scheme, the map consists of the very quantities that are summed to form the prediction and requires no additional computation.

### 3.3 Staged Optimization

ASNet is trained in three stages to establish discriminative supervoxel representations before learning their contribution-based interpretation. Stage 1 trains ASP with an auxiliary Gated Classifier

Table 1: Comparison with existing methods on ADNI and AIBL. Results are mean±std (%) over five folds. The best and second-best results are in **bold** and underlined, respectively.

| Method | ADNI | | | | | | AIBL | |
|---|---|---|---|---|---|---|---|---|
| | $^{18}$F-FDG | | $^{18}$F-AV1451 | | $^{18}$F-AV45 | | Amyloid PET | |
| | AUC | ACC | AUC | ACC | AUC | ACC | AUC | ACC |
| *AD vs. CN* | | | | | | | | |
| GF-Net | 90.35±3.09 | 85.59±5.59 | 84.71±3.85 | 81.18±4.40 | 84.80±2.56 | 80.56±2.15 | 87.51±3.88 | 80.46±2.88 |
| M3T | 95.16±2.56 | 87.10±4.90 | 90.90±3.75 | 84.71±7.53 | 90.35±5.36 | 83.33±4.39 | 88.52±4.36 | 83.41±5.54 |
| MCENN | 92.07±3.77 | 88.82±4.00 | 89.83±6.26 | 82.35±6.44 | 86.90±2.46 | 83.06±2.22 | 86.25±4.24 | 82.69±5.17 |
| LEAR | 91.00±4.51 | 84.09±4.78 | 86.16±5.79 | 78.24±3.53 | 89.10±4.26 | 84.44±3.66 | 86.88±7.48 | 83.08±7.13 |
| MAD-Former | 91.49±3.57 | 84.09±4.58 | 84.01±8.12 | 75.88±7.76 | 88.88±1.77 | 82.50±4.08 | 90.57±4.46 | 83.39±4.54 |
| LA-GMF | 95.04±3.53 | 89.25±5.73 | 91.49±4.58 | 84.12±6.86 | 92.38±2.69 | 85.83±5.37 | 91.82±3.81 | 83.04±5.08 |
| AAGN | 95.19±2.53 | 89.03±4.20 | 91.70±3.64 | 84.71±4.83 | 92.28±2.30 | 85.83±4.33 | 89.70±2.87 | 83.78±2.61 |
| 3DSC-TF | 94.98±1.57 | 87.96±2.39 | 91.56±2.12 | 82.94±3.90 | **92.78±1.84** | 84.17±2.58 | 90.26±3.66 | 84.15±4.67 |
| PRGNN | 94.44±2.69 | 86.88±4.83 | 90.84±3.98 | 85.88±4.83 | 89.01±1.91 | 83.33±5.73 | 89.31±1.93 | 81.93±3.14 |
| **ASNet (Ours)** | **96.20±2.53** | **90.54±2.09** | **94.12±3.28** | **89.41±2.35** | 92.68±1.33 | **86.67±2.58** | **93.15±2.82** | **85.19±3.51** |
| *MCI vs. CN* | | | | | | | | |
| MCENN | 64.82±6.43 | 59.46±6.27 | 69.40±6.73 | 64.00±6.78 | 62.15±4.24 | 57.69±5.04 | 68.24±3.17 | 65.20±0.76 |
| LEAR | 57.38±4.31 | 58.38±2.57 | 65.42±5.71 | 61.50±4.42 | 63.49±4.20 | 60.77±3.85 | 71.19±4.08 | 67.90±2.48 |
| MAD-Former | 59.10±6.35 | 59.10±1.46 | 65.41±4.46 | 62.33±3.09 | 64.37±7.60 | 60.51±8.75 | 71.03±6.71 | 66.21±5.18 |
| LA-GMF | 58.07±3.03 | 57.30±1.67 | 71.80±3.46 | 66.50±3.47 | 57.34±5.32 | 54.36±3.30 | 64.94±14.75 | 61.15±9.29 |
| AAGN | 65.41±3.16 | 60.54±5.68 | 71.18±3.99 | 65.17±3.56 | 64.06±3.00 | 62.56±6.11 | 71.21±9.05 | 63.54±7.56 |
| 3DSC-TF | 66.43±2.79 | 62.34±2.01 | 72.38±2.76 | 66.00±1.22 | **65.64±3.76** | 61.79±3.84 | 71.55±7.68 | 65.89±7.14 |
| PRGNN | 63.98±3.96 | 61.26±7.35 | 72.75±4.23 | 66.33±4.43 | 63.05±4.07 | 62.31±7.83 | 68.44±6.30 | 62.18±3.80 |
| **ASNet (Ours)** | **67.30±1.52** | **64.50±3.81** | **73.35±4.64** | **68.50±3.74** | 64.41±2.61 | **63.08±2.62** | **72.94±4.90** | **68.59±4.44** |

Table 2: Component and atlas ablations on ADNI (mean±std, %). HOA, BNA, and AAL replace ASP.

| Variant | AUC | ACC |
|---|---|---|
| AAL | 83.75±3.91 | 75.27±6.49 |
| BNA | 85.63±2.94 | 78.49±2.26 |
| HOA | 85.86±2.53 | 79.14±3.30 |
| ASP+G-Cls | 94.14±2.61 | 87.53±1.75 |
| ASP+AEC | 96.07±2.69 | 89.68±3.64 |
| **ASNet** | **96.20±2.53** | **90.54±2.09** |

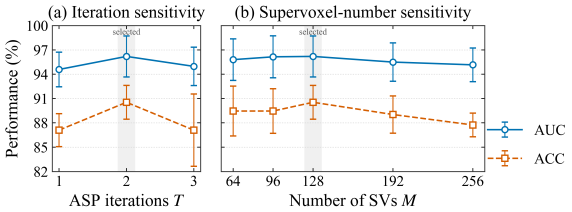



Figure 3: Sensitivity to (a) ASP iterations $T$ and (b) the number of supervoxels $M$ on ADNI FDG PET for AD vs. CN. Points and error bars show the mean and standard deviation over five folds. Shaded areas mark the default settings.

(G-Cls) under diagnostic supervision, which gives the partition and the supervoxel features an initial discriminative signal. Stage 2 freezes ASP and trains AEC to transform supervoxel representations into class-wise regional contributions, using a classification loss and a distillation loss that keep the AEC prediction consistent with the model learned in Stage 1. Stage 3 fine-tunes all components jointly, so the partition and the contribution-based classifier adapt to each other. G-Cls is used only during training. At inference, AEC alone produces both the diagnosis and the contribution map.

# 4 Experiments

## 4.1 Experimental Setup

**Datasets.** We evaluate ASNet on PET data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) and the Australian Imaging, Biomarkers and Lifestyle (AIBL) study. The ADNI cohort includes three tracers that capture complementary aspects of AD pathology: $^{18}$F-FDG for glucose metabolism, $^{18}$F-AV45 for amyloid, and $^{18}$F-AV1451 for tau. For each tracer, we construct two binary tasks, AD vs. CN and mild cognitive impairment (MCI) vs. CN. The AIBL cohort contains amyloid PET acquired with $^{11}$C-PiB, $^{18}$F-AV45, and $^{18}$F-Flutemetamol. Since all three tracers target amyloid pathology, we pool them into a single multi-tracer amyloid cohort and consider the same two tasks. This yields four PET settings and eight evaluation settings in total. Data preprocessing and sample statistics are summarized in Appendix A.

**Training and evaluation.** ASNet is implemented in PyTorch and optimized with AdamW (batch size 2, weight decay $10^{-4}$). The three training stages run for 50, 30, and 10 epochs, respectively. The learning rate is $10^{-4}$ and is reduced to $10^{-5}$ for ASP in Stage 3. We use $M = 128$ supervoxels and $T = 2$ ASP iterations. The CN reference $\mathcal{R}$ is computed only from the training subjects of each fold. We report AUC and ACC, averaged over five-fold cross-validation and reported as mean$\pm$std.

**Baselines.** We compare ASNet with representative methods grouped by spatial unit. Grid-based methods, operating on whole volumes, fixed patches, or slices, include GF-Net (Zhang et al., 2022), M3T (Jang & Hwang, 2022), MAD-Former (Ye et al., 2024), 3DSC-TF (Qiang et al., 2025), LA-GMF (Xu et al., 2024), and MCENN (Liu et al., 2023). Atlas-based methods include AAGN (Jiang & Miao, 2024) and PRGNN (Kim et al., 2025). We additionally include LEAR (Oh et al., 2023), which reinforces the classifier with counterfactual guidance. All methods are trained and evaluated with identical preprocessing and cross-validation splits.

### 4.2 Diagnostic Performance

**Overall comparison.** Table 1 reports results on all eight evaluation settings. ASNet achieves the best ACC in every setting and the best AUC in six of the eight, and it ranks first or second on every metric. No baseline shows comparable consistency. The strongest competitor changes from setting to setting: AAGN is second in AUC on FDG PET for AD vs. CN, 3DSC-TF leads in AUC on ADNI amyloid PET, and LA-GMF is second in AUC on AIBL. This pattern suggests that each fixed spatial unit suits some tracers better than others, whereas a learned partition adapts to each setting. ASNet also shows low fold-to-fold variability, with the lowest ACC standard deviation on FDG and tau PET for AD vs. CN (2.09 and 2.35) and the lowest AUC standard deviation on ADNI amyloid PET (1.33).

**Effect of tracer.** For AD vs. CN, the largest margin appears on tau PET, where ASNet improves AUC from 91.70 (AAGN) to 94.12 and ACC from 85.88 (PRGNN) to 89.41. The gains are smaller but consistent on FDG PET (AUC 96.20 vs. 95.19) and on the multi-tracer AIBL cohort (AUC 93.15 vs. 91.82), which shows that the learned partition transfers across tracers and acquisition protocols. On ADNI amyloid PET, 3DSC-TF attains a marginally higher AUC (92.78 vs. 92.68), a difference well within one standard deviation, while ASNet retains the best ACC (86.67 vs. 85.83). We attribute this ordering to the spatial characteristics of each tracer. Tau deposition tends to be focal and heterogeneous across individuals, which is precisely where fixed units are most likely to misalign with the abnormality. Amyloid retention is more diffuse across the cortex, which leaves less room for gains from adaptive partitioning.

**MCI vs. CN.** Distinguishing MCI from CN is considerably harder for all methods, as expected for prodromal changes that are subtle and heterogeneous. ASNet again achieves the best ACC in all four settings, with gains of 2.16 and 2.00 points on FDG and tau PET, respectively, and the best AUC in three settings. On ADNI amyloid PET, 3DSC-TF yields a higher AUC (65.64 vs. 64.41), consistent with the AD vs. CN results on the same tracer. Several baselines are unstable on this task, for example LA-GMF on AIBL (AUC 64.94$\pm$14.75), whereas the standard deviations of ASNet remain below 5 points in every setting.

### 4.3 Ablation and Sensitivity Analysis

**Learned versus atlas partitions.** Table 2 isolates the contribution of each component. Replacing ASP with the Automated Anatomical Labeling atlas (AAL) (Tzourio-Mazoyer et al., 2002), the Brainnetome atlas (BNA) (Fan et al., 2016), or the Harvard-Oxford atlas (HOA) (Desikan et al., 2006) while keeping AEC reduces AUC to 83.75–85.86 and ACC to 75.27–79.14, over 10 points below ASP+AEC on both metrics. This gap is not explained by granularity: BNA contains 246 regions, more than the 128 supervoxels used by ASNet, yet performs far below the learned partition. Although the three atlases vary in granularity and boundary criteria, all fall well short of ASP. Thus, the gap stems from the fixed, population-level design of atlas parcels rather than from any specific atlas, directly supporting the central premise of this work.

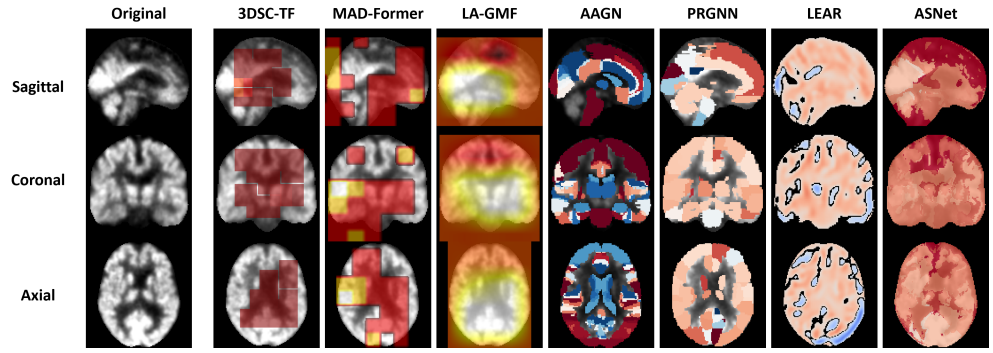



Figure 4: Visual interpretations for a representative AD subject on FDG PET; colors are not quantitatively comparable across methods due to differing interpretation representations.

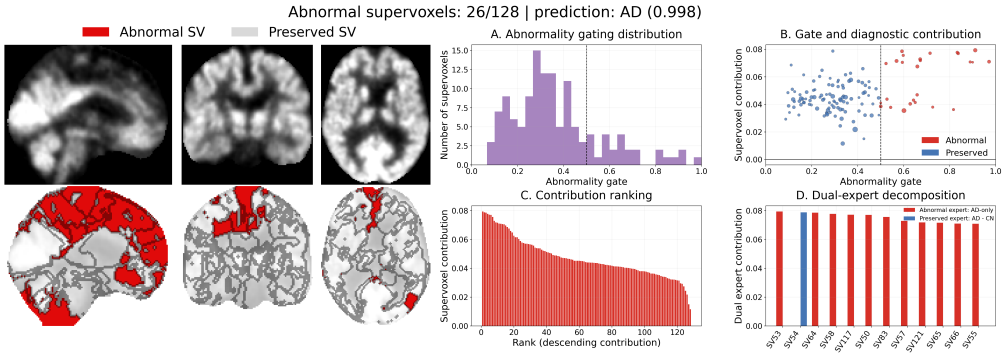



Figure 5: Abnormality gating and regional contributions for the subject in Fig. 4. Left: PET slices with supervoxel assignments; red denotes abnormal-gated supervoxels and gray relatively preserved ones. Right: (A) abnormality score $p_j$ distribution, (B) $p_j$ versus regional contribution $r_j$, (C) ranked contributions across supervoxels, and (D) top-ranked contributions and their producing experts. Dashed lines in (A) and (B) mark the gate threshold $p_j = 0.5$.

**Evidence classifier and training.** On the same ASP backbone, replacing the auxiliary G-Cls head with AEC raises AUC from 94.14 to 96.07 and ACC from 87.53 to 89.68. The reference-anchored, additive classifier therefore improves accuracy rather than trading it for interpretability. The full ASNet, which additionally adopts the staged optimization of Sec. 3.3, yields a further gain that is most visible in ACC (90.54) and in its lower variance (2.09 vs. 3.64).

**Sensitivity.** Fig. 3 examines the two main hyperparameters of ASP. Both AUC and ACC peak at $T = 2$. A single iteration leaves the partition close to its spatial initialization, whereas a third iteration slightly lowers the mean and widens the spread across folds, which suggests diminishing returns from further reassignment. For the number of supervoxels, AUC varies by only about one point between $M = 64$ and $M = 256$, while ACC peaks at $M = 128$ and decreases for larger $M$. A likely reason is that smaller supervoxels contain fewer voxels, which makes their regional deviation in Eq. 3 noisier. ASNet is therefore robust to the choice of $M$ over a wide range, and we use $M = 128$ and $T = 2$ in all experiments.

### 4.4 Interpretability Analysis

**Comparison of explanation forms.** Fig. 4 compares the explanations of ASNet with those of patch-based, atlas-based, and counterfactual methods for a representative AD subject on FDG PET. The form of each explanation is dictated by the spatial unit of the underlying model. 3DSC-TF and MAD-Former produce blockwise responses whose extent follows the patch grid, and LA-GMF yields a smooth map that spreads over most of the brain without a clear regional extent. AAGN and PRGNN assign scores to atlas parcels, so the boundaries of the highlighted areas coincide with template boundaries regardless of the subject's uptake. The counterfactual map of LEAR concentrates along tissue boundaries and appears fragmented. ASNet instead assigns a contribution to each learned supervoxel, whose boundaries follow neither a grid nor a template. More importantly, the

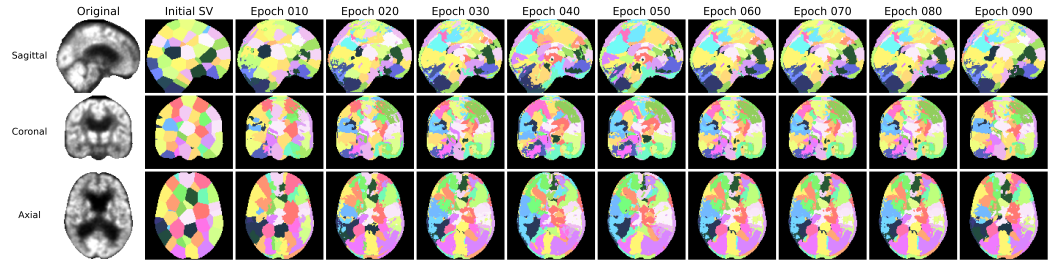



Figure 6: Supervoxel evolution across epochs for a representative sample. Rows show sagittal, coronal, and axial views. Colors distinguish supervoxels; more results are in the Appendix C.

values shown for ASNet are the regional terms $r_j$ that sum to the logit margin in Eq. 6, whereas the maps of the other methods are produced by attention, attribution, or counterfactual procedures whose relation to the class scores is indirect.

**Decomposing an individual prediction.** Fig. 5 provides a detailed view of the decision for the same subject, which ASNet classifies as AD with a probability of 0.998. In panel (A), abnormality scores concentrate around 0.3 with a tail above the threshold, and the hard gate routes 26 of the 128 supervoxels to the abnormal expert. In the sagittal view, these supervoxels form spatially contiguous cortical clusters rather than scattered voxels. Panel (B) shows that the gate governs expert selection rather than evidence magnitude: abnormal-gated supervoxels span a wide range of contributions, and several preserved supervoxels contribute as much as abnormal ones. Panel (C) shows that all contributions are positive and decay gradually with rank, indicating that the AD decision rests on a distributed set of regions rather than a single dominant one. Panel (D) confirms this pattern: among the 12 top-ranked supervoxels, 11 are routed to the abnormal expert and one to the preserved expert with a comparable contribution. Since $\sum_j r_j$ equals the logit margin, each bar represents a share of the final decision, yielding a region-level account of the prediction that attention maps and post hoc attributions do not provide.

**Evolution of the learned partition.** Fig. 6 shows how the partition of a representative subject evolves during training. Coordinate-based initialization produces compact cells of roughly uniform size. As training proceeds, these cells develop into regions of varied sizes and irregular shapes, while the partition becomes more stable in later epochs. This evolution shows how ASP adaptively reshapes supervoxels according to the subject's PET features, providing individualized regions for diagnosis and interpretation.

## 5 Conclusion

We presented ASNet, a framework for interpretable PET-based AD diagnosis that replaces fixed anatomical atlases with adaptive supervoxels learned under diagnostic supervision. ASP refines voxel-to-supervoxel associations to follow subject-specific uptake patterns, and AEC converts each learned region into explicit additive contributions. Across three tracers, two cohorts, and two diagnostic tasks, ASNet achieves the highest accuracy in all eight settings, and ablations show that the learned partition, rather than any particular atlas or finer granularity, drives most of this gain. The central implication is that the spatial unit itself, not only the architecture or attention mechanism, is a critical design choice for PET-based diagnosis. By learning the partition under diagnostic supervision, ASNet aligns regional evidence with the abnormality that drives each prediction rather than with boundaries derived from structural imaging of reference populations. Because each contribution is defined relative to a common normative reference, the resulting evidence supports comparison across subjects with different uptake patterns and across time points of the same subject, which is essential for longitudinal monitoring and may help reveal early or atypical presentations that fixed atlases tend to obscure. Future work will validate ASNet on larger, multi-center, and more diverse cohorts, quantitatively assess learned contributions against expert annotations and clinical outcomes, and extend adaptive supervoxel learning to multimodal PET and MRI.

### AI USE STATEMENT

In preparing this manuscript, generative AI tools were used to assist with LaTeX formatting and language editing. No generative AI tool was used to generate experimental data or to replace the authors' verification of the methods, results, claims, and references. The authors reviewed all AI-assisted material and take responsibility for the final content of this work.

## REFERENCES

Reduan Achtibat, Sayed Mohammad Vakilzadeh Hatefi, Maximilian Dreyer, Aakriti Jain, Thomas Wiegand, Sebastian Lapuschkin, and Wojciech Samek. AttnLRP: Attention-aware layer-wise relevance propagation for transformers. In *International Conference on Machine Learning (ICML)*, pp. 135–168, 2024.

Rishabh Agarwal, Levi Melnick, Nicholas Frosst, Xuezhou Zhang, Ben Lengerich, Rich Caruana, and Geoffrey E. Hinton. Neural additive models: Interpretable machine learning with neural nets. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 4699–4711, 2021.

Jing Bai, Zhengyang Zhang, Yue Yin, Weikang Jin, Talal Ahmed Ali Ali, Yong Xiong, and Zhu Xiao. LGG-NeXt: A next generation CNN and transformer hybrid model for the diagnosis of Alzheimer's disease using 2D structural MRI. *IEEE Journal of Biomedical and Health Informatics*, 29(4):2808–2818, 2025.

Fabian Bongratz, Tom Nuno Wolf, Jaume Gual Ramon, and Christian Wachinger. X-SiT: Inherently interpretable surface vision transformers for dementia diagnosis. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pp. 680–690, 2025.

Tong Chen, Minheng Chen, Jing Zhang, Yan Zhuang, Chao Cao, Xiaowei Yu, Yanjun Lyu, Lu Zhang, Li Su, Tianming Liu, and Dajiang Zhu. A unified continuous staging framework for Alzheimer's disease and Lewy body dementia via hierarchical anatomical features. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pp. 13–23, 2025.

Wenju Cui, Yilin Leng, Yunsong Peng, Chen Bai, Lei Li, Xi Jiang, Gang Yuan, and Jian Zheng. A novel dynamic neural network for heterogeneity-aware structural brain network exploration and Alzheimer's disease diagnosis. *IEEE Transactions on Neural Networks and Learning Systems*, 36(9):15764–15778, 2025.

Angelo Del Sole, Francesca Clerici, Arturo Chiti, Michela Lecchi, Claudio Mariani, Laura Maggiore, Lisa Mosconi, and Giovanni Lucignani. Individual cerebral metabolic deficits in Alzheimer's disease and amnestic mild cognitive impairment: An FDG PET study. *European Journal of Nuclear Medicine and Molecular Imaging*, 35(7):1357–1366, 2008.

Rahul S. Desikan, Florent Ségonne, Bruce Fischl, Brian T. Quinn, Bradford C. Dickerson, Deborah Blacker, Randy L. Buckner, Anders M. Dale, R. Paul Maguire, Bradley T. Hyman, Marilyn S. Albert, and Ronald J. Killiany. An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *NeuroImage*, 31(3):968–980, 2006.

Yueqin Diao, Huihui Fang, Hanyi Yu, Yuning Wang, Yaling Tao, Ziyan Huang, Si Yong Yeo, and Yanwu Xu. Prior-guided prototype aggregation learning for Alzheimer's disease diagnosis. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pp. 483–492, 2025.

Lingzhong Fan, Hai Li, Junjie Zhuo, Yu Zhang, Jiaojian Wang, Liangfu Chen, Zhengyi Yang, Congying Chu, Sangma Xie, Angela R. Laird, Peter T. Fox, Simon B. Eickhoff, Chunshui Yu, and Tianzi Jiang. The human Brainnetome atlas: A new brain atlas based on connectional architecture. *Cerebral Cortex*, 26(8):3508–3526, 2016.

Xingyu Gao, Runyi Guo, and Zhe Zhao. Progression-aware generative model enhancing baseline visit prediction of early Alzheimer's disease. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pp. 493–501, 2025.

Stine Hansen, Samuel Kuttner, Michael Kampffmeyer, Tom-Vegard Markussen, Rune Sundset, Silje Kjærnes Øen, Live Eikenes, and Robert Jenssen. Unsupervised supervoxel-based lung tumor segmentation across patient scans in hybrid PET/MRI. *Expert Systems with Applications*, 167: 114244, 2021.

Mattias P. Heinrich, Ivor J. A. Simpson, Bartłomiej W. Papież, Michael Brady, and Julia A. Schnabel. Deformable image registration by combining uncertainty estimates from supervoxel belief propagation. *Medical Image Analysis*, 27:57–71, 2016.

Wenzheng Hu, Zhenghua Guan, Peng Yang, Jiaqiang Li, Yi Liu, Shushen Gan, Tuo Cai, Ao Zhang, Tengda Zhang, Junlong Qu, Shaolong Wang, Gege Cai, Xiang Dong, Tianfu Wang, and Baiying Lei. Anatomy-guided multimodal graph networks for Alzheimer's disease: Integrative analysis of cross-modal brain connectivity signatures. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pp. 66–75, 2025.

Hesheng Huang, Witold Pedrycz, Kaoru Hirota, and Fei Yan. A multiview-slice feature fusion network for early diagnosis of Alzheimer's disease with structural MRI images. *Information Fusion*, 119:103010, 2025a.

Huaibo Huang, Xiaoqiang Zhou, Jie Cao, Ran He, and Tieniu Tan. Vision transformer with super token sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22690–22699, 2023.

Shuo Huang, Lujia Zhong, and Yonggang Shi. Multistage alignment and fusion for multimodal multiclass Alzheimer's disease diagnosis. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pp. 375–385, 2025b.

Benjamin Irving, Amalia Cifor, Bartłomiej W. Papież, Jamie Franklin, Ewan M. Anderson, Michael Brady, and Julia A. Schnabel. Automated colorectal tumour segmentation in DCE-MRI using supervoxel neighbourhood contrast characteristics. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pp. 609–616, 2014.

Varun Jampani, Deqing Sun, Ming-Yu Liu, Ming-Hsuan Yang, and Jan Kautz. Superpixel sampling networks. In *European Conference on Computer Vision (ECCV)*, 2018.

Jinseong Jang and Dosik Hwang. M3T: Three-dimensional medical image classifier using multi-plane and multi-slice transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20686–20697, 2022.

Yifan Jia, Luoyu Wang, Yitian Tao, Zihao Zhu, Yawen Zhang, and Han Zhang. A multi-view, hybrid, hypergraph learning framework with causal perturbation for early Alzheimer's disease diagnosis. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 2026.

Hongchao Jiang and Chunyan Miao. Anatomy-aware gating network for explainable Alzheimer's disease diagnosis. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pp. 90–100, 2024.

Yufeng Jiang, Hexiao Ding, Hongzhao Chen, Jing Lan, Xinzhi Teng, Gerald W. Y. Cheng, Yunlin Mao, Zongxi Li, Haoran Xie, Jung Sun Yoo, and Jing Cai. Decoding the Alzheimer's continuum: Interpretable multi-gate routing for diagnosis and transition prediction. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pp. 172–182, 2026.

Daesung Kim, Seungbeom Seo, Boosung Kim, Kyobin Choo, Youngjun Jun, and Mijin Yun. PRGNN: Pyramidal region graph neural network for region-based brain PET classification. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pp. 554–563, 2025.

Kassymzhomart Kunanbayev, Vyacheslav Shen, and Dae-Shik Kim. Training ViT with limited data for Alzheimer's disease classification: An empirical study. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pp. 334–343, 2024.

Robert Laforce, Duygu Tosun, Pia Ghosh, Manja Lehmann, Cindee M. Madison, Michael W. Weiner, Bruce L. Miller, William J. Jagust, and Gil D. Rabinovici. Parallel ICA of FDG-PET and PiB-PET in three conditions with underlying Alzheimer's pathology. *NeuroImage: Clinical*, 4:508–516, 2014.

Yilin Leng, Wenju Cui, Chen Bai, Zirui Chen, Yanyan Zheng, and Jian Zheng. Dynamic structural brain network construction by hierarchical prototype embedding GCN using T1-MRI. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pp. 120–130, 2023.

Fan Li, Shilun Zhao, Shuwei Bai, Yuxiao Liu, Kai Zhang, Yin Xu, Ya Zhang, Kaicong Sun, and Dinggang Shen. Brain-heart-gut guided multi-constraint knowledge distillation for early Alzheimer's disease diagnosis. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pp. 45–54, 2025.

Qingfeng Li, Xiaodan Xing, Ying Sun, Bin Xiao, Hao Wei, Quan Huo, Minqing Zhang, Xiang Sean Zhou, Yiqiang Zhan, and Zhong Xue. Novel iterative attention focusing strategy for joint pathology localization and prediction of MCI progression. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pp. 307–315, 2019.

Chaoqiang Liu, Fei Huang, Anqi Qiu, and Alzheimer's Disease Neuroimaging Initiative. Monte Carlo ensemble neural network for the diagnosis of Alzheimer's disease. *Neural Networks*, 159: 14–24, 2023.

Fei Liu, Huabin Wang, Mohamed Hisham Jaward, Shiuan-Ni Liang, Huey Fang Ong, and Jiayuan Cheng. Anatomical graph-based multilevel distillation for robust Alzheimer's disease diagnosis with missing modalities. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pp. 74–83, 2025.

Jingyu Liu, Weigang Cui, Yipeng Chen, Yulan Ma, Qunxi Dong, Ran Cai, Yang Li, and Bin Hu. Deep fusion of multi-template using spatio-temporal weighted multi-hypergraph convolutional networks for brain disease analysis. *IEEE Transactions on Medical Imaging*, 43(2):860–873, 2024.

Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 11525–11538, 2020.

Donghuan Lu, Karteek Popuri, Gavin Weiguang Ding, Rakesh Balachandar, and Mirza Faisal Beg. Multiscale deep neural network based analysis of FDG-PET images for the early diagnosis of Alzheimer's disease. *Medical Image Analysis*, 46:26–34, 2018.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, pp. 4765–4774, 2017.

Francesco Mercaldo, Marcello Di Giammarco, Fabrizio Ravelli, Fabio Martinelli, Antonella Santone, and Mario Cesarelli. Alzheimer's disease evaluation through visual explainability by means of convolutional neural networks. *International Journal of Neural Systems*, 34(2):2450007, 2024.

Kwanseok Oh, Jee Seok Yoon, and Heung-Il Suk. Learn-explain-reinforce: Counterfactual reasoning and its guidance to reinforce an Alzheimer's disease diagnosis model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4843–4857, 2023.

Zaixin Ou, Caiwen Jiang, Yuxiao Liu, Yuanwang Zhang, Zhiming Cui, and Dinggang Shen. A graph-embedded latent space learning and clustering framework for incomplete multimodal multiclass Alzheimer's disease diagnosis. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pp. 45–55, 2024.

Yuru Pei, Yunai Yi, Gengyu Ma, Tae-Kyun Kim, Yuke Guo, Tianmin Xu, and Hongbin Zha. Spatially consistent supervoxel correspondences of cone-beam computed tomography images. *IEEE Transactions on Medical Imaging*, 37(10):2310–2321, 2018.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. FiLM: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Minh Tuan Pham, Quoc Anh Le, Thanh Trung Le, Le Trong Phan, Mouloud Adel, Salah Bourennane, Eric Guedj, Ismail Burak Parlak, Guy Nagels, and Nguyen Linh Trung. Novel PET-driven brain mapping for diagnosis of Alzheimer's disease: Integrated feature extraction and ranking capabilities. *APSIPA Transactions on Signal and Information Processing*, 15(1):488–531, 2026.

Yan-Rui Qiang, Qin-Yi Zhou, Jia-Ni Li, Ming-Yu Xie, Xiaodong Cui, and Shao-Wu Zhang. Classification of Alzheimer's disease by jointing 3D depthwise separable convolutional neural network and transformer. *Expert Systems with Applications*, 286:127720, 2025.

Antoine Rogeau, Florent Hives, Cécile Bordier, Hélène Lahousse, Vincent Roca, Thibaud Lebouvier, Florence Pasquier, Damien Huglo, Franck Semah, and Renaud Lopes. A 3D convolutional neural network to classify subjects as Alzheimer's disease, frontotemporal dementia or healthy controls using brain 18F-FDG PET. *NeuroImage*, 288:120530, 2024.

Hyun Gee Ryoo, Hongyoon Choi, Kuangyu Shi, Axel Rominger, Dong Young Lee, Dong Soo Lee, and Alzheimer's Disease Neuroimaging Initiative. Distinct subtypes of spatial brain metabolism patterns in Alzheimer's disease identified by deep learning-based FDG PET clusters. *European Journal of Nuclear Medicine and Molecular Imaging*, 51(2):443–454, 2024.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 618–626, 2017.

Jaeyoon Sim, Minjae Lee, Guorong Wu, and Won Hwa Kim. Multi-modal graph neural network with transformer-guided adaptive diffusion for preclinical Alzheimer classification. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pp. 511–521, 2024.

Thomas Yu Chow Tam, Litian Liang, Ke Chen, Haohan Wang, and Wei Wu. A quantitative approach for evaluating disease focus and interpretability of deep learning models for Alzheimer's disease classification. In *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 5104–5111, 2024.

N. Tzourio-Mazoyer, B. Landeau, D. Papathanassiou, F. Crivello, O. Etard, N. Delcroix, B. Mazoyer, and M. Joliot. Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *NeuroImage*, 15(1):273–289, 2002.

Olga Veksler, Yuri Boykov, and Paria Mehrani. Superpixels and supervoxels in an energy optimization framework. In *European Conference on Computer Vision (ECCV)*, pp. 211–224, 2010.

Yifeng Wang, Ke Chen, and Haohan Wang. ADAPT: Adaptive profiling transformers for efficient Alzheimer's disease diagnosis. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 2026.

Liang Wu, Shunbo Hu, Duanwei Wang, Changchun Liu, and Li Wang. RClaNet: An explainable Alzheimer's disease diagnosis framework by joint registration and classification. *IEEE Journal of Biomedical and Health Informatics*, 28(4):2338–2349, 2024.

Tingsong Xiao, Lu Zeng, Xiaoshuang Shi, Xiaofeng Zhu, and Guorong Wu. Dual-graph learning convolutional networks for interpretable Alzheimer's disease diagnosis. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pp. 406–415, 2022.

Jinghao Xu, Chenxi Yuan, Xiaochuan Ma, Huifang Shang, Xiaoshuang Shi, and Xiaofeng Zhu. Interpretable medical deep framework by logits-constraint attention guiding graph-based multiscale fusion for Alzheimer's disease analysis. *Pattern Recognition*, 152:110450, 2024.

Jiayu Ye, An Zeng, Dan Pan, Yiqun Zhang, Jingliang Zhao, Qiuping Chen, and Yang Liu. MADFormer: A traceable interpretability model for Alzheimer's disease recognition based on multipatch attention. *IEEE Journal of Biomedical and Health Informatics*, 28(6):3637–3648, 2024.

Shengjie Zhang, Xiang Chen, Bohan Ren, Haibo Yang, Ziqi Yu, Xiao-Yong Zhang, and Yuan Zhou. 3D global Fourier network for Alzheimer's disease diagnosis using structural MRI. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pp. 34–43, 2022.

Shilun Zhao, Fan Li, Zhichao Liang, Kaicong Sun, Shuwei Bai, Weilin Zhou, Xin Lin, Zihao Wang, and Dinggang Shen. A hierarchical multi-task framework for dementia diagnosis via pathological feature learning from multi-organ data. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 2026.

Wenyong Zhu, Liang Sun, Jiashuang Huang, Liangxiu Han, and Daoqiang Zhang. Dual attention multi-instance deep learning for Alzheimer's disease diagnosis with structural MRI. *IEEE Transactions on Medical Imaging*, 40(9):2354–2366, 2021.

# A Dataset Statistics and Image Preprocessing

**Preprocessing.** Each PET image is normalized by clipping positive voxel intensities to the 1st to 99.5th percentile range and dividing by the 99.5th percentile. The normalized images are affinely registered to the SPM12 template space and resampled to a $121 \times 145 \times 121$ grid with an isotropic voxel size of 1.5 mm. Finally, an intracranial mask derived from the SPM12 tissue probability maps removes non-brain regions.

**Statistics.** Table 3 lists the number of subjects in each class for every tracer and task. The ADNI tau and amyloid tasks are exactly class-balanced, whereas the FDG and AIBL tasks are moderately imbalanced. For AIBL, the three amyloid tracers are pooled into a single cohort.

Table 3: Statistics of the PET datasets used in our experiments.

| Dataset | PET Type | Tracer | Task | Class Distribution | Total |
|---|---|---|---|---|---|
| ADNI | Tau | $^{18}$F-AV1451 | AD vs. CN | 85 / 85 | 170 |
| | | | MCI vs. CN | 300 / 300 | 600 |
| | Amyloid | $^{18}$F-AV45 | AD vs. CN | 180 / 180 | 360 |
| | | | MCI vs. CN | 195 / 195 | 390 |
| | Metabolic | $^{18}$F-FDG | AD vs. CN | 230 / 235 | 465 |
| | | | MCI vs. CN | 320 / 235 | 555 |
| AIBL | Amyloid | $^{11}$C-PiB<br>$^{18}$F-AV45<br>$^{18}$F-Flutemetamol | AD vs. CN | 122 / 149 | 271 |
| | | | MCI vs. CN | 147 / 149 | 296 |

# B Notation

Table 4: Summary of notation.

| Symbol | Description |
|---|---|
| $\Omega$, $i$ | Set of intracranial voxels and voxel index |
| $M$, $j$ | Number of supervoxels and supervoxel index |
| $T$, $t$ | Number of ASP iterations and iteration index |
| $\mathbf{p}_i$, $\mathbf{v}_i^{(t)}$ | Normalized coordinate and feature of voxel $i$ |
| $\mathbf{c}_j^{(t)}$, $\mathbf{s}_j^{(t)}$ | Center and feature of supervoxel $j$ |
| $\bar{\mathbf{s}}_j^{(t)}$ | Supervoxel feature after multi-head self-attention |
| $\mathcal{N}_i^{(t)}$ | Candidate supervoxels of voxel $i$ |
| $a_{ij}^{(t)}$, $\mathbf{A}^{(t)}$ | Association weight and association matrix |
| $\lambda$, $d_k$ | Spatial prior strength and projection dimension |
| $\Omega_j$ | Voxels assigned to supervoxel $j$ after $T$ iterations |
| $\mathcal{R} = \{\mu_i^{\mathrm{CN}}, \sigma_i^{\mathrm{CN}}\}$ | Voxel-wise CN reference (mean and standard deviation) |
| $d_i^{\mathrm{v}}$, $d_j^{\mathrm{sv}}$ | Voxel and supervoxel deviations from the CN reference |
| $p_j$, $g_j$ | Abnormality score and hard gate |
| $\tau$, $\alpha$ | Abnormality threshold and gate sharpness |
| $\mathbf{u}$ | Whole-brain context vector |
| $\mathbf{e}_j^A$, $\mathbf{e}_j^P$ | Evidence of the abnormal and preserved experts |
| $\mathcal{M}$ | Set of valid (non-empty) supervoxels |
| $\mathbf{C}_j$, $\mathbf{z}$ | Class-wise contribution of supervoxel $j$ and class logits |
| $r_j$ | Regional contribution score, $C_{j,\mathrm{AD}} - C_{j,\mathrm{CN}}$ |

# C MORE RESULTS

We provide additional examples of supervoxel evolution across training epochs, individual explanations, and longitudinal explanations. The signed regional contribution is defined as $r_j = C_{j,\mathrm{AD}} - C_{j,\mathrm{CN}}$; positive and negative values indicate evidence toward AD and CN, respectively.

## C.1 ADDITIONAL SUPERVOXEL EVOLUTION ACROSS EPOCHS

Figures 7–9 provide additional epoch-wise visualizations for FDG, AV1451, and AV45 PET.




Figure 7: Additional supervoxel evolution visualizations on $^{18}$F-FDG PET. The five panels are arranged from top to bottom.

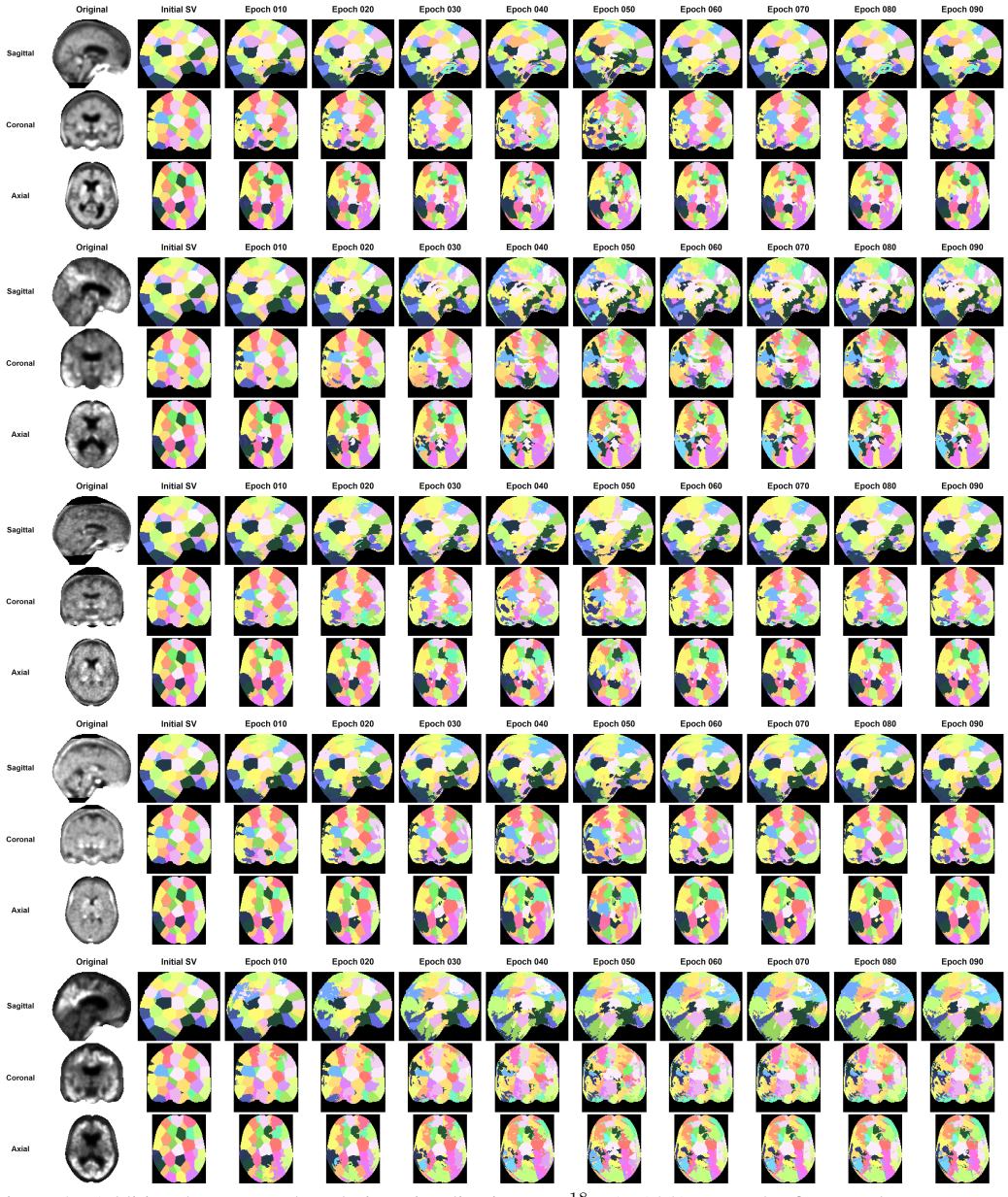



Figure 8: Additional supervoxel evolution visualizations on $^{18}$F-AV1451 PET. The five panels are arranged from top to bottom.




Figure 9: Additional supervoxel evolution visualizations on $^{18}$F-AV45 PET. The five panels are arranged from top to bottom.

### C.2 Additional Individual Explanations

Figures 10–15 provide six representative AD and CN examples. For each subject, the upper visualization shows the abnormal supervoxels selected by the classifier together with the gate and contribution summaries. The lower visualization projects the signed supervoxel contributions onto three anatomical slices. Together, the two views connect region selection to the evidence used for diagnosis.

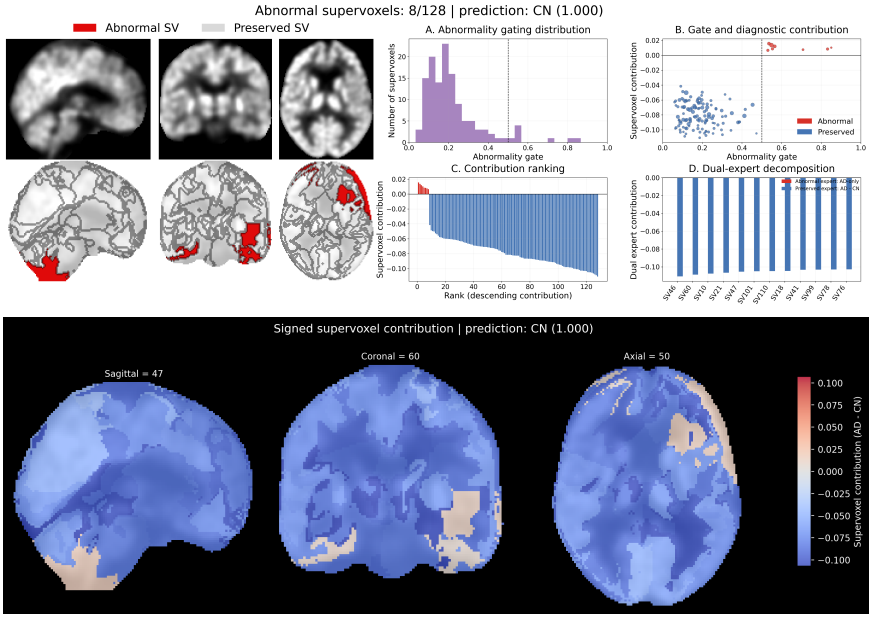



Figure 10: Individual explanation for subject A (CN). Upper: selected abnormal supervoxels and gate/contribution summaries. Lower: signed supervoxel contributions shown in three anatomical slices.

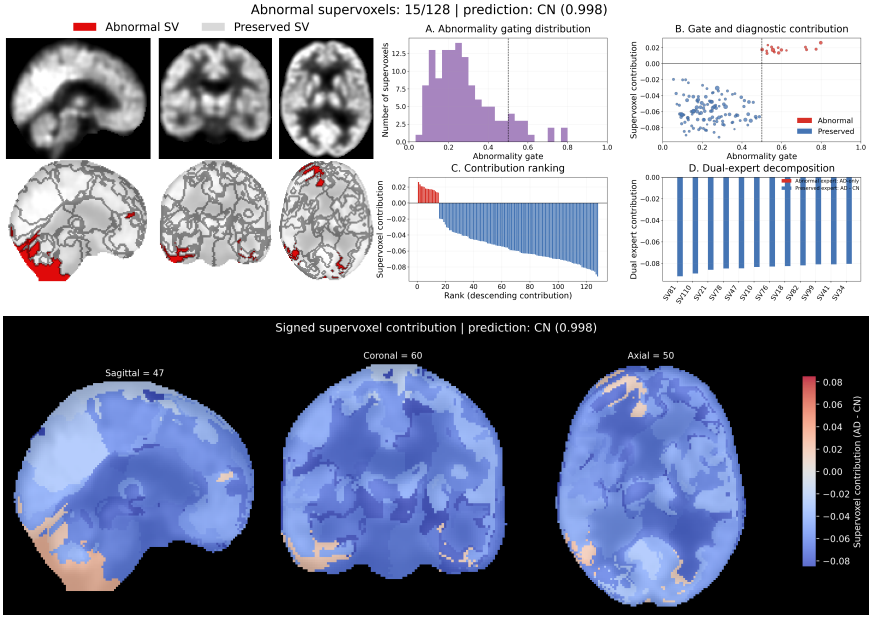



Figure 11: Individual explanation for subject B (CN). Upper: selected abnormal supervoxels and gate/contribution summaries. Lower: signed supervoxel contributions shown in three anatomical slices.

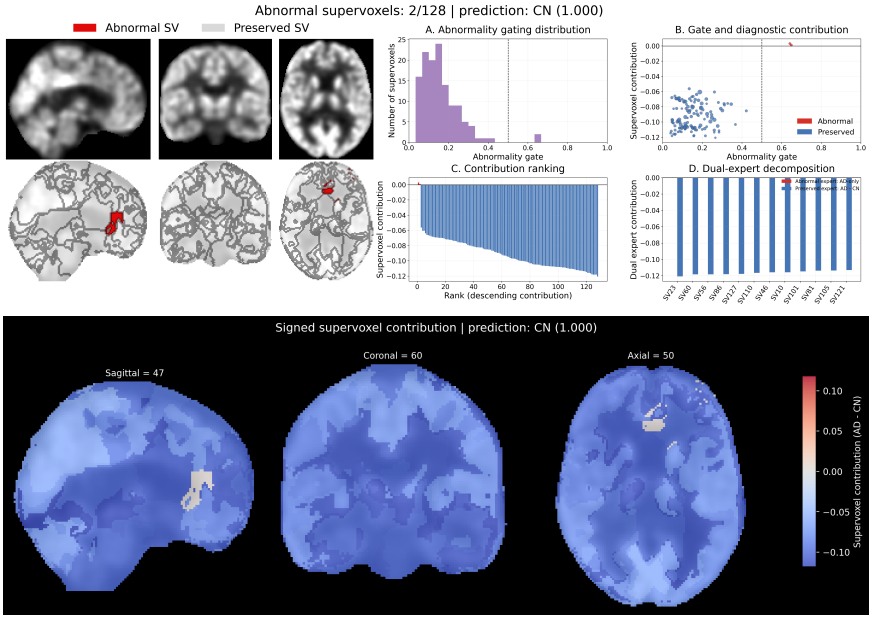



Figure 12: Individual explanation for subject C (CN). Upper: selected abnormal supervoxels and gate/contribution summaries. Lower: signed supervoxel contributions shown in three anatomical slices.

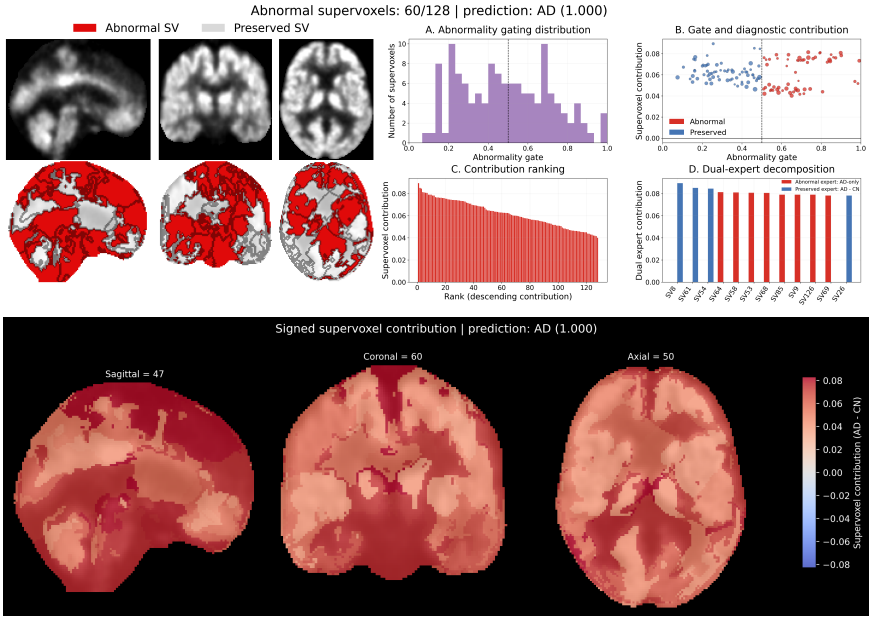



Figure 13: Individual explanation for subject D (AD). Upper: selected abnormal supervoxels and gate/contribution summaries. Lower: signed supervoxel contributions shown in three anatomical slices.

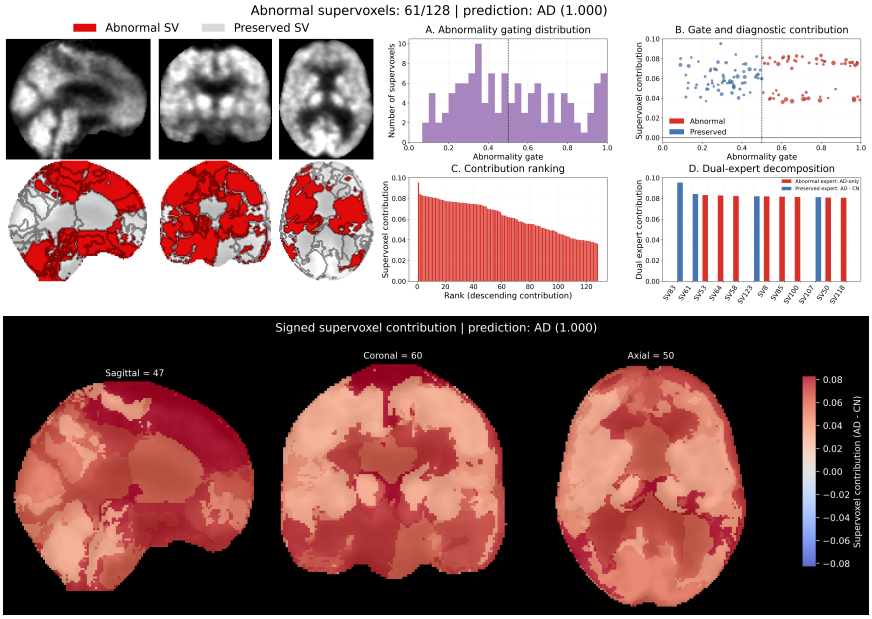



Figure 14: Individual explanation for subject E (AD). Upper: selected abnormal supervoxels and gate/contribution summaries. Lower: signed supervoxel contributions shown in three anatomical slices.

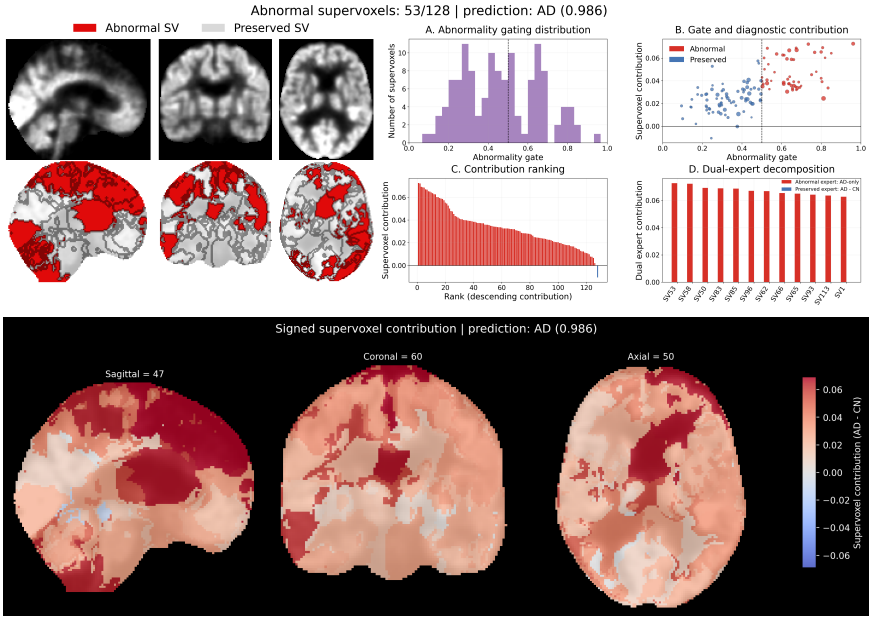



Figure 15: Individual explanation for subject F (AD). Upper: selected abnormal supervoxels and gate/contribution summaries. Lower: signed supervoxel contributions shown in three anatomical slices.

### C.3 LONGITUDINAL EXPLANATIONS IN AIBL

Figures 16–19 follow two AIBL subjects from an earlier CN visit to a later AD visit. For each subject, we show the selected abnormal supervoxels and their signed regional contributions at both visits. Notably, at the earlier CN visit, subject G already exhibits a substantial abnormal-gated fraction (45/128), which increases to 74/128 at the later AD visit. For subject H, the abnormal-gated regions appear in similar locations across visits, while the signed regional contributions shift from predominantly CN-supporting to AD-supporting. These examples illustrate how regional diagnostic evidence evolves within individual subjects over time.

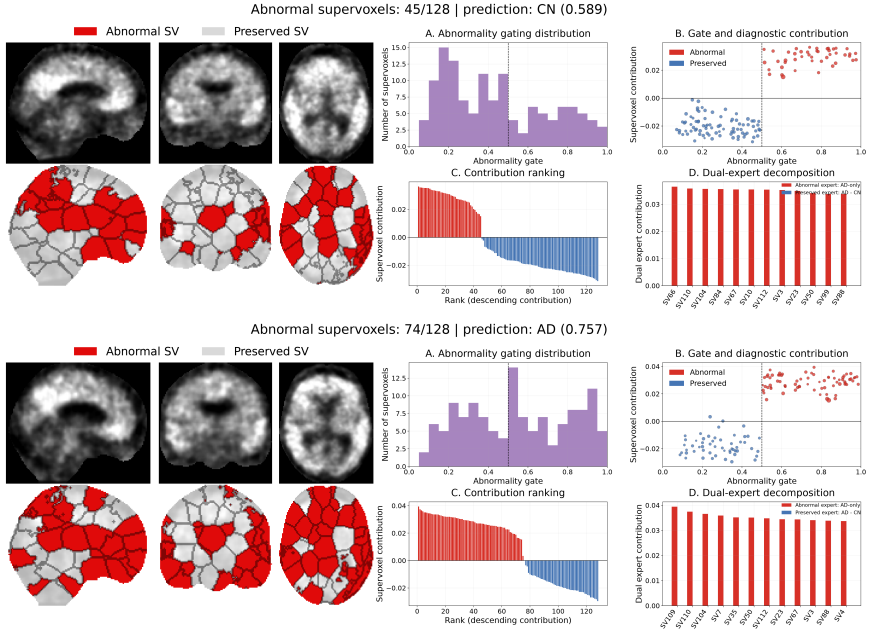



Figure 16: Longitudinal change in selected abnormal supervoxels for AIBL subject G. Upper: CN visit at month 18. Lower: AD visit at month 54. Both panels include the regional gate and contribution summaries.

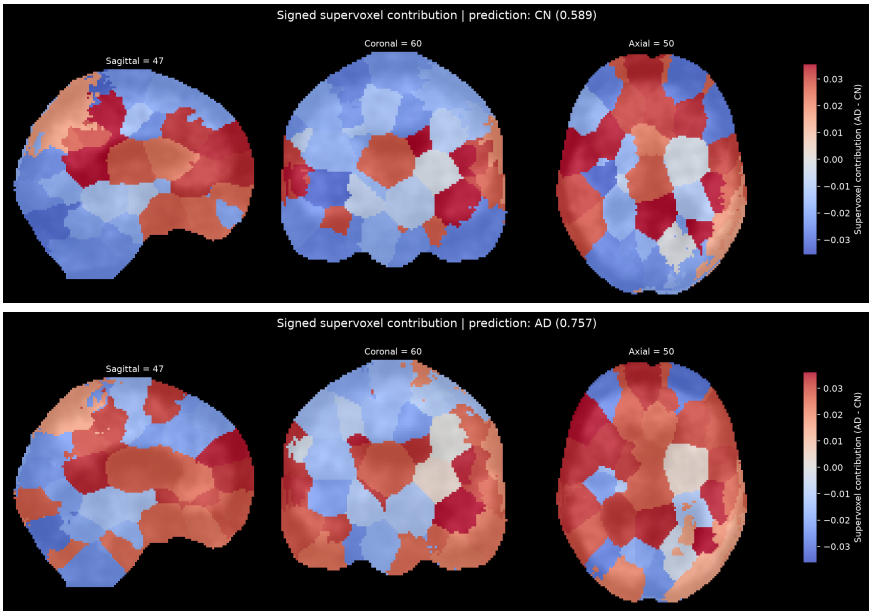



Figure 17: Longitudinal change in signed supervoxel contributions for AIBL subject G. The CN visit at month 18 is above the AD visit at month 54; each visit is shown in three anatomical slices.

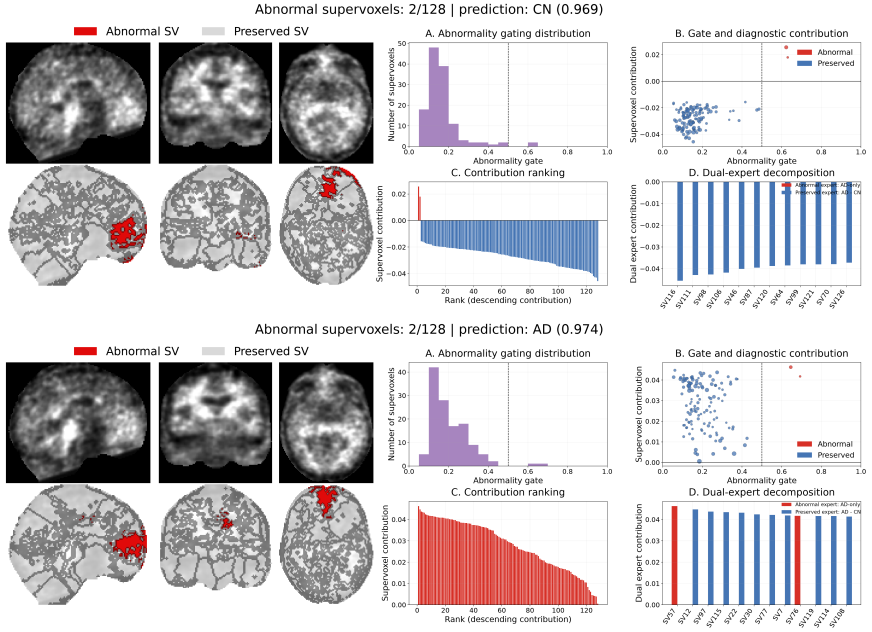



Figure 18: Longitudinal change in selected abnormal supervoxels for AIBL subject H. Upper: CN visit at month 18. Lower: AD visit at month 36. Both panels include the regional gate and contribution summaries.

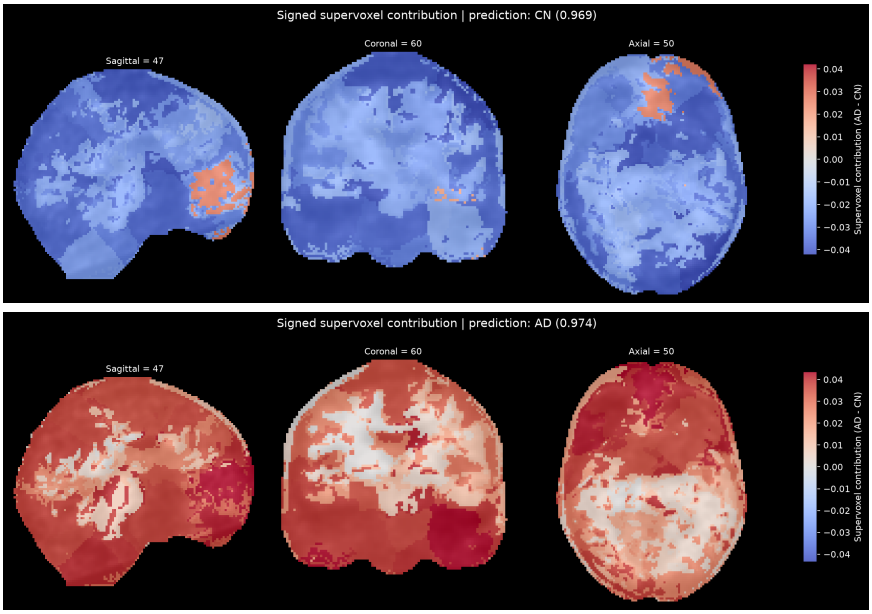



Figure 19: Longitudinal change in signed supervoxel contributions for AIBL subject H. The CN visit at month 18 is above the AD visit at month 36; each visit is shown in three anatomical slices.