Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

An explainable attention-based multimodal MRI framework for Parkinson's disease classification: a pilot feasibility and internal validation study

Zohreh Ganji
Mashhad University of Medical Sciences
Farzaneh Nikparast
Mashhad University of Medical Sciences
Naser Shoeibi
Mashhad University of Medical Sciences
Ali Shoeibi
Mashhad University of Medical Sciences
Emad Shieali
Mashhad University of Medical Sciences
Seyyed Mohammad Hashemi
Mashhad University of Medical Sciences
Hoda Zare
zareh@mums.ac.ir

Mashhad University of Medical Sciences

Research Article

Keywords: Parkinson's Disease, Neuromelanin-sensitive MRI, Arterial Spin Labeling (ASL), Perfusion MRI, Deep Learning, Explainable AI (XAI)

Posted Date: June 5th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9570147/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background

Early diagnosis of Parkinson's disease (PD) is challenging due to clinical heterogeneity and subtle MRI changes. This study explored the feasibility of an attention-enhanced hybrid deep learning (DL) framework integrating structural MRI, neuromelanin-sensitive MRI (NM-MRI), and arterial spin labeling (ASL) perfusion to diagnose PD.

Methods

In this retrospective study, 54 participants (29 PD patients, 25 healthy controls) underwent multimodal neuroimaging (T1-weighted, NM-MRI, and ASL). A hybrid framework utilizing a 2.5D ResNet18 backbone augmented with Convolutional Block Attention Modules (CBAM) and a Multi-head Attention tabular branch was developed. Post-hoc explainable AI (XAI) was validated using fidelity and reliability metrics.

Results

The multimodal model achieved an AUC of 0.91 (95% CI, 0.87–0.96), outperforming unimodal models (T1-only: 0.81, NM-MRI-only: 0.77, ASL-only: 0.79). Feature analysis ranked putaminal and caudate cerebral blood flow and nigral neuromelanin as dominant features. Saliency maps showed high anatomical coverage and reliability in the caudate, pallidum, and thalamus for PD patients.

Conclusions

This pilot study suggests that an attention-based multimodal DL framework can extract PD-related signals more effectively than unimodal approaches, even in small cohorts. The model provides anatomically consistent explanations, though larger multi-center studies are needed to confirm generalizability.

1. INTRODUCTION

Parkinson's disease (PD) is currently the second most prevalent neurodegenerative disorder globally, with an escalating incidence fueled by aging demographics and complex environmental interactions. The pathological hallmark of PD is the progressive loss of dopaminergic neurons within the substantia nigra pars compacta (SNpc), resulting in striatal dopamine depletion and the subsequent manifestation of classic motor symptoms such as bradykinesia, rigidity, and postural instability. However, it is now widely recognized that non-motor features, including REM sleep behavior disorder (RBD), autonomic dysfunction, and cognitive decline—often precede motor onset by decades. This clinical progression aligns with the Braak hypothesis, which describes a caudal-to-rostral propagation of α-synuclein

aggregates from brainstem nuclei, such as the locus coeruleus, toward midbrain and cortical regions (1, 2).

Despite the refinement of clinical diagnostic criteria, achieving high diagnostic accuracy remains elusive, particularly during the prodromal phase. Recent clinicopathologic studies reveal that misclassification rates persist between 13% and 18%, frequently involving atypical parkinsonian syndromes such as multiple system atrophy (MSA) and progressive supranuclear palsy (PSP) (3, 4). In community settings, the initial misdiagnosis rate can exceed 25%, leading to significant therapeutic delays (5). This diagnostic gap is particularly detrimental to disease-modifying trials, which require the accurate identification of participants in pre-motor phases when neuroprotective interventions are most effective (6, 7).Advanced neuroimaging offers a potential solution to these challenges, yet single-modality approaches have distinct limitations. Conventional T1-weighted structural MRI often lacks the sensitivity to detect early-stage changes at the individual level (8). Neuromelanin-sensitive MRI (NM-MRI) has emerged as a robust proxy for catecholaminergic integrity, showing significant signal reduction in the SNpc and locus coeruleus (9, 10).

However, NM-MRI is constrained by field-strength dependence and variable patterns across atypical syndromes. Similarly, Arterial Spin Labeling (ASL) perfusion MRI reveals critical cerebral blood flow (CBF) alterations—potentially reflecting early neurovascular uncoupling—but findings are often confounded by medication effects, transit delays, and motion artifacts (11, 12).

While multimodal integration—combining macro-structural changes (T1), cellular proxies (NM-MRI), and perfusion dynamics (ASL)—offers theoretical synergy, empirical evidence remains fragmented. Current registration and scanner variability often hinder the reproducibility of integrated models. Furthermore, although deep learning (DL) has shown promise in automated pattern detection, most models remain "black boxes." The clinical adoption of AI in neuroradiology is stifled by these opacity concerns and the risk of overfitting on small, heterogeneous clinical datasets (13–16).

It is important to note that the framework proposed here is designed as a knowledge-informed fusion model rather than a purely data-discovery tool. By extracting quantitative region-of-interest (ROI) metrics (e.g., CBF, neuromelanin CNR/SNR) a priori, we explicitly encode expert radiological priors into the system. The contribution of this work is not to replace the radiologist’s feature extraction, but to develop an attention-based mechanism that dynamically weighs these established biomarkers against raw spatial patterns (from the CNN branch) to improve classification confidence. This approach mitigates the “black box” nature of pure deep learning models, particularly in small-data regimes where end-to-end feature learning is prone to instability and overfitting.

Within this context, we developed an attention-enhanced hybrid DL framework that combines a 2.5D convolutional neural network (CNN) with a tabular branch encoding imaging-derived quantitative metrics and age. Our primary aim was to test, in a limited single-center cohort, whether such a multimodal architecture can extract PD-related pathological signal more effectively than unimodal baselines. A secondary aim was to systematically characterize the model’s behavior using multiple XAI techniques

and quantitative indices of anatomical fidelity and attribution reliability, while explicitly discussing methodological and interpretative limitations.

Given the modest sample size and retrospective design, we frame this work as a feasibility and hypothesis-generating study rather than definitive evidence for a clinical diagnostic tool.

2. MATERIALS AND METHODS

2.1. Study Population and Ethical Considerations

This retrospective study was approved by the Institutional Review Board of Mashhad University of Medical Sciences (Approval No. : IR.MUMS.MEDICAL.REC.1403.319) and was conducted in accordance with the principles of the Declaration of Helsinki and local regulatory requirements. All participants were enrolled in the Persian Organizational Cohort Study (POCS), a longitudinal population-based cohort, and had previously provided written informed consent for participation and linkage of clinical data. For the present secondary analysis of neuroimaging and pre-existing cohort data, the IRB granted a waiver of additional written consent.

The study cohort comprised 54 individuals: 29 patients with idiopathic PD and 25 age- and sex-matched healthy controls (HC). PD diagnosis was established by movement disorder specialists according to the MDS clinical diagnostic criteria. Exclusion criteria included major structural brain abnormalities on MRI, prior neurosurgery, and significant comorbid neurological or psychiatric disorders. Clinical metadata, including extensive blood and urine panels, were available from POCS but were not used directly as inputs to the deep learning model in this study.

Demographic characteristics are summarized in Table 1. The PD group was significantly older than the HC group (58.52 ± 11.13 vs 52.68 ± 5.79 years, P = .022). Gender distribution was balanced (P = .824). Because age is both a risk factor for PD and a strong determinant of brain structure and perfusion, we treated age as a primary covariate both in classical statistical analyses and as an explicit feature in the DL framework. Information on disease duration, motor severity (e.g., UPDRS), and antiparkinsonian medication state (e.g., ON/OFF levodopa) was available clinically but was not comprehensively modeled in the DL analysis; this limitation is addressed in the Discussion.

2.2. MRI Acquisition and Preprocessing

All MRI examinations were conducted on a Philips Ingenia Ambition S 1.5T scanner using a 20-channel head coil. The protocol was optimized for field stability, parallel imaging acceleration (SENSE), and patient tolerability in PD cohorts prone to tremor. Three core sequences were acquired: structural T1-weighted (sMRI), neuromelanin-sensitive (NM-MRI), and pseudo-continuous arterial spin labeling perfusion (pCASL ASL-MRI).

Structural imaging used a 3D T1-weighted Turbo Field Echo (TFE) sequence with isotropic 1 mm resolution (TR = 1900 ms, TE = 2.5 ms, flip angle = 90°, FOV = 256×256 mm², 176 slices, acquisition time

~ 4:30 min), enabling precise volumetry and multi-planar reconstruction.

NM-MRI employed a 2D T1-weighted Turbo Spin Echo sequence with on-resonance magnetization transfer contrast prepulse to enhance neuromelanin hyperintensity in substantia nigra and locus coeruleus. Parameters balanced high in-plane resolution (0.67×0.67 mm²), SNR (NSA = 5), and short duration via SENSE acceleration (TR = 600 ms, TE = 12 ms, slice thickness = 2.8 mm with 0.3 mm gap, FOV = 150×150 mm², 11 slices, acquisition time ~ 5:23 min).

Perfusion imaging utilized pCASL with single-shot EPI readout, post-label delay 1800 ms, labeling duration 1650 ms, and SENSE factor 2.5 for motion robustness (TR = 4000 ms, TE = 12 ms, voxel size 3.0×3.0×4.0 mm³, FOV = 150×150 mm², 11 slices, 30 label/control pairs, acquisition time ~ 4:15 min).

No explicit pharmacologic washout was enforced; patients were scanned in their routine clinical medication states. Because antiparkinsonian medications can alter CBF and possibly NM-MRI signal, this is considered a potential confounder and is further addressed in the Limitations.

Preprocessing was implemented in Python using standard neuroimaging libraries (ANTs, FSL, and PyTorch-based tools). All preprocessing steps were performed strictly on a per-subject basis, using only that subject's own images and without computing any dataset-level statistics, to avoid data leakage across cross-validation folds. The pipeline comprised the following steps:

- Image quality control: All MRI datasets were reviewed using a standardized visual quality control (QC) procedure. For each subject, T1-weighted, NM-MRI and ASL images were visually inspected by an experienced neuroradiologist and a medical physicist for susceptibility artefacts, signal dropout, slice corruption and motion-related blurring. Scans with gross artefacts that could compromise visualization of the basal ganglia or midbrain were to be excluded. In this cohort, no scans met the exclusion criteria on visual inspection and therefore no subjects were excluded on the basis of image quality.
- Bias Field Correction: N4ITK algorithm addressed RF inhomogeneities via B-spline modeling, preventing intensity artifacts that could bias deep learning features (17).
- Brain Extraction: Brain extraction was performed using HD-BET, a deep learning–based tool trained on heterogeneous brain pathologies (18). HD-BET was run independently for each subject to generate a brain mask, which was then applied to all modalities (T1, NM-MRI, ASL) of that subject. No group-level templates or derived masks were shared across subjects.
- Intra-Subject Registration: For each subject, NM-MRI and ASL images were rigidly registered to the corresponding T1-weighted volume using a mutual information cost function. This ensured anatomical alignment across modalities while remaining entirely within-subject, thereby avoiding any cross-subject interaction at this stage.
- Intensity Normalization (Subject-wise Z-scoring): To standardize the dynamic range while preserving relative contrast, we applied z-score normalization separately for each subject and each modality. Specifically, for each brain-extracted volume, voxel intensities within the brain mask were normalized

to zero mean and unit standard deviation computed from that subject's own brain voxels. This subject-wise normalization avoids leakage because no global mean or variance was estimated across subjects.

- Standard Space Registration and Atlas Propagation: Each subject's T1-weighted image was non-linearly registered to the MNI152 template (2 mm) using the ANTs SyN algorithm. The resulting transformation was then applied to that subject's NM-MRI and ASL volumes. The Harvard–Oxford subcortical atlas, defined in MNI space, was warped into each subject's native space using the inverse transformation. This atlas propagation was used solely to (i) define regions of interest (ROIs) for quantitative feature extraction (e.g., ROI-wise CBF, neuromelanin SNR/CNR) and (ii) quantify saliency coverage in XAI analyses. The atlas did not directly enter the training of the deep learning model; it served as an anatomical reference for ROI-based measurements, thus avoiding circular analysis in the learning process.
- ASL-specific Preprocessing and CBF Quantification: For ASL data, label/control pairs were processed using the BASIL toolbox from FSL to obtain quantitative cerebral blood flow (CBF) maps. For each subject, motion correction and pairwise label–control subtraction were performed, followed by calibration using a proton density–weighted reference image. Quantification was carried out with BASIL's single-compartment kinetic model, using standard assumptions for labeling efficiency, blood T1, and partition coefficient in line with current consensus recommendations for pCASL (19). Grey-matter and white-matter partial volume estimates derived from the structural T1 image were used by BASIL to perform partial-volume–corrected CBF estimation. This yielded whole-brain, grey-matter, and white-matter CBF maps in native ASL space for each subject. The resulting quantitative CBF maps were then co-registered to the subject's T1-weighted image and subsequently warped to MNI space using the transforms described above. All BASIL processing and quantification were performed independently per subject, without using any group-level statistics, thereby avoiding cross-subject information leakage at the preprocessing stage.
- ROI-Aware Slice Selection and 2.5D Input Formation: For each subject and modality, we selected central axial slices that maximized coverage of Parkinson-relevant structures, including the substantia nigra pars compacta (SNpc), basal ganglia (Caudate, Putamen, Pallidum), and Thalamus. Based on visual inspection and anatomical landmarks, five adjacent axial slices centered on the SNpc were chosen for the primary 2.5D input.

Given the NM-MRI slice thickness of 2.8 mm, a 5-slice stack corresponds to a rostro-caudal coverage of approximately 14 mm. Anatomical studies indicate that the SNpc height is typically 8–12 mm, while adjacent striatal structures (Caudate, Putamen) extend up to about 20 mm. A 5-slice stack therefore: (i) ensures complete coverage of the nigrostriatal complex without truncation; (ii) includes sufficient surrounding context from the Pallidum and Thalamus, which are relevant for assessing potential neurovascular decoupling; and (iii) balances anatomical coverage with computational efficiency. Smaller stacks (e.g., 3 slices) risk cutting through the ROIs due to head tilt or minor misalignment, whereas larger stacks (> 7 slices) introduce redundant background tissue and increase memory usage without clear diagnostic benefit for the targeted subcortical regions.

For the CNN input, the five selected slices from each modality were stacked along the channel dimension (e.g., 3 modalities × 5 slices = 15 channels), forming a 2.5D representation per subject. This entire slice-selection and stacking process was performed independently for each subject, prior to model training, without any information exchange between subjects.

2.3. Biomarker Computation (Tabular Features)

Beyond raw images, we derived a set of quantitative MRI-based biomarkers for the tabular branch:

- Neuromelanin metrics: Semi-automated region-growing with intensity thresholding was used to delineate SNpc on NM-MRI slices. For each subject, signal-to-noise ratio (SNR) and contrast-to-noise ratio (CNR) were computed for the SNpc (and, where reliably visible, locus coeruleus) as:

- $SNR = \mu_{SN}/\sigma_{noise}$
- $CNR = (\mu_{SN} - \mu_{reference})/\sigma_{noise}$

where μ denotes mean intensity in the ROI, $\mu_{reference}$ is mean intensity in adjacent white matter, and σ_{noise} is the standard deviation of background noise.

- Perfusion metrics: Quantitative CBF maps were generated from ASL data using the BASIL toolbox (FSL), as described above. For each subject, BASIL provided voxel-wise CBF estimates for the whole brain, with additional grey-matter and white-matter–specific maps obtained via partial-volume correction. Using the Harvard–Oxford subcortical atlas warped into each subject's native space, mean CBF was extracted from atlas-defined ROIs in each hemisphere (Putamen, Caudate, Thalamus, and other basal ganglia structures as needed). In addition to regional subcortical CBF values used in the hybrid deep learning model, whole-grey-matter and whole-white-matter CBF were also computed for descriptive and sensitivity analyses.
- Age: Chronological age was included as an additional scalar feature in the tabular branch to account for age-related effects, with the understanding that this can both help control for and potentially introduce proxy learning if not interpreted cautiously.

These imaging-derived features were computed per subject and were temporally and anatomically synchronized with the MRI data used in the CNN branch, thereby reflecting the same disease state at the time of scanning. Although comprehensive blood and urine laboratory panels were available in the parent cohort, they were not integrated into the current DL model; the present framework, however, is structured to accommodate such systemic biomarkers in future extensions.

2.4. Multimodal Deep Learning Framework

We implemented a seven-stage development and evaluation pipeline to systematically progress from unimodal baselines to multimodal fusion and interpretability (Fig. 1).

Stage 1 – Unimodal Benchmarks:

We first trained separate models on individual modalities to establish baseline performance. A 2D ResNet18 (ImageNet-pretrained) processed single-modality inputs: (i) T1 slices, (ii) NM-MRI slices or substantia nigra patches, and (iii) tabular ASL CBF features alone via a small multilayer perceptron (MLP). This provided a frame of reference for assessing any added value from multimodal fusion.

Stage 2–2.5D CNN Architecture:

To incorporate limited volumetric context without the computational cost of full 3D CNNs, we adopted a 2.5D strategy: five adjacent axial slices per modality were stacked to form a multi-channel input. The first convolutional layer of ResNet18 was adapted to accommodate the increased channel dimension. This allowed the model to capture local through-plane patterns in nigrostriatal and basal ganglia regions.

Stage 3 – Multimodal Fusion Architecture:

The full hybrid architecture consisted of:

- Image branch: a ResNet18 backbone augmented with Convolutional Block Attention Modules (CBAM) for channel and spatial attention, yielding a 512-dimensional embedding from the stacked multimodal 2.5D input.
- Tabular branch: an MLP followed by a multi-head self-attention layer that projected the imaging-derived quantitative features (NM-MRI metrics, regional CBF measures, and age) into a learned embedding space.
- Fusion and classifier: late fusion concatenated the image embedding with the attention-projected tabular embedding and the raw tabular vector, forming a joint representation. A two-layer MLP with ReLU activation and dropout (0.3–0.4) produced the final PD vs HC probability.

Stage 4 – Fusion Strategy and Backbone Comparison:

We compared:

- Early fusion (concatenating modalities at the input level),
- Standard late fusion (concatenating post-encoder features without attention),
- Attention-enhanced late fusion (the final model described above).

Alternative backbones (ResNet34, DenseNet121, EfficientNet-B3) were evaluated in late-fusion configurations. Given the small sample, models with substantially larger parameter counts showed signs of overfitting and unstable cross-validation performance. ResNet18 with CBAM and attention-based fusion (ResNet18-Att) offered a favorable trade-off between capacity and stability.

Stage 5 – Probability Calibration and Thresholding:

We applied isotonic regression to calibrate output probabilities, using held-out validation folds within cross-validation. Thresholds for binary classification were derived using Youden’s index but are reported

primarily for descriptive purposes, given the cohort size.

Stage 6 – Post-hoc Explainability (Images):

To interrogate the CNN's spatial focus:

- Grad-CAM was computed from the last convolutional block, with multi-layer aggregation and mild gamma correction, to produce class-specific saliency maps.
- Integrated Gradients (50 steps) were computed from a zero-baseline using Captum to obtain pixel-wise attributions.
- In exploratory analyses, small-window occlusion and LIME-based perturbation were used qualitatively on selected cases, but only Grad-CAM and Integrated Gradients were taken forward to quantitative ROI coverage analysis.

Stage 7 – Post-hoc Explainability (Tabular Features):

KernelSHAP with a background set of randomly sampled HC and PD subjects was used to estimate feature attributions for the tabular branch and for the fused representation at the decision layer. We focused on mean absolute SHAP values for global importance ranking and SHAP beeswarm plots for directionality.

We note that SHAP and gradient-based methods are subject to well-documented limitations in stability, fidelity, and susceptibility to modeling artifacts. Consequently, our XAI results are interpreted as exploratory consistency checks rather than definitive causal explanations.

2.5. Quantitative XAI Validation: Fidelity and Reliability

To move beyond purely visual inspection of saliency maps, we defined two summary metrics:

- Fidelity (ROI coverage): For each subject and each method (Grad-CAM, Integrated Gradients), we thresholded saliency maps to retain the most relevant voxels (top quantile of attribution values) and computed the overlap (percentage of thresholded voxels) within nine a priori defined PD-relevant ROIs from the Harvard–Oxford subcortical atlas: Brainstem, Caudate (L/R), Putamen (L/R), Pallidum (L/R), and Thalamus (L/R). Fidelity was operationalized as the extent to which PD subjects showed higher attribution coverage in these ROIs than HCs. Group comparisons were performed using non-parametric tests, with age entered as a covariate in complementary regression models where appropriate. These analyses do not prove causality but test for alignment with known neuroanatomical patterns.
- Reliability (Stability Index): For each ROI and each diagnostic group, we computed the mean (μ) and standard deviation (σ) of saliency coverage across subjects and defined a simple Reliability Index RI = μ / σ (equivalent to 1/CV when CV = σ/μ). Higher RI reflects more consistent attribution across individuals in that group. This index is heuristic and has no universally accepted cut-off; in our context, values > 2 suggested relatively stable localization patterns, while lower values indicated

heterogeneous or noisy attributions. We did not perform stress-testing of attribution robustness under adversarial perturbations or randomized baselines; these omissions are noted as limitations.

Given the modest cohort size and the ongoing debate about the robustness of XAI in medical imaging, we interpret both fidelity and reliability indices as descriptive rather than definitive indicators of model trustworthiness.

2.6. Training Procedure and Evaluation Metrics

Training was implemented in PyTorch on NVIDIA GPUs. The following protocol was used:

- Cross-validation and Prevention of Slice-level Data Leakage: We used stratified five-fold cross-validation at the subject level. Subjects, not slices, were the unit of splitting: in each fold, approximately 80% of subjects were used for training and 20% for validation, preserving the PD/HC ratio. Crucially, all 2.5D stacks (i.e., the five selected slices per modality) and all tabular features belonging to a given subject were assigned to the same fold. No subject contributed data to both the training and evaluation sets in any fold. This strategy prevents slice-level data leakage, where slices or partial volumes from the same brain might otherwise appear in both train and test sets and artificially inflate performance.
- Data Augmentation: Given the small sample size and the relatively small anatomical extent of the substantia nigra and basal ganglia, data augmentation was restricted to conservative, morphology-preserving transformations:
- Small in-plane rotations (up to ± 10°) were applied around the center of each slice to simulate minor differences in head positioning without distorting the intrinsic shape of the SNpc or basal ganglia.
- Small in-plane translations (a few pixels in x and y) were used to improve robustness to slight misalignments and residual registration errors.
- Mild intensity jitter (random scaling and offset within a narrow range) was applied to account for realistic variability in scanner gain and noise, while preserving relative contrast between gray matter, white matter, and neuromelanin-rich regions.

All augmentations were rigid (no scaling, no shearing, no non-linear warping) and were applied identically across modalities for a given subject and slice index, thereby maintaining multimodal spatial correspondence. Horizontal flips (left–right) were not used as an augmentation strategy in the final experiments, to avoid artificially disrupting potential lateralization patterns in PD. This conservative augmentation regime aims to increase robustness while respecting the anatomical integrity of the small deep-brain structures that are critical for PD assessment.

- Optimization and Metrics : Models were trained with the AdamW optimizer (initial learning rate 1×10^{-4}, weight decay 1×10^{-5}). Batch size was set to 64 where GPU memory allowed and reduced proportionally for larger backbones. Early stopping based on validation loss with a patience of 10 epochs was used to mitigate overfitting.

The primary evaluation metric was area under the receiver operating characteristic curve (AUC), computed from pooled cross-validated predictions. We report mean AUC across folds with an approximate 95% confidence interval derived from the variability of fold-wise AUCs. Given the non-independence induced by cross-validation and the limited sample size, these intervals should be regarded as approximate and potentially optimistic. Secondary metrics included accuracy, sensitivity, specificity at the Youden-optimal threshold, and the Brier score before and after isotonic regression calibration.

All analyses are internal to this single dataset; no external test set or independent validation cohort was available.

2.7. Statistical Analysis

Statistical analysis was performed in R version 4.5.0. Group differences in imaging-derived quantitative metrics were assessed using analysis of covariance (ANCOVA) or linear regression models with diagnostic group as the main factor and age entered as a covariate, given the significant age difference between PD and HC. Sex was tested as an additional covariate but did not materially change the results and was therefore not retained in the final models. Group×Age interaction terms were explored to assess differential aging effects, but these analyses were considered exploratory due to limited power.

Diagnostic performance was quantified using area under the ROC curve (AUC), sensitivity, and specificity based on the Youden-optimal threshold. Confidence intervals for AUC were derived from the variability across five cross-validation folds. DeLong’s test was used to compare AUCs between fusion strategies and backbone architectures, using pooled cross-validated predictions.

Given the pilot nature and small sample size of this study, p-values for group comparisons and ROI-wise XAI metrics were treated as exploratory and were not adjusted for multiple comparisons; findings should therefore be interpreted as hypothesis-generating rather than confirmatory.

3. RESULTS

3.1. Demographic Characteristics

The demographic analysis of the cohort (Table 1) indicated a significant age difference between the study groups (P = .022). The PD group had a mean age of 58.52 ± 11.13 years, while the HC group was younger (52.68 ± 5.79 years). To mitigate this potential confounder, age was incorporated as a primary covariate in all downstream deep learning architectures and statistical analyses. The gender distribution was statistically balanced between the two groups (P = .824), ensuring no sex-related bias in the model's training.

Table 1
Demographic Characteristics of the Study Population

Variable	PD (n = 29)	HC (n = 25)	P-value
Age (years)	58.52 ± 11.13	52.68 ± 5.79	0.022*
Gender (Male/Female)	17/12	14/11	.824

3.2. Learning Dynamics and Model Convergence

Across cross-validation folds, the ResNet18-Att model demonstrated stable convergence. Training and validation losses generally plateaued by approximately 30 epochs. Mean validation accuracy stabilized around 87.2% (± 3.4% across folds). Figure 2 shows the training and validation loss and accuracy curves for each of the five folds, illustrating monotonic loss decrease and steadily improving accuracy, with validation curves consistently tracking below training curves and indicating limited but controlled overfitting for the chosen architecture.

3.3. Modality Benchmarking and Fusion Performance

Modality Benchmarking and Fusion Performance

Figure 3 illustrates an ablation across modality combinations within the five-fold cross-validation framework. In the unimodal setting (Stage 1), T1-weighted MRI, NM-MRI, and ASL-based tabular CBF achieved mean AUCs of approximately 0.81, 0.77, and 0.79, respectively. Pairwise combinations (T1 + NM, T1 + ASL, NM + ASL) yielded intermediate performance, with T1 + ASL showing the highest AUC among the two-modality models. The full multimodal configuration (T1 + NM + ASL, ResNet18-Att) achieved the best performance (mean AUC ≈ 0.91, 95% CI: 0.87–0.96), indicating that each modality contributes complementary information and that three-way fusion provides the highest discriminative power in this dataset.

Backbone comparisons (also summarized in Fig. 4) suggested diminishing returns for deeper architectures in this sample: while mean AUCs for DenseNet121 and EfficientNet-B3 were numerically similar to ResNet18-based models, they exhibited wider fold-to-fold variability and faster overfitting. We therefore focused subsequent analyses on the ResNet18-Att configuration. Given the absence of external validation and the methodological constraints of cross-validated CI estimation, this performance differences should be viewed as provisional and specific to this dataset.

3.4. Probability Calibration and Overall Performance

Figure 5 provides a performance summary of the calibrated ResNet18-Att model. The ROC curve shows an AUC of 0.91, with the Youden-optimal operating point (red dot) corresponding to an accuracy of approximately 0.87, sensitivity of 0.90, and specificity of 0.84. The precision–recall curve indicates high precision across most of the recall range (average precision ≈ 0.89). The confusion matrix at the chosen

threshold (0.667) summarizes the final PD vs HC decisions (21 true negatives, 26 true positives, 4 false positives, and 3 false negatives). The histogram of calibrated probabilities for PD and HC demonstrates good separation between groups after isotonic regression. Consistent with these plots, the Brier score improved from 0.19 before calibration to 0.11 after calibration, indicating better probabilistic reliability within this dataset.

3.5. Quantitative Validation of Model Explainability

The quantitative assessment of XAI Fidelity and Reliability (Fig. 6) confirmed that the model's diagnostic logic aligns with established neuroanatomical markers of PD.

- Fidelity Analysis: Statistical comparison of ROI coverage revealed that the model significantly prioritized the Left Caudate ($P < .001$), Left Pallidum ($P < .001$), and Left Thalamus ($P = .008$) in PD patients compared to healthy controls (Fig. 6A). These findings are directionally consistent with the asymmetric involvement of basal ganglia structures in early PD. We found less consistent differences in right-hemisphere ROIs. Nonetheless, these P-values are exploratory (no multiple comparison correction, limited sample) and saliency coverage is an indirect proxy for "pathophysiological focus" rather than a direct measure of causality.
- Reliability Analysis: The Reliability Index (μ/σ) exceeded 2.0 for Thalamus and Pallidum in PD subjects, suggesting relatively stable attributions across individuals for these ROIs. Other regions showed lower RI values, indicating more heterogeneous saliency patterns. Given the lack of explicit robustness testing (e.g., under input noise or adversarial perturbations), these RI values should be interpreted cautiously.

Overall, the model's attributions qualitatively and quantitatively aligned with several known PD-related structures; however, we did not compare these patterns against randomized saliency baselines or lesion-grounded "gold standard" maps, which limits the strength of claims about true pathophysiological fidelity.

3.6. Patient-Level Explainability and Case Analysis

To contextualize model behavior, we examined four illustrative cases using a composite XAI dashboard (Fig. 7). For privacy, we refer to anonymized cases rather than specific identifiers:

- **Case TP (True Positive)**: A 38-year-old male with clinically confirmed PD was correctly classified with high PD probability. Despite his younger age relative to the cohort mean, Grad-CAM and Integrated Gradients highlighted midbrain and basal ganglia regions, and SHAP attributed a notable contribution to reduced Putamen CBF and decreased NM-MRI signal metrics, consistent with early nigrostriatal involvement.
- **Case TN (True Negative)**: A 61-year-old female HC was correctly classified as non-PD with low predicted probability. Saliency maps were more diffuse and less concentrated in PD-relevant ROIs;

tabular SHAP values indicated relatively preserved CBF and neuromelanin metrics, suggesting that the model did not over-interpret age-related structural changes as PD.

- **Case FP (False Positive)**: An older HC subject was misclassified as PD with moderate probability. Attribution analysis suggested that high age and subtle global structural and perfusion changes were interpreted by the model as disease-related, indicating residual age confounding and the model's sensitivity to age-related brain changes.
- **Case FN (False Negative)**: A middle-aged PD patient with relatively subtle imaging abnormalities and near-normal ASL CBF metrics was misclassified as HC. Here, younger age and less pronounced neuromelanin signal loss likely reduced the model's "suspicion," highlighting the difficulty of detecting very early or mild PD and the model's reliance on relatively prominent imaging deviations.

These qualitative analyses illustrate that, within this dataset, the model often—but not always—focuses on clinically plausible regions and features. They also highlight failure modes related to age, mild disease, and potential measurement limitations at 1.5T.

3.7. Global Feature Importance (Tabular SHAP Analysis)

Global SHAP analysis (Fig. 8) indicated that:

- Putamen CBF was the most influential feature in the tabular/fusion decision space, with the largest mean absolute SHAP value.
- Secondary contributors included Caudate CBF, NM-MRI SNR and CNR in the substantia nigra, and Thalamus CBF.
- SHAP beeswarm plots showed that lower Putamen CBF values tended to push predictions toward the PD class, consistent with prior reports of striatal hypoperfusion in PD.
- Similarly, lower neuromelanin signal metrics were associated with higher PD probability.

However, SHAP attributions are model-dependent rather than mechanistic; they are sensitive to feature correlations and background sample choice and should not be interpreted as causal effect sizes. Nonetheless, the prominence of Putamen CBF and nigral neuromelanin metrics in the attribution hierarchy supports the proposed "neurovascular–structural decoupling" hypothesis in PD, at least within this small sample.

4. DISCUSSION

This pilot feasibility study evaluated a knowledge-informed, attention-enhanced hybrid deep learning framework that integrates T1-weighted structural MRI, neuromelanin-sensitive MRI (NM-MRI), and ASL perfusion with imaging-derived quantitative metrics and age for PD versus healthy control classification. In a single-center cohort of 54 participants, the multimodal model achieved higher internal cross-validated performance (AUC 0.92) than any unimodal branch (T1: 0.81; NM-MRI: 0.77; ASL-CBF: 0.79), and explainability analyses showed that predictions were predominantly driven by putaminal and

caudate CBF and neuromelanin metrics in the substantia nigra, with saliency concentrated in basal ganglia and thalamic regions, particularly on the left. These findings indicate that, even in a limited dataset, attention-based fusion of complementary MRI contrasts and quantitative biomarkers can enhance PD-related signal extraction beyond what single sequences typically provide.

Our unimodal T1-based performance is broadly consistent with prior structural MRI studies, which have demonstrated modest but reproducible atrophy patterns in basal ganglia and fronto-limbic regions and classifier AUCs on the order of 0.8 when using sMRI features alone (20–23). Likewise, NM-MRI and ASL-only approaches have shown that loss of neuromelanin signal in SNpc/LC and regional CBF changes in striatal and cortical networks can distinguish PD from controls and correlate with motor and non-motor features, but reported diagnostic performance as single modalities is typically in the low-to-mid 0.8 range (9, 21, 24, 25). Within this context, the internal AUC of 0.91 obtained by our hybrid model, under subject-level cross-validation in the same cohort, is in line with recent multimodal efforts that have reported gains by combining structural MRI with diffusion, functional, or genetic data (26, 27) (Fereshtehnejad et al., 2017; Chen et al., 2021; Cao & Long, 2025). Given the small sample size and lack of external validation, this AUC should be interpreted as a promising internal benchmark rather than a definitive comparison, but it supports the central premise that integrating structural, neuromelanin, and perfusion information is advantageous.

A key feature of our framework is its explicitly knowledge-informed design. Instead of relying on unconstrained end-to-end learning, we first computed ROI-based neuromelanin SNR/CNR and regional ASL-derived CBF values that mirror standard radiological measurements and then fused these expert-defined biomarkers with CNN-derived image embeddings via attention. In this way, the network learns how to weight and combine established features with high-dimensional spatial patterns, rather than attempting to replace expert knowledge. The prominence of putaminal and caudate CBF and nigral neuromelanin metrics in SHAP rankings is therefore an intended and desirable outcome, reflecting a “neurovascular–structural” synergy that is coherent with both NM-MRI and ASL literature. The attention mechanism allows the model to adaptively emphasize perfusion markers when neuromelanin signal is subtle or noisy (e.g., at 1.5T), and vice versa, which is particularly useful in small and heterogeneous clinical datasets.

The region-focused 2.5D input further embeds anatomical priors. By restricting the CNN input to a five-slice axial slab centered on the midbrain and basal ganglia, we target the substantia nigra pars compacta, striatum, pallidum, and thalamus—key nodes of the nigrostriatal and basal ganglia–thalamocortical circuits implicated in PD—while excluding much of the cortex and ventricular system. This design reduces the likelihood that the network bases its decisions on diffuse cortical atrophy or ventricular enlargement, which are strongly age-related and non-specific, and acts as a form of anatomically grounded regularization in a small-sample setting. Saliency-based XAI confirmed that, within this slab, the model consistently focused on PD-relevant subcortical ROIs, and the reliability index indicated stable attributions for the thalamus and pallidum across PD subjects. In larger, multi-center cohorts, this region-focused strategy can be relaxed toward whole-brain 2.5D/3D architectures with

subject-level attention pooling, following directions suggested by recent multimodal and multi-omics PD work (27)(e.g., Cao & Long, 2025), while retaining the interpretability benefits of knowledge-informed fusion.

Age and medication status are important contextual factors. The PD group was significantly older than the HC group, and age is a strong determinant of brain structure, perfusion, and neuromelanin content. We therefore treated age as a primary confounder by including it as a covariate in statistical analyses and as an explicit feature in the tabular branch, and by focusing the CNN input on deep structures rather than the full cortex. These choices likely mitigated, but did not completely eliminate, age-related confounding; some contribution of age-dependent variation to the decision boundary cannot be excluded. All PD patients were receiving dopaminergic therapy and were scanned in their usual medicated (ON) state. Prior ASL work has reported that antiparkinsonian medications may modulate cerebral blood flow (28), implying that our CBF measures reflect a combination of disease- and treatment-related effects. We did not model medication type, dose, or duration, so the present results should be interpreted as characterizing the imaging phenotype of treated PD under typical clinical conditions rather than drug-naïve disease. At the same time, a uniform ON-state reduces variability from ON/OFF fluctuations and aligns the imaging protocol with real-world practice.

As a pilot feasibility study, this work is naturally constrained by a modest sample size from a single center, with all imaging acquired on one 1.5T scanner using a fixed NM-MRI and ASL protocol. These characteristics limit formal generalizability and mean that the reported AUC represents a robust internal estimate rather than a final clinical performance benchmark. We attempted to maximize methodological rigor within these constraints by using subject-level stratified five-fold cross-validation, fully per-subject preprocessing (avoiding cohort-level statistics), conservative augmentation, and a compact backbone to reduce overfitting and data leakage. Nested cross-validation for hyperparameter tuning and formal robustness testing of the XAI components were not implemented and remain important directions for future work. Likewise, although our XAI analyses aligned well with known PD anatomy, more extensive sanity checks and perturbation-based evaluations in larger, independent cohorts will be needed to fully establish the stability and clinical reliability of these explanations.

Despite these constraints, the study provides a concrete proof-of-concept that an attention-enhanced, knowledge-informed multimodal framework can, even in a pilot cohort, (i) outperform unimodal MRI models, (ii) focus on anatomically and pathophysiologically plausible subcortical structures, and (iii) offer interpretable feature- and region-level attributions. The proposed pipeline—region-focused 2.5D inputs, expert-derived MRI biomarkers, attention-based late fusion, calibrated outputs, and quantitative XAI—offers a practical template that can be scaled and stress-tested in larger, harmonized, multi-center datasets. Future studies should validate this approach externally, incorporate richer clinical and treatment variables, extend to whole-brain and prodromal-stage applications (e.g., iRBD), and apply more rigorous XAI evaluation, with the longer-term goal of developing robust and interpretable AI tools to support imaging-based assessment of Parkinson’s disease.

5. CONCLUSION

This pilot study demonstrates the feasibility and methodological value of a knowledge-informed, attention-enhanced multimodal deep learning framework that fuses structural MRI, neuromelanin-sensitive MRI, and ASL-derived perfusion with quantitative imaging biomarkers for PD classification. Within a small single-center cohort, the hybrid model outperformed unimodal MRI approaches and generated explainable attributions that aligned with known nigrostriatal and basal ganglia pathology. While the current results are internal and exploratory, they indicate that intelligently combining perfusion, neuromelanin, and structural information in an interpretable architecture is a promising direction for future, larger-scale studies aimed at developing clinically useful AI-assisted tools for Parkinson's disease.

Declarations

Acknowledgements

The authors would like to acknowledge the Advanced Medical Center of Avan for their valuable cooperation in acquiring imaging data. We also gratefully acknowledge the Persian Cohort Center at Imam Reza Hospital for providing access to the participant data used in this study.

Funding Sources

This research was supported by the Vice-Chancellor for Research Department at Mashhad University of Medical Sciences (Grant No. 4030777).

Disclosure Statement

The authors report no conflict of interest.

Ethical Statement

This research was conducted in accordance with the principles embodied in the Declaration of Helsinki and in accordance with local statutory requirements. The study was approved by the Research Ethics Committee of the Mashhad University of Medical Sciences (approval number IR.MUMS.MEDICAL.REC.1403.319). All participants provided written informed consent to participate in the study.

Declaration Of Generative AI And AI-Assisted Technologies in the Manuscript Preparation Process

During the preparation of this work the authors used ChatGPT (OpenAI) in order to proofread and improve the clarity of English-language sentences. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

Data Availability Statement

Individual-level data used in this study originate from the Persian Organizational Cohort Study (POCS) at Mashhad University of Medical Sciences (MUMS). Because these data are governed by institutional and ethical restrictions to protect participant privacy, they are not publicly available. De-identified derived data (e.g., ROI-wise imaging features) may be made available upon reasonable request, subject to approval by the relevant MUMS/POCS oversight committees and execution of an appropriate data-use agreement. Requests should be directed to the Persian Cohort/MUMS data access office.

All code, including preprocessing and training/inference scripts, model checkpoints, and figure-generation notebooks, is publicly available in the authors' GitHub repository: https://github.com/zohrehganji/MultiMRI-PD. The repository includes environment files and step-by-step instructions to facilitate reproducibility of the analyses reported in this manuscript.

Author Contributions

Z.G. and F.N.: Data curation, methodology, and writing of the original draft. N.Sh. and A.Sh.: Clinical Data Validation and Resources. E.Sh.: Data curation. SM.H.: Methodology, H.Z.: Supervision, funding acquisition, writing, reviewing, and editing.

References

1. Berg D, Borghammer P, Fereshtehnejad S-M, Heinzel S, Horsager J, Schaeffer E, et al. Prodromal Parkinson disease subtypes—key to understanding heterogeneity. Nat Reviews Neurol. 2021;17(6):349–61.
2. Postuma RB, Berg D. Prodromal Parkinson's disease: the decade past, the decade to come. Mov Disord. 2019;34(5):665–75.
3. Smithies S, Ebneabbasi A, Conci R, Murley A, Savulich G, Stockton K et al. Cognitive profile accurately predicts underlying pathology in Corticobasal Syndrome. medRxiv. 2025:2025.12. 17.25342469.
4. Kathpalia KV, Duodu MG, Raj R, Albkerat AA. Recent progress in biomarkers for neurodegenerative disorders. Neurodegeneration Revolution. 2025:423–36.
5. Breen DP, Lang AE. Tracking the course of prodromal Parkinson's disease. Brain. 2017;140(2):259–62.
6. Eisenstein T, Groenewald K, van Hillegondsberg L, Al Hajraf F, Zerenner T, Lawton MA, et al. Cholinergic degeneration in prodromal and early Parkinson's disease: a link to present and future disease states. Brain. 2026;149(1):226–37.
7. Su D, Cui Y, He C, Yin P, Bai R, Zhu J et al. Projections for prevalence of Parkinson's disease and its driving factors in 195 countries and territories to 2050: modelling study of Global Burden of Disease Study 2021. BMJ. 2025;388.
8. Theis H, Pavese N, Rektorová I, Van Eimeren T. Imaging biomarkers in prodromal and earliest phases of Parkinson's disease. J Parkinson's Disease. 2024;14(s2):S353–65.

9. Trujillo P, Aumann MA, Claassen DO. Neuromelanin-sensitive MRI as a promising biomarker of catecholamine function. Brain. 2024;147(2):337–51.
10. Liu P, Wang X, Zhang Y, Huang P, Jin Z, Cheng Z, et al. PENCIL imaging: A novel approach for neuromelanin sensitive MRI in Parkinson's disease. NeuroImage. 2024;291:120588.
11. Joshi D, Prasad S, Saini J, Ingalhalikar M. Role of arterial spin labeling (ASL) images in Parkinson's disease (PD): a systematic review. Acad Radiol. 2023;30(8):1695–708.
12. Pan Y, Qin Z, Liu J, Wang W, Zhao Y. Hemodynamic subtype profiling in tremor-dominant and akinetic–rigid Parkinson's disease using multi-delay pseudo-continuous arterial spin labeling imaging. Front Hum Neurosci. 2025;19:1617996.
13. Dentamaro V, Impedovo D, Musti L, Pirlo G, Taurisano P. Enhancing early Parkinson's disease detection through multimodal deep learning and explainable AI: insights from the PPMI database. Sci Rep. 2024;14(1):20941.
14. Camacho Camacho MI. Explainable prediction of Parkinson's disease in a large multimodal database. 2023.
15. Ruga T, Caroprese L, Vocaturo E, Zumpano E. Explainable Artificial Intelligence for Parkinson's Disease. Explainable Machine Intelligence in Healthcare. 2025:135 – 49.
16. Pires LMdO. Assessing brain functional connectivity in Parkinson's disease using explainable Artificial Intelligence methods 2023.
17. Tustison NJ, Avants BB, Cook PA, Zheng Y, Egan A, Yushkevich PA, et al. N4ITK: improved N3 bias correction. IEEE Trans Med Imaging. 2010;29(6):1310–20.
18. Isensee F, Schell M, Pflueger I, Brugnara G, Bonekamp D, Neuberger U, et al. Automated brain extraction of multisequence MRI using artificial neural networks. Hum Brain Mapp. 2019;40(17):4952–64.
19. Alsop DC, Detre JA, Golay X, Günther M, Hendrikse J, Hernandez-Garcia L, et al. Recommended implementation of arterial spin-labeled perfusion MRI for clinical applications: a consensus of the ISMRM perfusion study group and the European consortium for ASL in dementia. Magn Reson Med. 2015;73(1):102–16.
20. Zeighami Y, Ulla M, Iturria-Medina Y, Dadar M, Zhang Y, Larcher KM-H, et al. Network structure of brain atrophy in de novo Parkinson's disease. elife. 2015;4:e08440.
21. Shinde S, Prasad S, Saboo Y, Kaushick R, Saini J, Pal PK, et al. Predictive markers for Parkinson's disease using deep neural nets on neuromelanin sensitive MRI. NeuroImage: Clin. 2019;22:101748.
22. Laansma MA, Zhao Y, van Heese EM, Bright JK, Owens-Walton C, Al-Bachari S, et al. A worldwide study of subcortical shape as a marker for clinical staging in Parkinson's disease. npj Parkinson's Disease. 2024;10(1):223.
23. Dadar M, Miyasaki J, Duchesne S, Camicioli R. White matter hyperintensities mediate the impact of amyloid ß on future freezing of gait in Parkinson's disease. Parkinsonism Relat Disord. 2021;85:95–101.

24. Matsuura K, Ii Y, Maeda M, Tabei K-i, Satoh M, Umino M, et al. Neuromelanin-sensitive magnetic resonance imaging in disease differentiation for parkinsonism or neurodegenerative disease affecting the basal ganglia. Parkinsonism Relat Disord. 2021;87:75–81.
25. Arslan DB, Gurvit H, Genc O, Kicik A, Eryurek K, Cengiz S, et al. The cerebral blood flow deficits in Parkinson's disease with mild cognitive impairment using arterial spin labeling MRI. J Neural Transm. 2020;127(9):1285–94.
26. Fereshtehnejad S-M, Zeighami Y, Dagher A, Postuma RB. Clinical criteria for subtyping Parkinson's disease: biomarkers and longitudinal progression. Brain. 2017;140(7):1959–76.
27. Cao J, Long X. RACF: A Multimodal Deep Learning Framework for Parkinson's Disease Diagnosis Using SNP and MRI Data. Appl Sci. 2025;15(8):4513.
28. Zhang J, Fan Y, Liang H, Zhang Y. Global, regional and national temporal trends in Parkinson's disease incidence, disability-adjusted life year rates in middle-aged and older adults: a cross-national inequality analysis and Bayesian age-period-cohort analysis based on the global burden of disease 2021. Neurol Sci. 2025;46(4):1647–60.

Figures

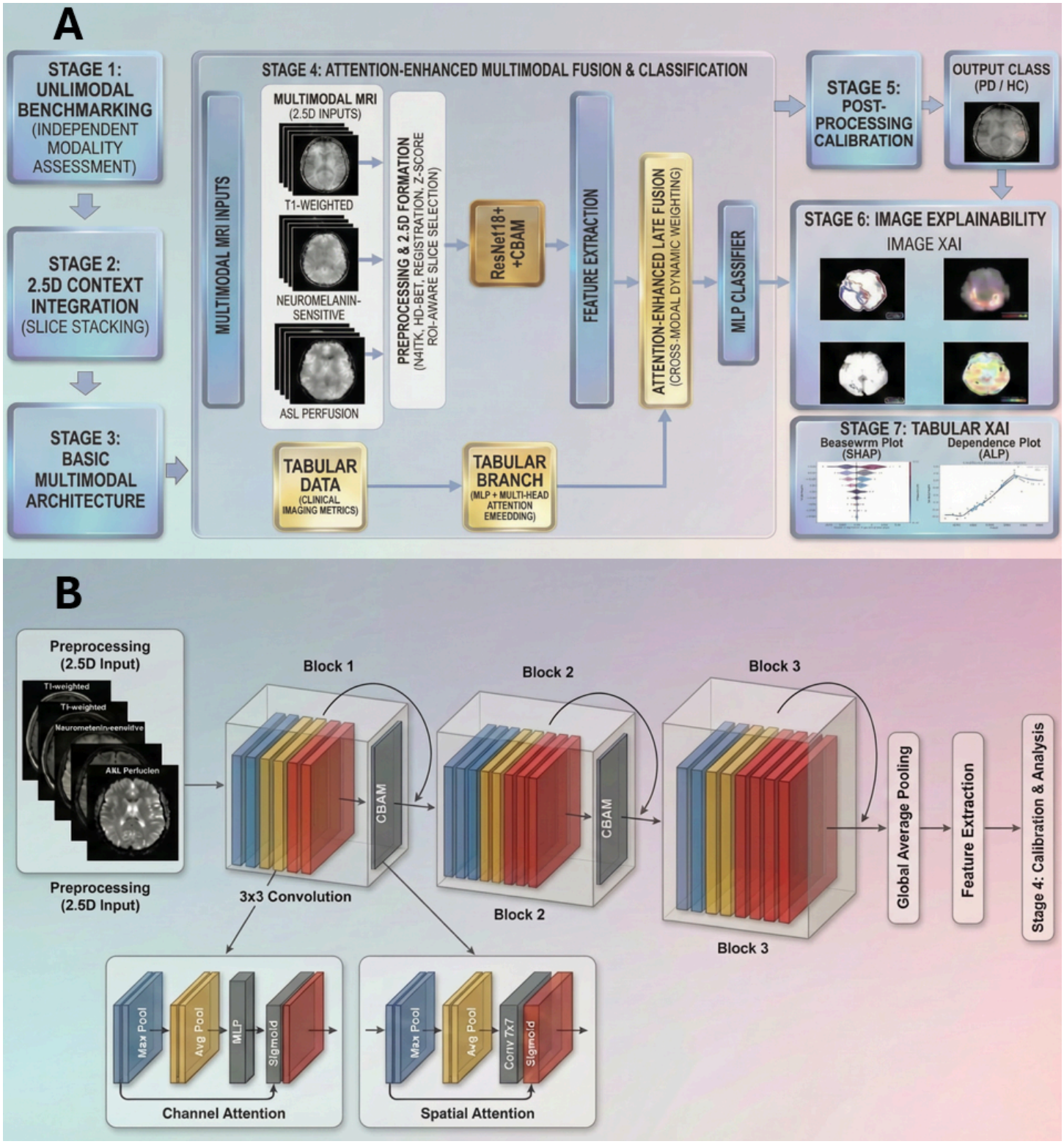


Figure 1

Overview of the seven-stage multimodal hybrid deep learning pipeline, from MRI acquisition and preprocessing to multimodal fusion, calibration, and explainable AI.

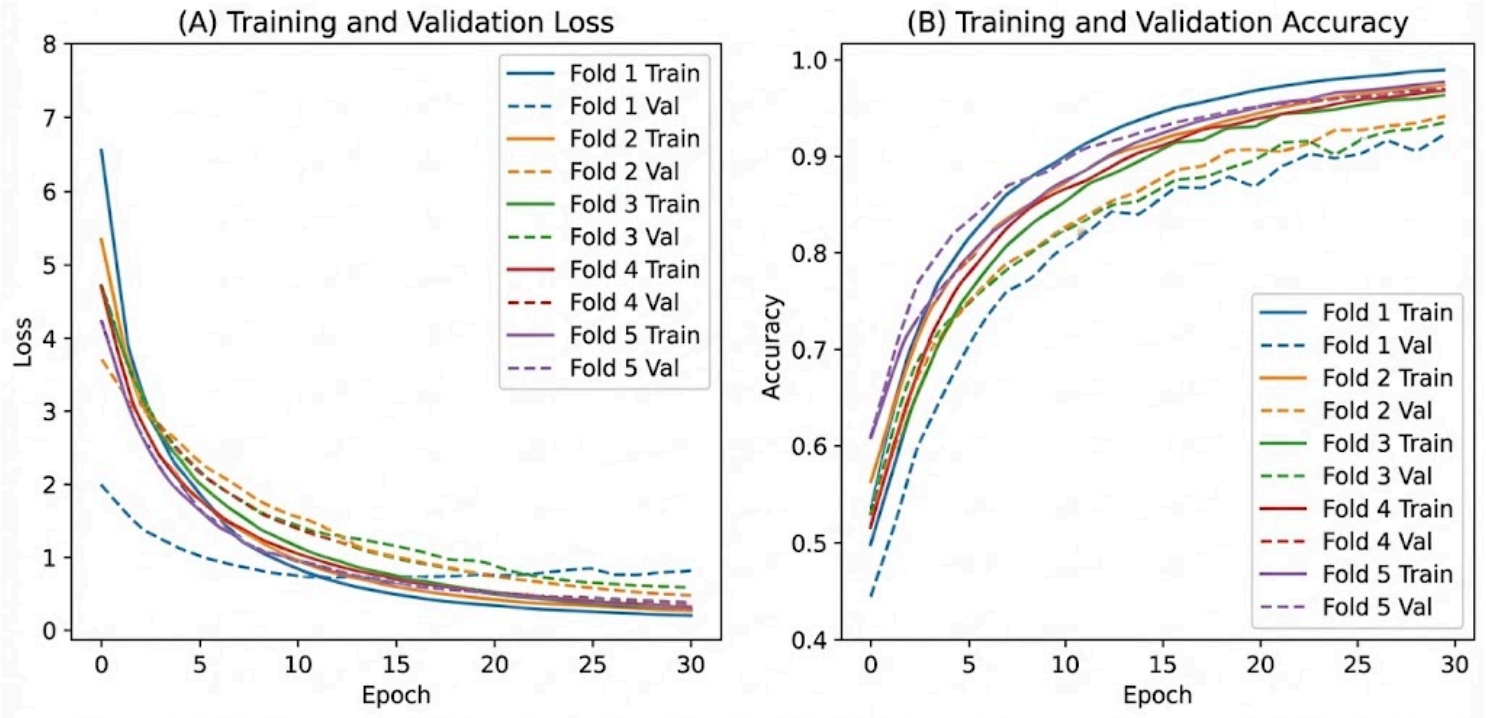


Figure 2

Training vs. validation loss (A) and accuracy (B) across epochs, demonstrating stable convergence with limited overfitting.

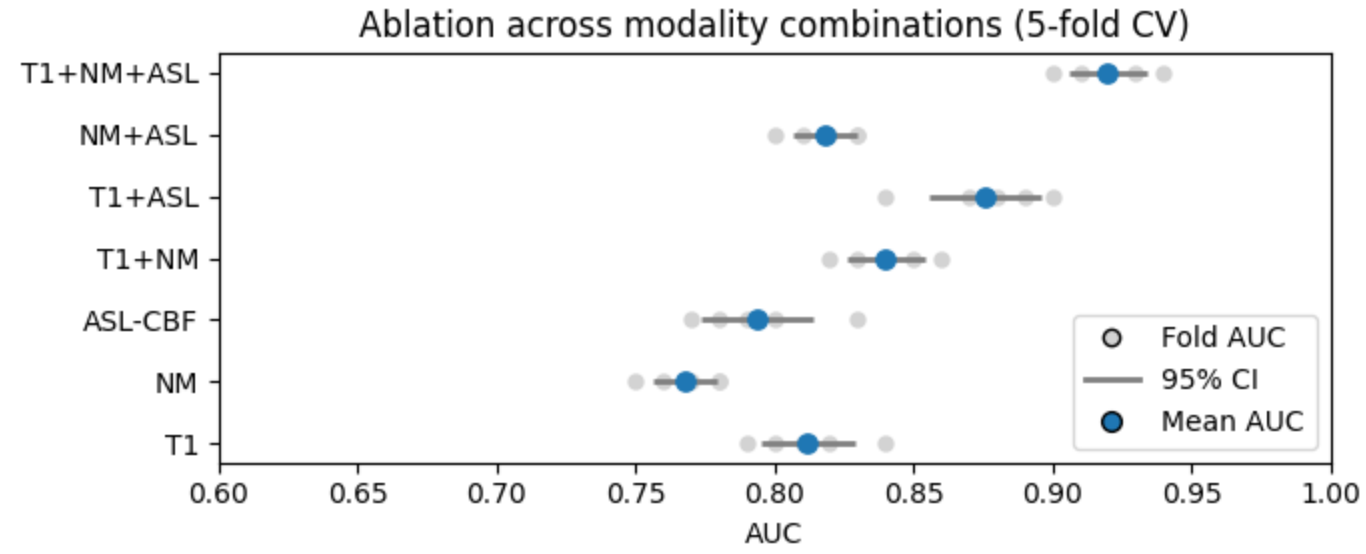


Figure 3

Ablation across modality combinations (5-fold CV). Mean AUC for each model (blue dots) with 95% confidence intervals (gray horizontal lines), and individual fold-wise AUC values (light gray circles) for unimodal models, pairwise combinations, and the full T1+NM+ASL hybrid.

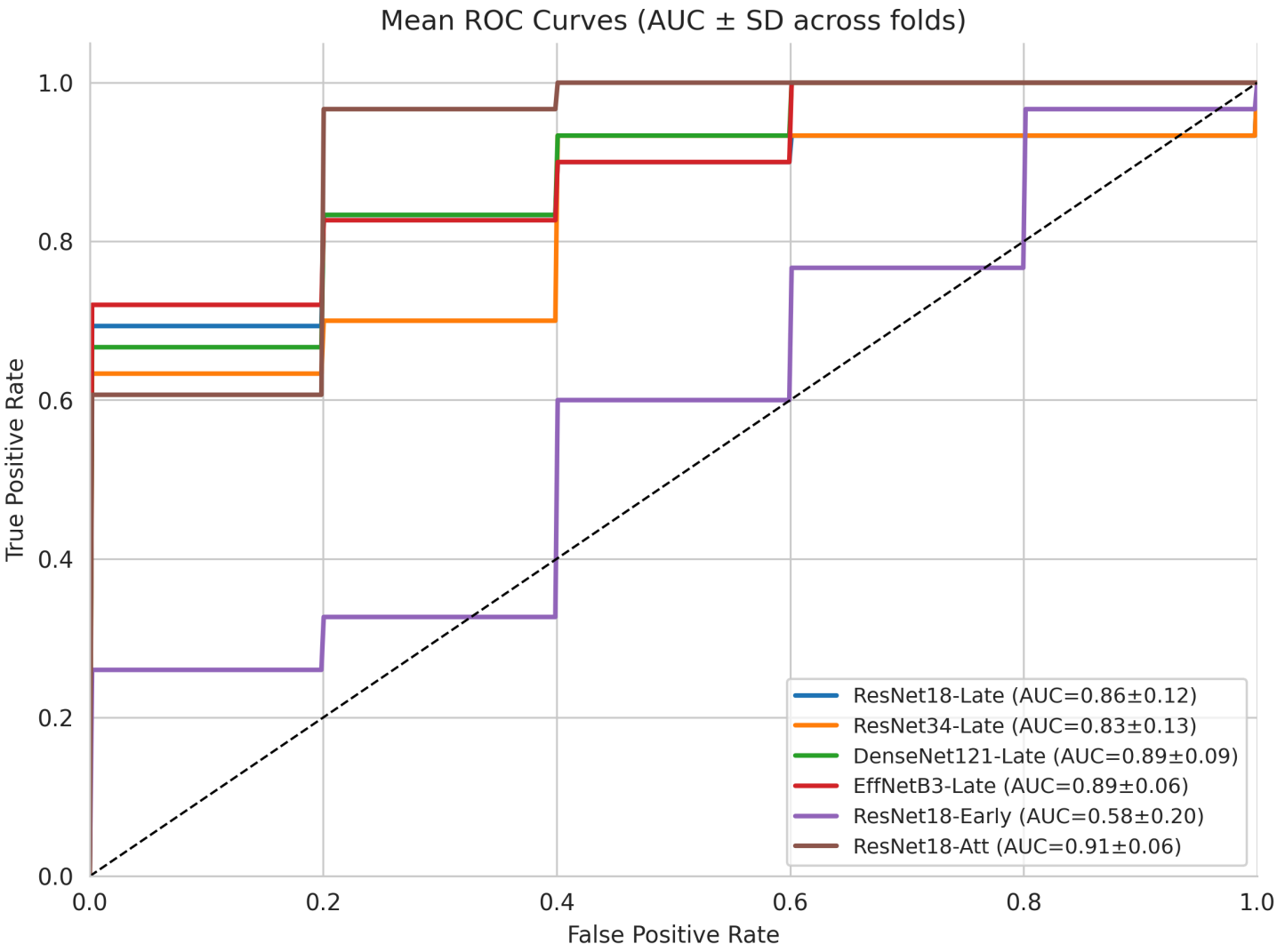


Figure 4

Comparison of fusion strategies and CNN backbones: mean AUC (±SD or 95% CI) for early fusion, standard late fusion, and attention-enhanced late fusion across different architectures, highlighting the superior and more stable performance of ResNet18-Att.

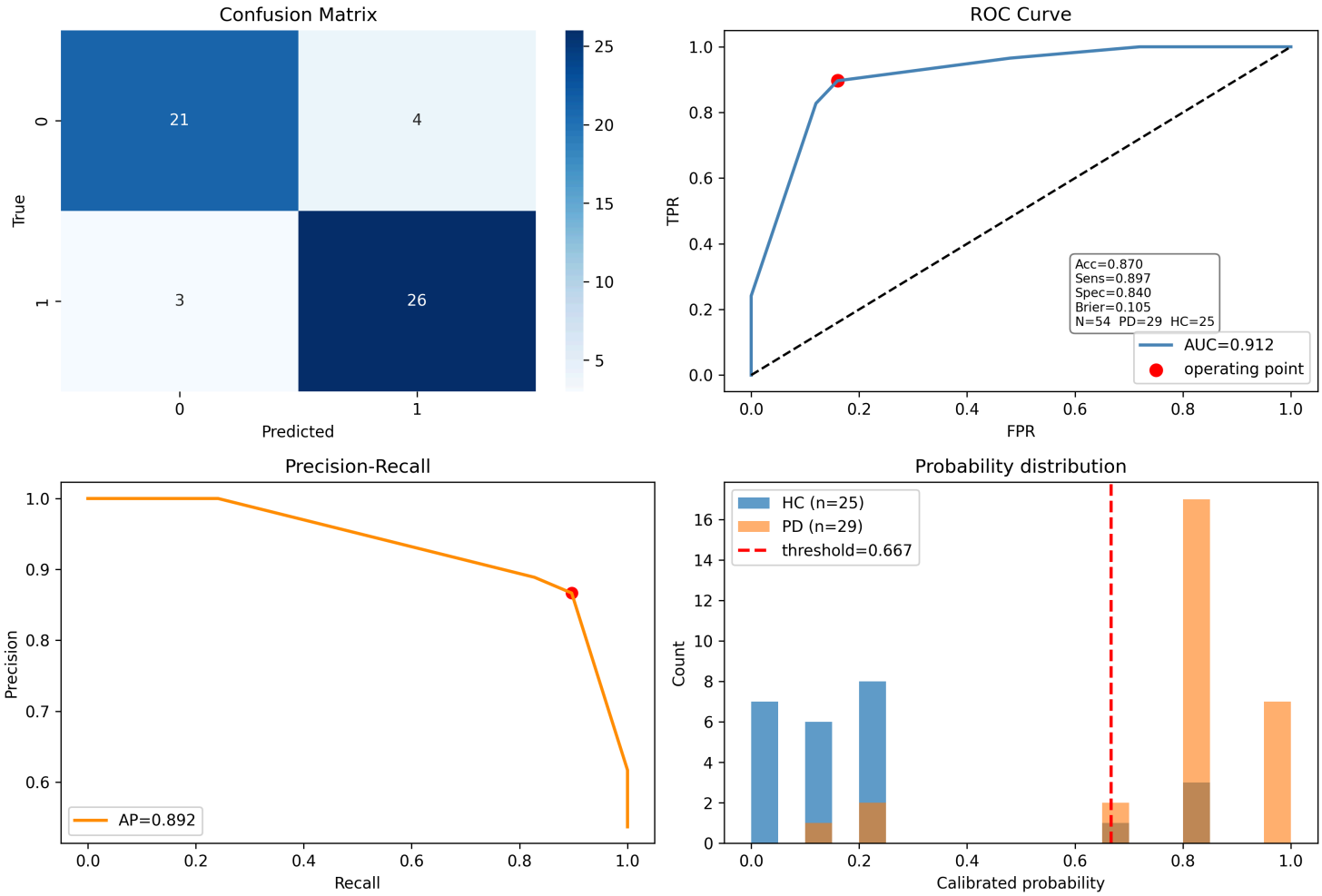


Figure 5

Performance summary of the calibrated ResNet18-Att model: confusion matrix, ROC curve (AUC ≈0.91), precision–recall curve, and distribution of calibrated PD vs HC probabilities at the Youden-optimal threshold (0.667).

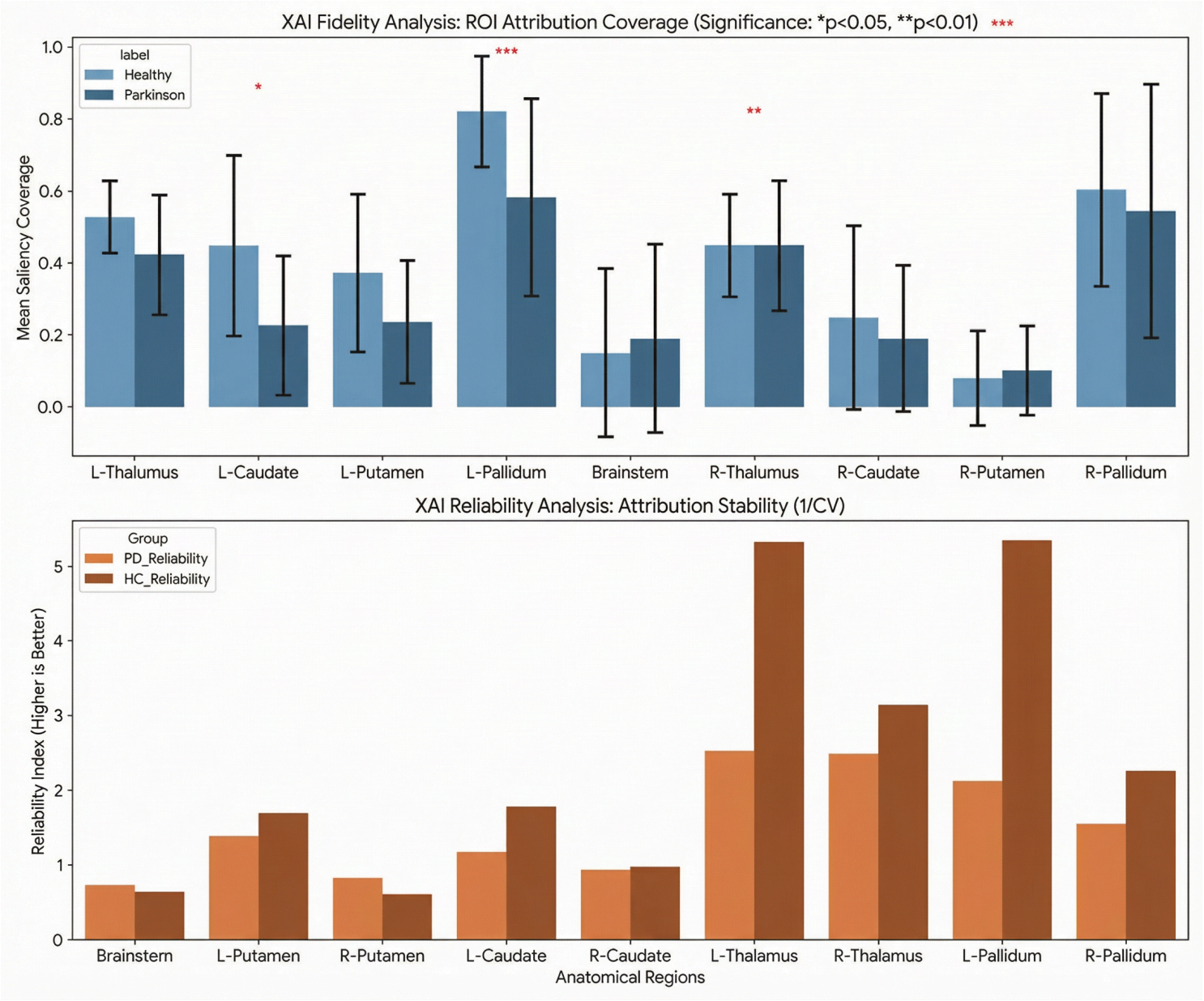


Figure 6

Quantitative XAI metrics: (A) mean saliency coverage (fidelity) in PD-relevant ROIs for PD vs HC; (B) reliability index (μ/σ) per ROI in PD, showing stable attributions for thalamus and pallidum.

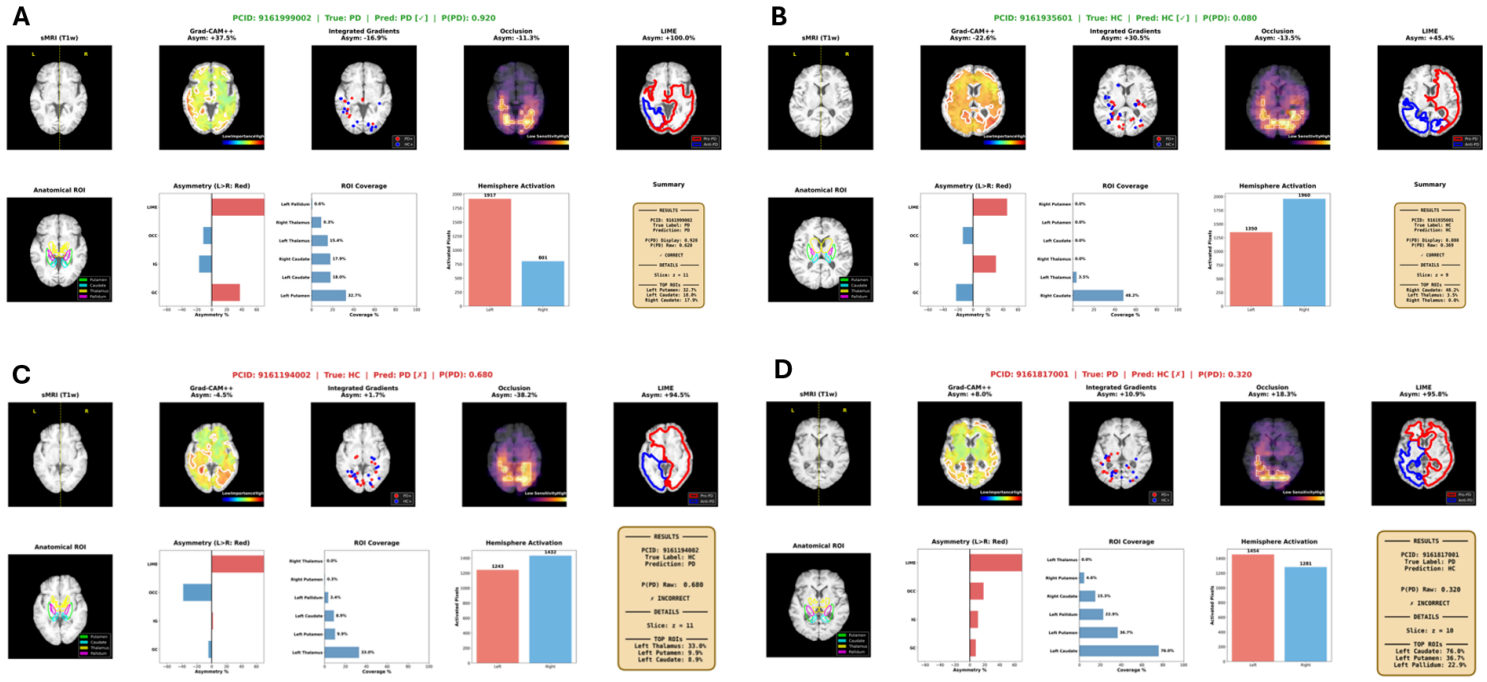


Figure 7

Patient-level XAI dashboard for representative (A) true positive, (B) true negative, (C) false positive, and (D) false negative cases, combining Grad-CAM, Integrated Gradients, and SHAP feature attributions.

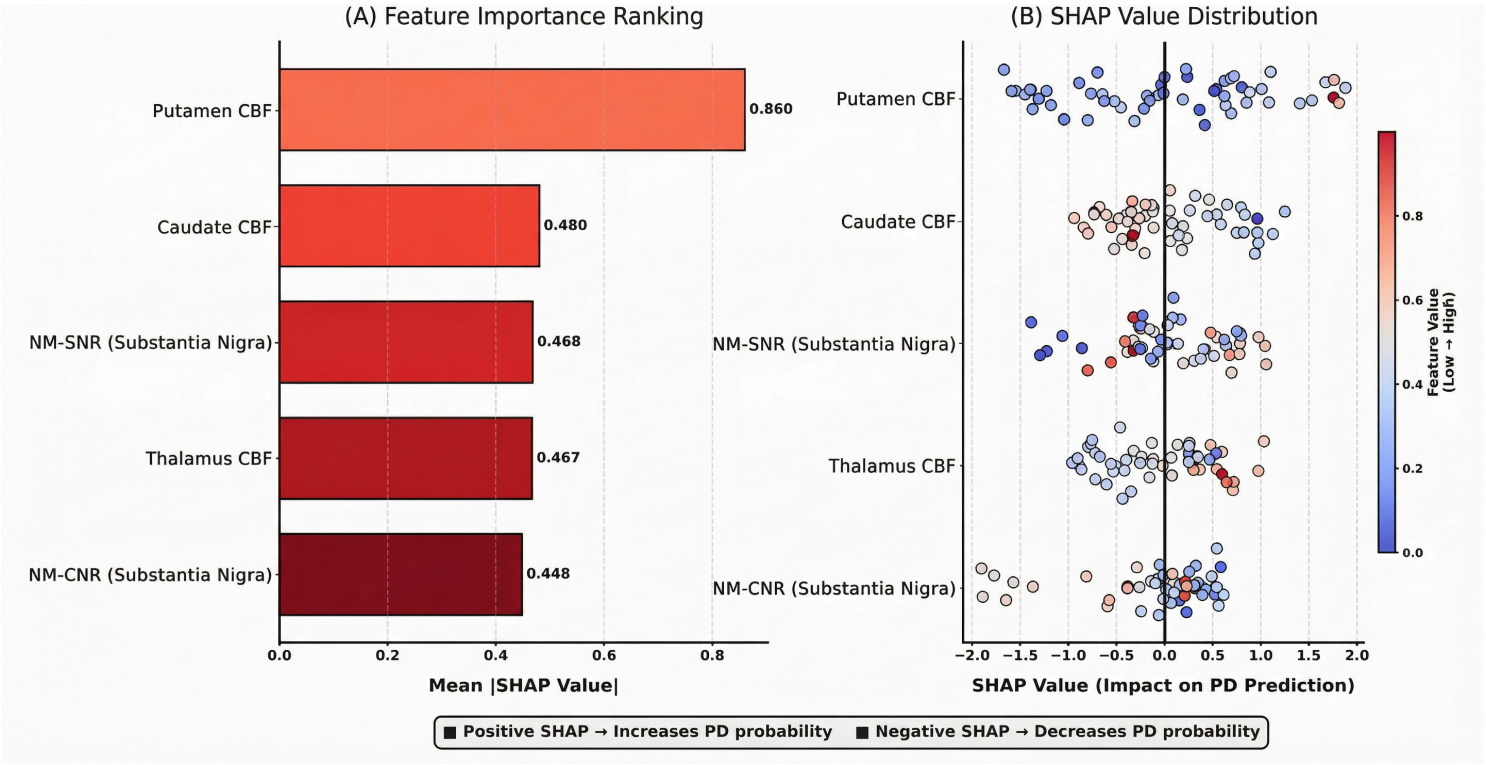


Figure 8

Global SHAP analysis of tabular features: (A) mean absolute SHAP values highlighting putaminal and caudate CBF and nigral NM metrics as dominant contributors; (B) beeswarm plot showing that lower putaminal CBF and neuromelanin signal shift predictions toward the PD class.