Hierarchical Differential MedViT-3D: Specialized Hybrid Transformer for Multiclass Diagnosis of Neurodegenerative Diseases

Éloi Navet[1 [0009−0009−0705−3678]], Rémi Giraud[2 [0000−0003−3965−1719]], Boris Mansencal[1 [0000−0002−9190−4819]], and Pierrick Coupé[1 [0000−0003−2709−3350]]

[1] Univ. Bordeaux, CNRS, Bordeaux INP, LaBRI, UMR 5800, F-33400 Talence, France
eloi.navet@u-bordeaux.fr
[2] Univ. Bordeaux, CNRS, Bordeaux INP, IMS, UMR 5218, F-33400 Talence, France

Abstract. Differential diagnosis of neurodegenerative diseases using 3D T1w MRI remains a challenge, particularly when distinguishing between clinically similar dementia subtypes. While deep learning has achieved high performance on binary screening (e.g., controls vs. Alzheimer's Disease), standard architectures often struggle with the class imbalance and subtle anatomical variations, leading to a critical trade-off between global specificity and focal sensitivity. In this work, we propose a hybrid Dual-Stream framework using a Hierarchical Differential MedViT-3D. First, a fine-scale stream adapts the MedViTV2 architecture to 3D and adds recent Differential Attention and Rotary Positional Embeddings (RoPE) mechanisms to capture focal anomalies that standard attention misses. Second, to counter the "vanishing specificity" of such high-sensitivity models, we fuse this focal stream with global volumetric priors and introduce a hierarchical loss strategy grouping pathologies. We evaluate our method on a challenging 7-class diagnostic task using over 4,000 scans aggregated from 16 datasets. Our experiments suggest that this dual-stream hierarchical specialization improves the empirical trade-off between rare-variant sensitivity and CN specificity, achieving the highest Top-2 balanced accuracy among the evaluated methods on the held-out OOD set. By explicitly modeling the balance between global context and local saliency, our approach provides a clinically motivated framework for computer-aided diagnosis in complex settings. Code is available here: https://github.com/EloiNavet/Hierarchical-Differential-MedViT-3D.

Keywords: Differential Diagnosis · Vision Transformers · Neurodegeneration · Late-Fusion · Differential Attention.

1 Introduction

Accurate differential diagnosis of neurodegenerative disorders from 3D structural MRI is clinically important because several diseases present with overlapping symptoms despite differences in their underlying pathology, prognosis, and

clinical management. While deep learning methods have achieved strong performance on binary screening tasks, such as distinguishing Alzheimer's Disease (AD) from Cognitively Normal (CN) individuals, extending these methods to multi-class differential diagnosis remains challenging. In this work, we consider a clinically relevant 7-class setting comprising AD, CN, Dementia with Lewy Bodies (DLB), three frontotemporal dementia (FTD) subtypes—behavioral variant FTD (bvFTD), Progressive Non-Fluent Aphasia (PNFA), and Semantic Dementia (SD)—and Progressive Supranuclear Palsy (PSP).

This task is further complicated by severe class imbalance, as several disorders are substantially rarer than AD and CN. Moreover, neurodegenerative patterns span multiple spatial scales, from relatively diffuse structural alterations to highly localized cortical degeneration. Models must therefore preserve coarse anatomical context while remaining sensitive to subtle regional abnormalities.

Standard spatial networks face a representational bottleneck. Convolutional neural network (CNN) pooling operations and the low-pass filtering nature of standard Vision Transformer (ViT) self-attention intrinsically favor global morphometry, often obscuring the fine-scale anomalies typical of focal lobar degenerations [27,37]. Furthermore, standard ViTs often struggle with geometric robustness and reproducibility in volumetric neuroimaging [23]. Recent mechanisms such as Differential Attention [39] address this limitation by computing attention as the difference between two softmax maps. This acts as a spatial high-pass filter, amplifying localized textural deviations while suppressing dominant background signals.

However, applying such aggressive focal extractors to 3D brain MRI reveals an empirical vulnerability that we term the *saliency trap*. In our experiments, increasing focal sensitivity is accompanied by a marked loss of CN specificity, suggesting that highly focal extractors may over-detect pathological patterns within normal anatomical variability. Restoring this balance requires explicit structural priors. Rather than relying on a low-resolution convolutional branch, which may still produce entangled representations, we leverage explicit regional brain volumes to provide complementary geometric information.

To address this specificity–sensitivity trade-off, we propose a Dual-Stream Late-Fusion framework that explicitly separates global anatomical priors from focal saliency extraction. Embracing representational complementarity, our main contributions are:

- **Fine-Scale Focal Extractor:** We introduce a 3D-adapted MedViTV2 [18] equipped with 3D Rotary Positional Embeddings (RoPE) and Differential Attention. We analyze its capacity to isolate rare focal dementias while characterizing the observed *saliency trap*-like behavior that limits its standalone diagnostic reliability.
- **Clinical Hierarchical Regularization:** To stabilize the focal stream against unstructured noise, we introduce a Hierarchical Risk-Aware Loss. This regularizes the latent space by grouping fine-grained diagnoses into broad neuroanatomical meta-classes and penalizing severe inter-group diagnostic errors.

– **Dual-Stream Late-Fusion Framework:** We propose a multi-scale fusion strategy combining explicit global priors, derived from regional volumes processed by a support vector machine (SVM), with fine-scale transformer class probabilities. Evaluated on an out-of-domain (OOD) test set of 1,304 subjects, our framework achieves a balanced accuracy of 62%, offering an adaptive compromise between baseline specificity and rare-variant sensitivity [1,30].

2 Methodology

Robust computer-aided diagnosis for dementia must reconcile conflicting scales of biological information: coarse-scale morphological deformations (e.g., global ventricular expansion) and fine-scale textural anomalies (e.g., localized cortical thinning). Standard end-to-end architectures often struggle to simultaneously optimize these diverging spatial frequencies. To overcome this representational bottleneck, we propose a Dual-Stream Late-Fusion framework (Fig. 1) that decouples feature extraction: a Global Stream establishes rigid anatomical priors, and a Fine-Scale Stream operates as a highly sensitive focal anomaly extractor.

2.1 Global Stream: Coarse-Scale Anatomical Priors

To provide robust geometric baselines and preserve high diagnostic specificity for CN subjects, the global stream operates exclusively on coarse-scale structural volumes. Following extraction protocols from large-scale studies [7], 133 regional brain volumes are extracted from 3D T1w MRIs [6]. These low-frequency volumetric features are fed into an SVM, whose class-posterior probabilities provide stable global priors $P_{global} \in [0,1]^C$ ($C = 7$). Regional volumes were normalized by total brain-mask volume and standardized using statistics estimated from the training folds; SVM hyperparameters were selected per fold with Optuna/TPE over linear, RBF, and polynomial kernels, with $C \in [10^{-4}, 1]$, $\gamma \in [10^{-5}, 10^{-1}]$ for RBF/polynomial kernels, and polynomial degree 2–5.

2.2 Fine-Scale Stream: Focal Saliency Extractor

The fine-scale stream leverages a 3D adaptation of MedViTV2 [18], a hybrid CNN-Transformer network utilizing Kolmogorov-Arnold Networks (KAN) and Dilated Neighborhood Attention. We adapt this foundation to specifically isolate fine-scale textural variations within raw 3D MRI patches.

3D Adaptation and RoPE. We extend MedViT's blocks to 3D. Since absolute positional embeddings lack translation invariance and struggle with volumetric geometry, we integrate 3D RoPE [31]. To accurately encode 3D spatial geometry, the embedding dimension d is divided into three equal segments for independent rotary transformations along the 3 axes. This natively encodes relative spatial distances, improving attention allocation across irregular morphological structures [36].

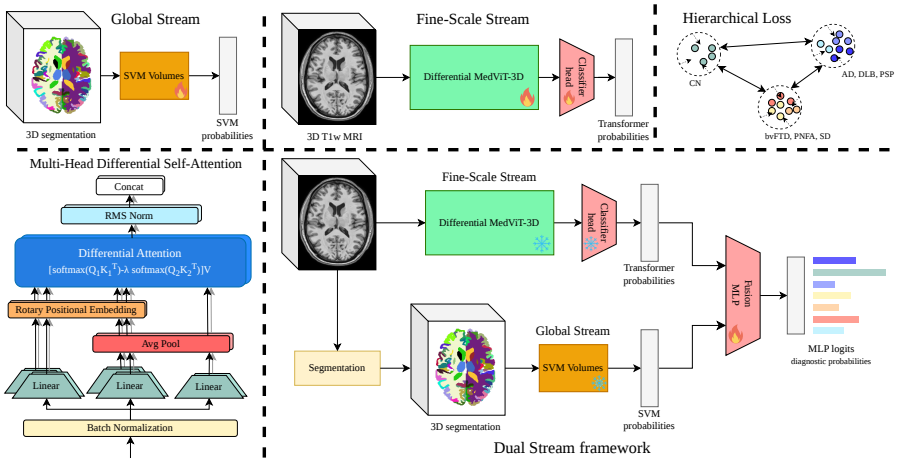

Fig. 1. Detailed overview of the proposed Hierarchical MedViT-3D Dual-Stream framework. **Top-Left:** The Global Stream trains an SVM on regional brain volumes to establish robust coarse-scale priors. **Top-Middle:** The Fine-Scale Stream depicts the independent end-to-end training of the MedViT-3D to extract focal textural anomalies from 3D MRI patches. **Top-Right:** The Hierarchical Loss used when training the Fine-Scale Stream. **Bottom-Left:** The Multi-Head Differential Self-Attention module. Adapted from MedViTV2 E-MHSA block, it is modified to incorporate 3D Rotary Positional Embeddings for spatial awareness, and Differential Attention acting as a high-pass filter. **Bottom-Right:** The Late-Fusion architecture. Flame icons denote independent base-model training phases. Snowflake icons denote the fusion phase where the SVM, MedViT-3D, and the off-the-shelf segmentation pipeline are fully frozen. An adaptive Fusion-MLP non-linearly aggregates these features to yield diagnostic probabilities. Abbreviations: MLP, multi-layer perceptron; RMSNorm, root-mean-square normalization.

Differential Attention Mechanism. To amplify subtle focal pathologies often overshadowed by healthy tissue, we integrate Differential Attention [39]. The attention score is the difference between two softmax maps:

$$\text{DiffAttn}(X) = \left(\text{Softmax}\left(\frac{Q_1 K_1^T}{\sqrt{d}}\right) - \lambda \text{Softmax}\left(\frac{Q_2 K_2^T}{\sqrt{d}}\right)\right) V, \tag{1}$$

where X is the input token sequence, Q_i and K_i are distinct projection matrices, V is the value matrix, and d is the head dimension. The learnable scalar λ is initialized following [39]. Clinically, this subtraction acts as a spatial high-pass filter: the first term captures raw structural details (signal + noise), while the second models the pervasive low-frequency background context (noise), effectively isolating anomalous textural deviations.

Hierarchical Risk-Aware Loss. Empirically, Differential Attention's high-pass nature can cause latent instability, over-predicting focal anomalies in healthy subjects. Inspired by biologically-informed regularizations [22], we constrain the

extractor using a Hierarchical Risk-Aware Loss ($\mathcal{L}_{HL}$). We map the 7 fine-grained diagnostic classes (y_{fine}) into 3 structural meta-classes (y_{coarse}) based on broad MRI phenotypes: (1) *Intact:* CN; (2) *Diffuse & Subcortical:* AD, DLB, PSP; and (3) *Focal Cortical:* bvFTD, PNFA, SD. To penalize catastrophic errors between these distinct phenotypical groups, the total loss combines a fine-grained label-smoothed cross-entropy with a coarse-level Focal Loss:

$$\mathcal{L}_{HL} = \mathcal{L}_{CE}^{LS}(y_{fine}, \hat{y}_{fine}) + \lambda_{HL}\mathcal{L}_{Focal}(y_{coarse}, \hat{y}_{coarse}, \gamma), \tag{2}$$

where $\mathcal{L}_{CE}^{LS}$ denotes label-smoothed cross-entropy, $\mathcal{L}_{Focal}$ denotes the coarse-level focal loss, and the coarse probability predictions are the marginal sums of their respective fine-class softmax scores (e.g., $\hat{y}_{coarse,Intact} = \hat{y}_{fine,CN}$). Applied at the meta-class level, the focal term enforces a focus on resolving "hard" inter-group errors. By penalizing confident but anatomically inconsistent predictions, this hierarchical constraint aims to reduce severe inter-group errors; calibration is reported separately. Parameters ($\lambda_{HL} = 2.0, \gamma = 2.0$) were set empirically [29].

2.3 Adaptive Late-Fusion

To resolve the specificity-sensitivity trade-off, a Multi-Layer Perceptron (MLP) acts as a probability-level late-fusion module. Let $P_{focal} \in [0,1]^C$ denote the softmax class-probability vector produced by the fine-scale MedViT classifier and $P_{global} \in [0,1]^C$ the SVM class-posterior probability vector. The fused logits are $Z_{fusion} = \text{MLP}([P_{global}, P_{focal}])$. The Fusion-MLP comprises three fully connected layers (14→64→32→7), with ReLU activations and dropout after each hidden layer. Trained on ID data, the MLP adaptively learns to rely on the SVM for coarse geometric context (safeguarding baseline specificity) and the MedViT for subtle anomalies (maintaining high sensitivity for focal dementias).

3 Experimental Setup

Multi-Centric Dataset and Clinical Cohorts. To ensure robust clinical validation, we aggregated 3D T1w MRI scans from 16 independent multi-centric cohorts, yielding 4,107 subjects across 7 diagnostic classes (see Table 1), as proposed in [7]. Training and testing cohorts are strictly separated to assess OOD generalization. The **In-domain (ID)** set (2,803 subjects) aggregates 12 cohorts (MIRIAD [17], PDBP [25], ALLFTD [4], DLBS [26], ABIDE [8], ADNI [21], ICBM [20], IXI [16], OASIS [19], SALD [38], CamCAN [32], NDAR [28]). The **Out-of-domain (OOD)** set (1,304 subjects) is a fixed test set comprising 4 entirely unseen cohorts (NACC [3], AIBL [10], NIFD [12], 4RTNI [34]). Detailed cohort-level demographic and acquisition characteristics are provided in Supplementary Table S1.

Preprocessing and Implementation Details. All MRIs underwent unified preprocessing with N4 bias correction [33], skull-stripping, affine MNI152 registration [2,11], 1 mm isotropic resampling, and Z-score normalization. Regional volumes were extracted via UNet models [6] for the global stream. Concurrently,

Table 1. Distribution of structural MRI scans across diagnostic classes. Subjects are separated into ID and OOD. Values are presented as: Total count, (Male/Female), and [Age min - Age max].

Subset	Diagnosis							Total
	AD	**CN**	**DLB**	**bvFTD**	**PNFA**	**SD**	**PSP**	
ID	428 (217/211) [55.0-96.0]	1926 (827/1099) [44.0-94.0]	126 (104/22) [50.0-90.0]	154 (97/57) [49.0-83.0]	41 (19/22) [55.0-82.0]	39 (23/16) [50.0-79.0]	89 (42/47) [52.0-80.0]	2803 (1329/1474) [44.0-96.0]
OOD	488 (184/304) [46.2-96.1]	528 (220/308) [30.1-100.2]	47 (39/8) [54.5-88.6]	90 (59/31) [45.0-76.3]	40 (17/23) [54.0-81.0]	44 (24/20) [50.0-85.2]	67 (29/38) [55.0-86.0]	1304 (572/732) [30.1-100.2]

normalized 3D images were cropped to $144 \times 168 \times 144$ voxels for the fine-scale stream.

To address class imbalance, a weighted random sampler (inversely proportional to ID frequencies) was applied during training. On-the-fly 3D augmentations included affine/elastic deformations, spatial cropping, noise injection, and 3D MixUp [5]. These augmentations were applied only during fine-scale stream training and were not used for regional volume extraction. The fine-scale neural model was optimized via AdamW (LR 3×10^{-4}, weight decay 0.05, cosine annealing).

To prevent data leakage during patient-level stratified 10-fold cross-validation, we adopted a strict cross-fitting strategy *within* the ID set. For each fold, base models (SVM and MedViT-3D) were trained exclusively on a 70% ID training split. To avoid overfitting, the adaptive late-fusion MLP was then optimized purely on the out-of-sample class-probability vectors from a 20% ID validation split (the remaining 10% being the internal test). The external OOD dataset was strictly excluded from all training, validation, and hyperparameter tuning phases. Reported OOD results represent the ensemble average of the 10 cross-validation models.

Baselines and Evaluation Metrics. We compare our approach to the global stream SVM, the Lifespan Tree (an unsupervised UMAP using volumes) [7], 3D ResNet-50 [15], 3D Swin-DPL [24], and vanilla MedViT-3D. Due to severe imbalance, we report Balanced Accuracy (BACC) alongside Top-1 and Top-2 class-wise sensitivities. For a given class, Top-k sensitivity is defined as the proportion of subjects for which the ground-truth diagnosis is among the k highest-scoring predicted classes. Top-k BACC is the unweighted mean of the corresponding class-wise Top-k sensitivities. We report Top-2 BACC as a measure of diagnostic shortlisting performance rather than as a substitute for Top-1 classification. To assess whether confidence scores reliably reflect true accuracy without relying on post-hoc temperature scaling, we measured the Expected Calibration Error (ECE) [14].

4 Results

4.1 Architectural Ablation and the *saliency trap*

We evaluated successive architectural configurations (Table 2). Integrating RoPE, Differential Attention, and the Hierarchical Loss (HL) progressively shifted the network's predictive bias away from the majority healthy class and toward under-represented pathological variants.

While RoPE yielded morphological regularization, Differential Attention pivoted the network's focus toward focal variants (e.g., doubling PNFA Top-1 sensitivity to 62%). Constrained by the HL, the fine-scale stream achieved its highest sensitivity for rare focal dementias. However, focal specialization exposed a *saliency trap*-like behavior: improved focal sensitivity came with a severe drop in CN specificity.

The Late-Fusion MLP mitigated this sensitivity-specificity trade-off. By re-incorporating global SVM priors, the Dual-Stream framework partially restored CN specificity. While not matching the near-perfect specificity of pure volumetric models, it established a better balance across the imbalanced 7-class distribution (overall Top-1 BACC of 62%). This gain came at the cost of reduced AD Top-1 sensitivity compared with the fine-scale stream alone, and ECE remained close to the main baselines rather than being the best among ablations.

Table 2. Ablation study on the OOD test set (1,304 subjects). Top-1/Top-2 per-class Sensitivities, BACC, and ECE (%). 95% bootstrap CIs are provided in brackets. ↑=higher-is-better, ↓=lower-is-better. Top-1 and Top-2 best results are respectively in green and blue. Best and second-best overall are respectively **bold** and underlined. DiffAttn denotes Differential Attention; HL denotes Hierarchical Loss.

Model Configuration	Params	ECE ↓	Top	BACC ↑	AD ↑	CN ↑	DLB ↑	bvFTD ↑	PNFA ↑	SD ↑	PSP ↑
MedViT-3D (Vanilla)	34.70M	49	@1	59 [56-63]	71 [66-75]	**88** [85-90]	6 [0-14]	**62** [52-72]	46 [30-63]	66 [51-80]	78 [67-87]
			@2	78 [74-81]	90 [88-93]	**96** [95-98]	26 [14-39]	**81** [72-89]	68 [52-83]	93 [85-100]	90 [82-96]
+ RoPE	34.70M	48	@1	57 [53-61]	72 [68-76]	86 [83-89]	9 [2-18]	**62** [53-73]	27 [13-42]	66 [51-80]	76 [66-86]
			@2	77 [73-80]	**91** [88-93]	**96** [94-97]	26 [14-39]	79 [70-87]	65 [49-80]	93 [85-100]	90 [82-96]
+ RoPE + DiffAttn	31.57M	**39**	@1	57 [53-60]	75 [71-79]	49 [45-54]	6 [0-14]	**62** [52-72]	62 [46-78]	64 [49-78]	78 [67-87]
			@2	75 [72-78]	90 [88-93]	72 [68-75]	21 [10-33]	74 [65-83]	**87** [74-97]	**95** [88-100]	88 [80-95]
+ RoPE + DiffAttn + HL	31.57M	41	@1	60 [56-63]	**76** [72-80]	49 [45-53]	6 [0-14]	58 [47-68]	**73** [58-87]	82 [70-93]	75 [64-85]
			@2	75 [72-78]	**91** [88-93]	72 [69-76]	15 [6-26]	74 [65-83]	84 [71-95]	**95** [88-100]	**92** [86-98]
Dual-Stream (Ours) (MLP + RoPE + DiffAttn + HL + SVM) +`#{seg params}`	31.57M	47	@1	**62** [58-66]	59 [55-63]	63 [59-67]	**21** [10-33]	**62** [52-72]	60 [43-76]	**86** [75-96]	**84** [74-92]
			@2	**82** [79-86]	78 [75-82]	84 [80-87]	**83** [71-93]	76 [66-84]	73 [58-87]	**95** [88-100]	88 [80-95]

`#{seg params}` depends on the segmentation pipeline used to obtain the volumes, here [6].

4.2 Comparison with other Baseline Architectures

Baselines evaluations (Table 3) highlight the global versus fine-scale dichotomy. The SVM demonstrates strong CN specificity but fails on focal (PNFA) or diffuse (DLB) variants. Conversely, deep 3D architectures mitigate this limitation, yielding more distributed sensitivity across the 7 classes.

Table 3. Comparison of the proposed Dual-Stream framework against baselines on the OOD test set (1,304 subjects). Results present Top-1 and Top-2 per-class, BACC, and ECE (%). 95% bootstrap confidence intervals are in brackets. Top-1 and Top-2 best results are respectively in green and blue. Best and second-best are respectively **bold** and underlined.

Method	Params	ECE ↓	Top	BACC ↑	AD ↑	CN ↑	DLB ↑	bvFTD ↑	PNFA ↑	SD ↑	PSP ↑
Lifespan Tree [7]	`#{seg params}`	N/A	@1	55 [52-59]	44 [39-48]	74 [70-78]	**32** [19-46]	43 [33-54]	33 [18-48]	**91** [81-98]	71 [61-82]
			@2	71 [68-75]	60 [56-65]	80 [77-83]	72 [59-85]	59 [48-69]	50 [34-66]	**95** [88-100]	82 [72-91]
SVM (Volumes Only)	`#{seg params}`	**47**	@1	46 [42-49]	43 [38-47]	**97** [96-98]	0 [0-0]	51 [41-61]	17 [6-30]	57 [41-72]	58 [46-70]
			@2	65 [61-69]	77 [73-81]	**100** [99-100]	13 [4-23]	70 [60-79]	43 [27-58]	75 [61-87]	79 [69-88]
3D ResNet-50 [15]	46.17M	49	@1	56 [52-60]	68 [63-72]	93 [91-95]	2 [0-7]	60 [50-70]	24 [11-39]	66 [51-80]	79 [69-88]
			@2	76 [73-80]	89 [86-92]	98 [97-99]	28 [15-41]	78 [69-86]	59 [43-75]	91 [81-98]	**91** [84-97]
3D Swin DPL [24]	41.13M	48	@1	56 [52-60]	**71** [66-75]	81 [77-84]	4 [0-11]	**64** [54-74]	30 [15-46]	70 [56-84]	70 [59-81]
			@2	74 [71-78]	**91** [88-93]	94 [92-96]	24 [12-37]	**84** [77-92]	49 [32-66]	**95** [88-100]	84 [74-92]
MedViT-3D (Vanilla)	34.70M	49	@1	59 [56-63]	**71** [66-75]	88 [85-90]	6 [0-14]	62 [52-72]	46 [30-63]	66 [51-80]	78 [67-87]
			@2	78 [74-81]	90 [88-93]	96 [95-98]	26 [14-39]	81 [72-89]	68 [52-83]	93 [85-100]	90 [82-96]
Ours	31.57M +`#{seg params}`	**47**	@1	**62** [58-66]	59 [55-63]	63 [59-67]	21 [10-33]	62 [52-72]	**60** [43-76]	86 [75-96]	**84** [74-92]
			@2	**82** [79-86]	78 [75-82]	84 [80-87]	**83** [71-93]	76 [66-84]	**73** [58-87]	**95** [88-100]	88 [80-95]

`#{seg params}` depends on the segmentation pipeline used to obtain the volumes, here [6].

By bridging these representational scales, our Dual-Stream framework achieves higher overall Top-2 BACC than the Lifespan Tree baseline and single-stream architectures (82%). While the Lifespan Tree maintains a Top-1 sensitivity advantage on variants with highly specific regional volume loss (such as SD), our model exhibits strong Top-2 sensitivity recovery. This suggests that even when hesitating on the primary diagnosis due to structural overlaps, our network frequently retains the correct pathology among its top predictions. Furthermore, despite its multi-scale complexity, the framework remains computationally tractable, yielding an inference time of ~ 1 second per 3D scan.

Confusion matrices (Fig. 2) highlight these trade-offs. The SVM accurately identifies CN subjects but misclassifies several pathological cases as CN or AD. Conversely, the fine-scale stream strongly detects rare FTDs, yet its horizontal spread across the CN row is consistent with the *saliency trap*: Differential Attention may mistake normal anatomical variability for pathology. The Dual-Stream framework partially filters this noise through global priors, but also increases DLB over-prediction for some AD and CN subjects.

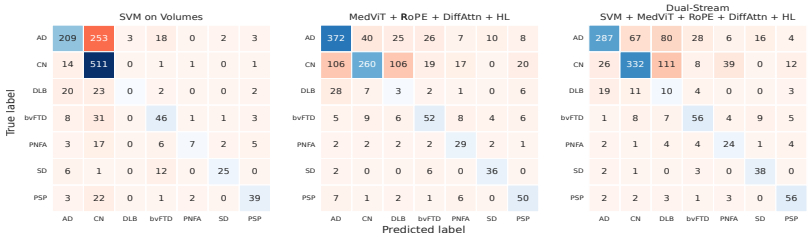

Fig. 2. Confusion matrices evaluated on the OOD test set. See text for detailed analysis.

5 Discussion

The *saliency trap* in Deep Neuroimaging. We observe a *saliency trap*-like behavior associated with Differential Attention in structural MRI. The configuration incorporating Differential Attention exhibits higher sensitivity to focal lobar degenerations (notably PNFA and SD), accompanied by a shift toward over-detection and increased misclassification of normal anatomical variability in healthy subjects. This suggests a limitation of single-stream processing in imbalanced data and motivates our Dual-Stream design [13,9].

Synergy of Multi-Scale Representations. Our framework mitigates this trade-off through representational complementarity. Trained on regional volumes, the SVM acts as a coarse anatomical prior, preserving high CN specificity and providing complementary low-frequency structural information. The Late-Fusion MLP then ensembles these stable priors with the fine-scale MedViT predictions. This yields an improved empirical trade-off, mitigating some Transformer false positives while retaining true-positive detections on challenging focal variants.

Limitations and Future Work. First, relying on automated volumetry introduces morphological uncertainties in highly atrophied brains. Second, the present ablation does not fully disentangle the respective contributions of the volumetric prior, the fusion module, and each transformer modification; controls such as vanilla MedViT+SVM fusion or alternative tabular learners would further clarify the source of the gains. Consequently, the present results establish the performance of the complete Dual-Stream framework but do not isolate the contribution of each individual component to the final performance. Third, our framework is trained entirely from scratch. Recent literature suggests that leveraging self-supervised 3D foundation models (e.g., OpenMind [35]) could drastically improve data efficiency and OOD robustness. Although holding out complete cohorts provides a stringent test of cross-dataset generalization, performance was not stratified by site, age, or sex. Consequently, the present analysis cannot isolate the contribution of demographic and acquisition differences to OOD performance variations. Calibration was assessed using ECE only; classwise reliability, Brier score, and post-hoc calibration were not evaluated and remain directions for future work. Finally, Top-1 DLB detection remains poor, and the confusion matrices show increased DLB over-prediction for some AD and CN subjects. This may reflect OOD cohort/protocol shift or residual confounding in this rare class, and future multi-modal fusion integrating functional imaging (e.g., PET) or neuropsychological assessments is necessary to close this diagnostic gap.

6 Conclusion

We introduced a Dual-Stream Late-Fusion framework for complex differential diagnosis of neurodegenerative diseases. We empirically observed that exploiting Differential Attention in 3D Vision Transformers can increase sensitivity to

rare focal variants. By coupling this sensitive fine-scale stream with global volumetric priors to mitigate the observed *saliency trap*-like behavior, our method establishes competitive out-of-domain performance on a large multi-centric 7-class dataset. Our approach offers a clinically motivated multi-scale framework for the differential diagnosis of neurodegenerative diseases, with calibration and cohort-specific failure modes remaining important directions for future work.

References

1. Ahmed, S.R., et al.: Alzheimer's disease prediction using MRI-PET fusion through multi-modal deep learning with late fusion techniques. In: 2025 3rd International Conference on Cyber Resilience (ICCR). pp. 1–8 (2025)
2. Avants, B.B., et al.: A reproducible evaluation of ANTs similarity metric performance in brain image registration. NeuroImage **54**(3), 2033–2044 (2011)
3. Beekly, D.L., et al.: The National Alzheimer's Coordinating Center (NACC) Database: The Uniform Data Set. Alzheimer Disease & Associated Disorders **21**(3), 249 (Jul 2007)
4. Boeve, B.F., et al.: Studying the natural history of frontotemporal lobar degeneration (FTLD): The ARTFL LEFFTDS longitudinal FTLD (ALLFTD) protocol. Alzheimer's & Dementia **16**(S6), e045482 (2020)
5. Cardoso, M.J., et al.: Monai: An open-source framework for deep learning in healthcare. arXiv:2211.02701 (2022)
6. Coupé, P., et al.: AssemblyNet: A large ensemble of CNNs for 3D whole brain MRI segmentation. NeuroImage (2020)
7. Coupé, P., et al.: Lifespan tree of brain anatomy: Diagnostic values for motor and cognitive neurodegenerative diseases. Human Brain Mapping **46**(13), e70336 (2025)
8. Di Martino, A., et al.: Enhancing studies of the connectome in autism using the autism brain imaging data exchange II. Scientific Data **4**(1), 170010 (2017)
9. Dosovitskiy, A., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2021)
10. Ellis, K.A., et al.: The Australian Imaging, Biomarkers and Lifestyle (AIBL) study of aging: Methodology and baseline characteristics of 1112 individuals recruited for a longitudinal study of Alzheimer's disease. International Psychogeriatrics **21**(4), 672–687 (2009)
11. Fonov, V.S., et al.: Unbiased average age-appropriate atlases for pediatric studies. NeuroImage **54**(1), 313–327 (2011)
12. Frontotemporal Lobar Degeneration Neuroimaging Initiative: Frontotemporal Lobar Degeneration Neuroimaging Initiative (FTLDNI / NIFD) (2009)
13. Geirhos, R., et al.: Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In: ICLR (2019)
14. Guo, C., et al.: On calibration of modern neural networks (2017), https://arxiv.org/abs/1706.04599
15. He, K., et al.: Deep Residual Learning for Image Recognition. In: Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778. CVPR '16, IEEE (2016)
16. Kotikalapudi, R.: IXI – Information extraction from Images — cortical volume (June 2024), dataset. DOI: 10.5281/zenodo.11635168

17. Malone, I.B., et al.: MIRIAD—Public release of a multiple time point Alzheimer's MR imaging dataset. NeuroImage **70**, 33–36 (2013)
18. Manzari, O.N., et al.: Medical Image Classification with KAN-Integrated Transformers and Dilated Neighborhood Attention. arXiv:2502.13693 (2025)
19. Marcus, D.S., et al.: Open Access Series of Imaging Studies (OASIS): Cross-sectional MRI Data in Young, Middle Aged, Nondemented, and Demented Older Adults. Journal of Cognitive Neuroscience **19**(9), 1498–1507 (2007)
20. Mazziotta, J., et al.: A probabilistic atlas and reference system for the human brain: International Consortium for Brain Mapping (ICBM). Philosophical Transactions of the Royal Society B: Biological Sciences **356**(1412), 1293–1322 (Aug 2001)
21. Mueller, S.G., et al.: Ways toward an early diagnosis in Alzheimer's disease: The Alzheimer's Disease Neuroimaging Initiative (ADNI). Alzheimer's & Dementia **1**(1), 55–66 (2005)
22. Muminov, A., Pham, A.: CytoDINO: Risk-Aware and Biologically-Informed Adaptation of DINOv3 for Bone Marrow Cytomorphology. arXiv:2512.17930 (2025)
23. Navet, E., et al.: On the stability and robustness of vision transformers for neurodegenerative disease classification. In: Medical Imaging with Deep Learning - Validation Papers (2026)
24. Nguyen, H.D., et al.: 3D Transformer Based on Deformable Patch Location for Differential Diagnosis Between Alzheimer's Disease and Frontotemporal Dementia. In: Machine Learning in Medical Imaging. pp. 53–63. Springer Nature Switzerland, Cham (2024)
25. Ofori, E., et al.: Parkinson's disease biomarkers program brain imaging repository. NeuroImage **124**, 1120–1124 (Jan 2016)
26. Park, D.C., et al.: The Dallas Lifespan Brain Study: A Comprehensive Adult Lifespan Data Set of Brain and Cognitive Aging. Scientific Data **12**(1), 846 (2025)
27. Park, N., Kim, S.: How do vision transformers work? In: International Conference on Learning Representations (2022), https://openreview.net/forum?id=D78Go4hVcxO
28. Payakachat, N., Tilford, J.M., Ungar, W.J.: National database for autism research (NDAR): Big data opportunities for health services research and health technology assessment. PharmacoEconomics **34**(2), 127–138 (2016)
29. Singh, J., et al.: Batch-balanced focal loss: A hybrid solution to class imbalance in deep learning. Journal of Medical Imaging **10**(5), 051809 (2023)
30. Sriwiboon, N.: DSC-CNN: A dual-stream CNN with cognitive embedding fusion for early alzheimer's diagnosis. Research Square (2025)
31. Su, J., et al.: RoFormer: Enhanced transformer with rotary position embedding. arXiv:2104.09864 (2023)
32. Taylor, J.R., et al.: The Cambridge Centre for Ageing and Neuroscience (Cam-CAN) data repository: Structural and functional MRI, MEG, and cognitive data from a cross-sectional adult lifespan sample. NeuroImage **144**, 262–269 (2017)
33. Tustison, N., et al.: N4ITK: Improved N3 bias correction. In: ISBI (2010)
34. University of California, San Francisco, National Institutes of Health: 4-Repeat Tauopathy Neuroimaging Initiative (4RTNI), Observational MRI and Biomarker Study. https://clinicaltrials.gov/study/NCT01804452 (2011–2016)
35. Wald, T., et al.: An OpenMind for 3D medical vision self-supervised learning (2025). https://doi.org/10.48550/arXiv.2412.17041
36. Wald, T., et al.: Primus: Enforcing Attention Usage for 3D Medical Image Segmentation. arXiv:2503.01835 (2025)

37. Wang, P., Zheng, W., Chen, T., Wang, Z.: Anti-oversmoothing in deep vision transformers via the fourier domain analysis: From theory to practice (2022), https://arxiv.org/abs/2203.05962
38. Wei, D., et al.: Structural and functional brain scans from the cross-sectional Southwest University adult lifespan dataset. Scientific Data **5**(1), 180134 (2018)
39. Ye, T., et al.: Differential transformer. In: The Thirteenth International Conference on Learning Representations (2025)