Biomedical Signal Processing and Control 127 (2026) 111143

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc

β-amyloid–stratified six-stage Alzheimer's disease discrimination via multiband MRI histogram and GLCM features

José Menezes, Maria Inês Barbosa, Pedro Miguel Rodrigues *

Universidade Católica Portuguesa, CBQF – Centro de Biotecnologia e Química Fina – Laboratório Associado, Escola Superior de Biotecnologia, Rua de Diogo Botelho 1327, Porto, 4169-005, Portugal

ARTICLE INFO

Dataset link: http://adni.loni.usc.edu, http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

Keywords:
Alzheimer's disease
Beta-amyloid
Mild cognitive impairment
sMRI
Machine learning
Discrimination

ABSTRACT

Accurate and early characterization of Alzheimer's disease (AD) progression remains a major clinical challenge, particularly in biologically heterogeneous stages such as mild cognitive impairment (MCI). We propose an sMRI-based machine learning framework for multi-stage AD discrimination, encompassing six clinical groups (cognitively normal (CN, $n = 103$), subjective memory complaints (SMC, $n = 54$), early MCI (EMCI, $n = 151$), MCI ($n = 229$), late MCI (LMCI, $n = 117$), and AD ($n = 114$)), with explicit β-amyloid (Aβ) stratification of MCI subtypes (+/−). T1-weighted sMRI scans from 768 ADNI subjects were analyzed using histogram and GLCM texture features across multiple anatomical planes and multi-scale wavelet decompositions. The framework demonstrated robust performance across 36 pairwise comparisons under stratified 5-fold cross-validation with SMOTE and RUS, achieving $AUC > 0.95$ in well-separated tasks (e.g., AD vs EMCI$^-$) and in SMC vs AD and SMC vs MCI$^+$. More challenging "gray-zone" distinctions (e.g., EMCI$^-$ vs EMCI$^+$) showed moderate performance ($AUC \approx 0.62$–0.67), with RUS providing the most overall balanced results. Feature analysis revealed dominant contributions from GLCM descriptors, complemented by histogram features, with highest discriminability in Level-2 wavelet sub-bands (HL2 and LH2). Aβ stratification enabled a more biologically grounded interpretation of disease progression, improving discrimination in later stages while remaining challenging in early-stage comparisons. Overall, sMRI radiomics provides an interpretable framework for capturing stage-specific structural patterns in AD, highlighting the importance of biological stratification and robust validation.

1. Introduction

Alzheimer's disease (AD) is a neurodegenerative condition that progressively impairs cognitive abilities, including memory, language and perception, and ultimately leads to death. It is the leading cause of dementia in older adults, accounting for 60%–70% of cases, and is regarded as one of the most costly, lethal and challenging diseases of our time [1–3].

AD is characterized by the accumulation of extracellular β-amyloid (Aβ) plaques and intracellular neurofibrillary tangles composed of hyperphosphorylated tau protein [4]. In addition, it is associated with progressive neuronal loss across the hippocampus and cerebral cortex, while other brain regions remain largely spared until the later stages of the disease [5,6]. As AD progresses, its clinical presentation follows a sequence of stages that can vary in definition and number across studies. Usually, these stages include cognitively normal (CN), subjective memory complaints (SMC), early mild cognitive impairment (EMCI), mild cognitive impairment (MCI), late mild cognitive impairment (LMCI) and AD itself [7–9].

MCI is a transitional stage between normal cognitive aging and dementia, characterized by measurable cognitive decline that exceeds age-related expectations but does not yet significantly interfere with daily functioning. Although MCI is not invariably progressive nor always indicative of early AD, it is frequently regarded as the earliest clinical manifestation of AD pathophysiology, with MCI due to AD potentially progressing to dementia in approximately 40%–75% of cases [10–13]. Individuals with EMCI typically experience mild memory lapses while maintaining independence and those with MCI exhibit increasing forgetfulness, disorientation and communication difficulties over several years. In LMCI, patients begin requiring assistance with daily tasks as cognitive, motor and behavioral impairments intensify. During the dementia stage, profound cognitive decline leads to

* Correspondence to: Universidade Católica Portuguesa, CBQF - Centro de Biotecnologia e Química Fina, Escola Superior de Biotecnologia, Office EBI112, Rua Diogo Botelho 1327, 4169-005 Porto, Portugal.
E-mail address: pmrodrigues@ucp.pt (P.M. Rodrigues).

https://doi.org/10.1016/j.bspc.2026.111143
Received 27 January 2026; Received in revised form 29 June 2026; Accepted 22 July 2026
Available online 28 July 2026

loss of independence, disorientation, impaired recognition of relatives, and pronounced behavioral disturbances [7]. These symptoms tend to appear inconsistently and vary in intensity over time [3,14].

The diagnosis of AD is established through a comprehensive clinical history, psychiatric and neurological examination, neuropsychological assessment and supportive measures, such as structural imaging and blood tests, which serve to exclude other origins of dementia [5]. In both clinical practice and research, several standardized tools are used to measure cognitive decline and overall impairment: the Alzheimer's Disease Assessment Scale–Cognitive Subscale (ADAS-Cog), the Mini-Mental State Examination (MMSE), and the Clinical Dementia Rating scale (CDR-SOB) . However, using different assessment tools can make it difficult to compare findings across patients or studies directly [15]. For example, the MMSE is the most used cognitive test in clinical practice, but it shows limited sensitivity for early detection, with values ranging from 27% to 89% and specificities from 32% to 90% for predicting progression from MCI to AD dementia [16]. In contrast, the CDR-SOB shows intermediate performance, demonstrating scores of 71% to 74% sensitivity and 81% specificity while distinguishing MCI from early AD [17]. Regarding ADAS-Cog, this test demonstrates superior diagnostic performance, with sensitivity of 86.9% and specificity of 83.5% for MCI detection, and 92.2% sensitivity and 90.7% specificity for AD detection [18]. It also has the highest Youden's Index (0.829), a parameter optimized to identify optimal cut-off points for diagnostic biomarkers such as ptau and $A\beta$, suggesting that the ADAS-Cog may offer the most balanced discrimination between positive and negative cases [18,19].

Advances in structural and molecular imaging have yielded critical insights into the temporal dynamics of the pathophysiological processes in AD [20,21]. More recently, progress in imaging biomarkers have enhanced their integration into the diagnostic process for AD, with structural magnetic resonance imaging (sMRI) representing a particularly valuable approach [22]. sMRI plays a crucial role in the diagnosis and monitoring of AD, providing detailed visualization of structural, microstructural and functional brain alterations associated with disease progression. Its noninvasive nature, lack of radiation exposure, shorter scanning time and relatively low cost make sMRI especially well-suited for longitudinal studies [23]. These advantages also position sMRI as an ideal technique for machine learning (ML)-based discrimination of AD stages and for enhancing diagnostic differentiation among early, mild, and late cognitive impairment subtypes [24–26].

Building upon recent advances in neuroimaging and ML for AD characterization, the present study introduces a comprehensive MRI-based framework aimed at improving the classification of AD across its clinical continuum. While prior studies have demonstrated the utility of sMRI combined with ML for distinguishing broad diagnostic categories (see Section 2), these approaches typically rely on coarse group definitions (e.g., CN, MCI, AD) and thus fail to capture the substantial biological and clinical heterogeneity within intermediate stages, particularly within MCI. In this context, we explicitly hypothesize that incorporating Aβ status as a biological stratifier, alongside refined clinical staging, enhances the discriminability between disease stages and provides a more biologically grounded characterization of AD progression.

To address this, the proposed framework integrates detailed clinical staging (CN, SMC, EMCI, MCI, LMCI, AD) with amyloid biomarker stratification ($A\beta^{+}/A\beta^{-}$) of the MCI stages, enabling the systematic and explicit evaluation of how the presence of amyloid pathology modulates classification performance across successive disease stages. By subdividing MCI into EMCI, MCI and LMCI, and further stratifying these stages according to Aβ status, this study extends beyond conventional MRI-based classification paradigms. Specifically, it allows for a direct assessment of whether amyloid positivity is associated with increased separability between adjacent clinical stages, thereby quantifying the extent to which molecular pathology is reflected in structural neuroimaging patterns. This approach not only facilitates the identification of biologically meaningful subgroup differences across the AD continuum, but also directly addresses a critical and underexplored question: to what extent does Aβ burden influence the discriminative capacity of MRI-based models as the disease progresses?

Thus, the main contributions and objectives of this study are as follows:

- Development of a refined diagnostic framework that extends beyond the standard CN–MCI–AD paradigm by subdividing MCI into EMCI, MCI and LMCI stages, and further stratifying these groups based on Aβ positivity or negativity. This design enables the systematic evaluation of the impact of amyloid status on disease progression and group separability.
- Extraction of sMRI-derived features from T1-weighted scans, including histogram-based metrics and gray-level co-occurrence matrix (GLCM) features, computed across multiple whole-brain frequency bands obtained via 2D discrete wavelet transform (DWT2), considering both slice-wise analyses and combined whole-brain representations.
- Design and implementation of a ML framework integrating feature selection strategies, multiple classifiers and resampling techniques to handle class imbalance within the training process, enabling robust pairwise discrimination across all clinically and biologically defined groups.
- Systematic evaluation of classification performance using both hold-out and repeated stratified 5-fold cross-validation protocols, assessing model generalization and robustness through multiple evaluation metrics, including *Accuracy* (Acc), *Precision* ($Prec$), *Recall* (Rec), *Specificity* ($Spec$), *F1-score* ($F1$), Geometric Mean (*Gmean*) and *AUC*. This evaluation explicitly considers performance variations across disease stages and amyloid status.
- Focused assessment of discriminability between adjacent disease stages (e.g., EMCI vs. MCI), with particular emphasis on quantifying how Aβ status modulates classification performance.
- Identification of the most discriminative sMRI features associated with disease stages and amyloid status, supported by permutation-based feature importance analysis to enhance interpretability and provide mechanistic insight into model decision-making. This analysis further enables the characterization of how Aβ status influences feature relevance and drives classification performance across the disease continuum.

2. The state-of-the-art

To map the state-of-the-art in AD and cognitive-impairment classifications, we conducted a literature review from August 2025 to January 2026, using Google Scholar as the primary search engine. We prioritized articles from major publishers (Elsevier, MDPI, Springer Nature), conference proceedings and relevant institutional publications. The search combined terms such as "Alzheimer disease classification", "MCI vs. AD MRI features", "Alzheimer multimodal biomarkers", "amyloid PET deep learning", "plasma p-tau181 machine learning", "hippocampal volume radiomics", "subjective memory complaints" and "cognitive tests dementia ML". Table 1 synthesizes 26 recent studies (2017–2025) spanning multiple modalities — sMRI, PET, plasma biomarkers, clinical/cognitive data and multimodal pipelines for different classification tasks (CN vs. MCI vs. AD, EMCI/LMCI stratifications and amyloid-positivity detection). Methods range from traditional algorithms (SVC, RF, LR, KNN, DT) to deep architectures (CNN/ResNet) and ensemble or mixture-of-experts (ME) approaches. Best metrics range from 70% *Acc* in harder settings (e.g., EMCI vs. LMCI in small cohorts) to 100% in larger, multi-stage datasets, while reported *AUC*s span 0.59–0.98, reflecting the heterogeneity of cohorts (ADNI, NACC, BIOFINDER, IDEAS, AHEAD) and modalities. Notably, plasma biomarkers (p-tau181, $A\beta 42/40$) and multimodal combinations often reach $AUC \geq 0.90$, underscoring the promise of noninvasive and multi-scale integration for early detection and fine-grained stratification of AD.

Table 1
Summary of ML approaches for AD classification.

Article	Approach	Input data	Dataset	Classification task	Classifier	Validation method	Best result
Alatrany et al. [27], 2024	Clinical and Cognitive biomarkers	Demographics, clinical and neuropsychological tests, CDR, FAQ, NPI-Q, GDS	NACC (45,923 subjects)	CN vs MCI vs AD CN vs AD CN vs MCI MCI vs AD	SVC, RF, KNN, NB	Hold-out(80/20) & 5-Fold CV	84.7% *Acc* 97.5% *Acc* 88.1% *Acc* 86.6% *Acc*
Salunkhe et al. [28], 2021	MRI	GLCM features, age, hippocampal volume	ADNI (234 subjects)	CN vs AD	SVC, DT, RF	Hold-out(70/30) & 5-Fold CV	90.2% *Acc*
Oliveira et al. [29], 2024	MRI	GLCM features	ADNI (504 subjects)	CN vs AD CN vs MCI AD vs MCI CN vs MCI vs AD	LSVC, OvsR, ExTC, LR	Hold-out(80/20)	85.2% *Acc*, 0.839 *AUC* 95.1% *Acc*, 0.946 *AUC* 98.5% *Acc*, 0.971 *AUC* 87.1% Acc, 0.947 *AUC*
Altaf et al. [30], 2018	MRI and Clinical data	GLCM features, Clinical scores	ADNI (287 subjects)	CN vs AD CN vs MCI AD vs MCI AD vs MCI vs CN	SVC, KNN, DT, Ensemble Bagging	Hold-out(80/20) & 5-Fold CV	97.8% *Acc* 91.8% *Acc* 85.3% *Acc* 79.8% *Acc*
Beheshti et al. [31], 2017	MRI	Histogram features, FAQ scores	J-ADNI (201 subjects)	AD vs CN	LSVC	10-Fold CV	97.01% *Acc*, 0.97 *AUC*
Swarun Raj et al. [32] 2026	MRI	Pearson correlation + Phase synchronization + Linear SVM	ADNI (270 subjects)	AD vs CN	LSVC, LDA, DT, NB, KNN, QSVC, CSVC	10-Fold CV	93.3% *Acc*
Jitsuishi & Yamaguchi [33], 2022	Multimodal MRI	Tract-based Spatial Statistics	ADNI-3 (66 subjects)	EMCI vs LMCI	SVC, KNN, LR, DT, RF, GBM, AdaBoost	10-Fold CV	70% *Acc*, 0.79 *AUC*
Toshkhujaev et al. [34], 2020	MRI	Cortical thickness, Subcortical volume	GARD, ADNI, NACC, ARWIBO (570 subjects)	CN vs AD EMCI vs LMCI	RBF-SVC	Hold-out(70/30) & 10-Fold CV	97.37% *Acc*, 97.5% *AUC* 95.45% *Acc*, 96.43% *AUC*
Jin et al. [35], 2022	MRI, PET	Fused Deep Features	ADNI (720 subjects)	EMCI vs LMCI	3D ResNet	10 splits 5-Fold CV	83.78% *Acc*, 0.89 *AUC*
Tang et al. [36], 2021	MRI	Cortical & Subcutaneous Volume, Cortical Area Thickness, Surface Area, Hippocampal Subfield	ADNI (560 subjects)	CN vs EMCI CN vs LMCI CN vs AD EMCI vs LMCI EMCI vs AD LMCI vs AD	RF, SVC, DT	10-Fold CV	77.45% *Acc*, 0.59 *AUC* 87.56% *Acc*, 0.81 *AUC* 96.14% *Acc*, 0.92 *AUC* 81.25% *Acc*, 0.75 *AUC* 90.15% *Acc*, 0.85 *AUC* 84.54% *Acc*, 0.89 *AUC*
Inan et al. [37], 2024	MRI	16 selected 2D slices via K-Means++ and Gradient Boosting, EfficientNetV2S transfer features, Bottleneck dense features	ADNI (509 subjects) OASIS-1 (416 subjects)	CN vs AD(ADNI) CN vs MCIc(ADNI) CN vs MCInc(ADNI) CN vs AD(OASIS)	EfficientNetV2S + Dense layers (Deep Integrated)	20-Fold CV	83.64% *Acc* 82.69% *Acc* 71.40% *Acc* 91.54% *Acc*
Wang & Liu [38], 2021	MRI	J-HCPMMP (FreeSurfer processing)	ADNI (96 subjects)	HC vs EMCI HC vs LMCI HC vs AD EMCI vs LMCI LMCI vs AD EMCI vs AD	SVC, KNN, LDA, CNN, DT	5-Fold CV	85.60% *Acc* 92.90% *Acc* 96.80% *Acc* 83.30% *Acc* 84.90% *Acc* 89.50% *Acc*
Venugopalan et al. [39], 2021	MRI, Genetics, Clinical/EHR data	Demographics, Neurological exams and assessments, Medications, Imaging Volumes, Biomarkers	ADNI (2004 subjects)	CN vs MCI vs AD CN vs AD AD vs (CN, MCI)	3D CNN, KNN, SVC, RF, DT	10-Fold CV & External Validation	89% *Acc* 86% *Acc* 89% *Acc*
Zhang et al. [40], 2025	MRI	3D CNN features, Hierarchical Transformer Encoder (HTE) features, Adaptive Feature Fusion (AFFM)	ADNI (896 subjects)	AD vs CN AD vs MCI MCI vs CN	MPPL-SFDA (3DCNN + HTE + MPP-L)	Hold-out	92.7% *Acc*, 92.8% *AUC* 84.2% *Acc*, 84.1% *AUC* 82.7% *Acc*, 82.6% *AUC*
Alickovic et al. [41], 2020	MRI	Histogram features	ADNI (267 subjects)	AD vs CN	RF	10-Fold CV	85.8% *Acc*, 0.885 *AUC*
Palmqvist et al. [42], 2021	Biomarkers, Cognitive tests, APOE genotyping, MRI	Plasma ptau217 & ptau181, Plasma $\beta_{42}/A\beta_{40}$ ratio, Plasma neurofilament light, APOE $\epsilon 4$ genotype, Cognitive test features, Cortical thickness	BioFINDER, ADNI (883 subjects)	Prediction of progression to AD	LR	External validation	0.91 *AUC*
Palmqvist et al. [43], 2024	Biomarker	Plasma ptau217, Aβ42:Aβ40	BIOFINDER (1213 subjects)	AD vs CN AD vs other diagnoses	LR	Hold-out	0.97 *AUC*
Gupta et al. [44], 2019	MRI, Biomarkers, PET, APOE genotype	ROI gray-matter volumes, ROI regional mean intensities, Aβ42, total tau, ptau181p, APOE features	ADNI (158 subjects)	AD vs HC MCIs vs MCIc AD vs MCIs AD vs MCIc HC vs MCIc HC vs MCIs	SVC	10-Fold CV	98.4% *Acc*, 0.983 *AUC* 94.9% *Acc*, 0.936 *AUC* 96.7% *Acc*, 0.968 *AUC* 92.3% *Acc*, 0.946 *AUC* 94.1% *Acc*, 0.964 *AUC* 95.7% *Acc*, 0.952 *AUC*
Pan et al. [45], 2023	Biomarkers, Cognitive tests	Aβ42, Aβ40, ptau181, Aβ42/Aβ40 Ratio, NfL, T-tau, APOE Genotype, Cognitive Scores Demographics	Chinese cohort, ADNI (893 subjects)	$A\beta^+$ vs$^-$ in CN, SMD, MCI, AD, non-AD	DT	CV(1000 times) & External validation	0.94 *AUC*
Nallapu et al. [46], 2024	Biomarkers, MRI, PET	Aβ42, Aβ40, t-tau, ptau181, NfL, APOE, Hippocampal Volume, Demographics, $\epsilon 4$ count	ADNI (2003 subjects)	AD vs CN MCI progression to dementia	RF	Independent testing & 5-Fold CV	0.89 *AUC* 0.76 *AUC*
Kim et al. [47], 2021	MRI, PET	Histogram, GLCM, ISZM, age, sex, APOE	Samsung Medical Center dataset (440 subjects)	$A\beta^+$ vs $A\beta^-$ in MCI	LR	Bootstrap(70/30) & External Validation	0.76 *AUC*
Kim et al. [48], 2021	PET	FDG-PET Volumes, 291 slices per subject	ADNI, KBASE, Severance Hospital dataset (1536 subjects)	Aβ + vs −	CNN	Hold-out(80/20) & External validation	77% *Acc*, 0.86 *AUC*

(*continued on next page*)

Table 1 (*continued*).

Article	Approach	Input data	Dataset	Classification task	Classifier	Validation method	Best result
Lin YS. et al. [49], 2025	Biomarkers	p-tau217, p-tau181, GFAP, APOE ϵ4 status, age, sex	KBASE-V, AD biosignature study (270 subjects)	Aβ + vs −	LR	CV with independent testing	0.946 *AUC*
Meyer et al. [50], 2024	Biomarkers	ptau217, Aβ42/40, APOE proteotype	IDEAS cohort (583 subjects)	Aβ + vs −	LR	CV by original cohort	0.94 *AUC*
Rissman et al. [51], 2024	Biomarkers	Aβ42/Aβ40 ratio, %p-tau217, age, APOE proteotype	AHEAD 3-45 (1080 subjects)	Aβ + vs −	ME	10-Fold CV	0.95 *AUC*

Note: Acc – Accuracy; AD – Alzheimer's Disease; AdaBoost – Adaptive Boosting; ADNI – Alzheimer's Disease Neuroimaging Initiative; APOE – Apolipoprotein E; ARWIBO – Alzheimer's Disease Repository Without Borders; *AUC* – Area Under the Curve; Aβ – Amyloid beta; Aβ^{+}/Aβ^{-} – Amyloid positive/negative; BioFINDER – Biomarkers For Identifying Neurodegenerative Disorders Early and Reliably; CDR – Clinical Dementia Rating; CN – Cognitively Normal; CNN – Convolutional Neural Network; CSVC – Cubic Suport Vector Machine; DT – Decision Tree; EHR – Electronic Health Records; EMCI – Early Mild Cognitive Impairment; ExTC – Extremely Randomized Trees Classifier; FAQ – Functional Activities Questionnaire; FDG-PET – Fluorodeoxyglucose Positron Emission Tomography; FreeSurfer – Automated neuroimaging analysis software; GARD – German Alzheimer's Disease Neuroimaging Initiative; GBM – Gradient Boosting Machine; GDS – Geriatric Depression Scale; GFAP – Glial Fibrillary Acidic Protein; GLCM – Gray-Level Co-occurrence Matrix; HC – Healthy Control; IDEAS – Imaging Dementia—Evidence for Amyloid Scanning; ISZM – Intensity-Size Zone Matrix; J-ADNI – Japanese Alzheimer's Disease Neuroimaging Initiative; J-HCPMMP – Joint Human Connectome Project Multimodal Parcellation; KBASE/KBASE-V – Korean Brain Aging Study for Early Diagnosis and Prediction of Alzheimer's Disease; KNN – k-Nearest Neighbors; LDA – Linear Discriminant Analysis; LMCI – Late Mild Cognitive Impairment; LR – Logistic Regression; LSVC – Linear Support Vector Machine; MCIs – Stable Mild Cognitive Impairment; MCI – Mild Cognitive Impairment; MCIc – Converting Mild Cognitive Impairment; MCInc – Mild Cognitive Impairment not converted; ME – Misture-of-Experts; MPP-L – Multicentric Prototype Pseudo-labeling; MRI – Magnetic Resonance Imaging; NACC – National Alzheimer's Coordinating Center; NB – Naïve Bayes; NfL – Neurofilament light chain; NPI-Q – Neuropsychiatric Inventory Questionnaire; OvsR – One-vs-Rest classification; PET – Positron Emission Tomography; ptau – Phosphorylated tau; QSVC – Quadratic Support Vector Machine; RBF-SVC – Radial Basis Function Support Vector Machine; ResNet – Three-Dimensional Residual Neural Network; RF – Random Forest; ROI – Region of Interest; SVC – Support Vector Machine; TBSS – Tract-Based Spatial Statistics; t-tau – Total tau.

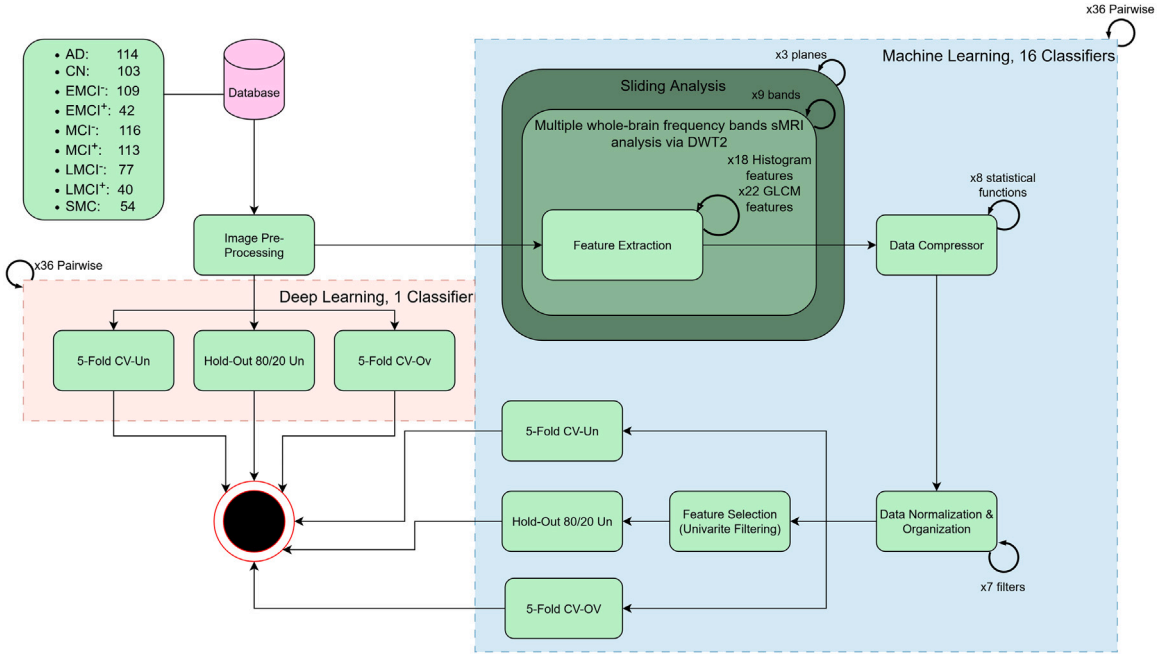

Fig. 1. Methodology flowchart illustrating the proposed framework for AD staging.
Note: DWT2 – Two-dimensional Discrete Wavelet Transform; Feat – Features; GLCM – Gray-Level Co-occurrence Matrix; sMRI – Structural Magnetic Resonance Imaging.

3. Methods

This project was carried out on a laptop running Windows 11, equipped with an Intel Core i7 13th generation CPU, Nvidia rtx 4070 GPU, 32 GB of RAM and 1 TB of storage. Fig. 1 illustrates the methodology followed.

3.1. Dataset

This study uses sMRI images obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (http://adni.loni.usc.edu), including data from all phases of the project. Participants were included if they had a sMRI examination consisting of a 3D Magnetization Prepared Rapid Gradient Echo (MPRAGE) T1-weighted sequence, acquired at field strengths of 1.5T or 3T, using scanners from Siemens, Philips, or General Electric [52–54]. A total of 768 participants were included and divided into 114 with AD, 103 with CN, 151 with EMCI, 229 with MCI, 117 with LMCI and 54 with SMC [55]. Although ADNI is a longitudinal study in which each participant undergoes multiple imaging sessions across different phases, all available phases were included while ensuring that no data from the same participant were redundantly used or shared across different subsets of the analysis (e.g., training and testing), thereby preventing data leakage. Table 2 presents additional information about the patients. Statistical analysis ($p < 0.05$), in the form of *p*-value, were derived from one-way ANOVA with Bonferroni correction [56] for multiple classes, revealed significant differences between all the diagnostic groups for age, cognitive assessment scores and biomarker concentration.

To allow the stratification of MCI groups by Aβ status, additional subgroups were created: EMCI^{-} ($n = 109$), EMCI^{+} ($n = 42$), MCI^{-} ($n = 116$), MCI^{+} ($n = 113$), LMCI^{-} ($n = 77$), and LMCI^{+} ($n = 40$). Aβ status was determined from cerebrospinal fluid (CSF) biomarkers. Participants were classified as Aβ-positive if the ptau181/Aβ42 ratio was ≥ 0.025 or, when this ratio was unavailable, if the Aβ42 concentration was < 980 pg/mL. Participants not meeting these criteria were classified

Table 2
Patient information and assessment tests.

Group	#	Gender		Age	Assessment tests				Biomarkers	
	patients	Female	Male		MMSE	CDR	ADAS-Cog	MoCA	$A\beta42$	ptau181
AD	114	47	67	74.44 ± 0.45	17.71 ± 0.30	1.64 ± 0.04	20.08 ± 0.42	16.03 ± 0.27	683.90 ± 43.13	36.17 ± 0.96
CN	103	59	44	72.46 ± 0.32	28.97 ± 0.21	0.11 ± 0.03	5.70 ± 0.30	26.24 ± 0.19	1324.93 ± 30.10	21.32 ± 0.67
EMCI	151	65	86	71.33 ± 0.45	27.17 ± 0.30	0.46 ± 0.04	8.06 ± 0.42	23.26 ± 0.27	1175.46 ± 42.62	24.49 ± 0.95
MCI	229	91	138	72.65 ± 0.39	27.11 ± 0.26	0.57 ± 0.03	10.50 ± 0.36	23.04 ± 0.24	938.34 ± 36.89	28.95 ± 0.82
LMCI	117	55	62	72.20 ± 0.60	23.05 ± 0.40	0.83 ± 0.05	12.43 ± 0.56	20.06 ± 0.37	882.23 ± 57.37	29.75 ± 1.27
SMC	54	33	21	72.29 ± 0.83	28.92 ± 0.55	0.04 ± 0.07	5.56 ± 0.78	26.75 ± 0.51	1232.49 ± 63.66	22.97 ± 1.41
p-value				5.51e−04	9.48e−10	1.80e−10	3.51e−260	1.74e−53	1.57e−51	1.34e−41

Note: AD – Alzheimer's Disease; ADAS-Cog – Alzheimer's Disease Assessment Scale–Cognitive Subscale; $A\beta42$ – amyloid beta 42; CDR – Clinical Dementia Rating; CN – Cognitively Normal; EMCI – Early Mild Cognitive Impairment; LMCI – Late Mild Cognitive Impairment; MCI – Mild Cognitive Impairment; MMSE – Mini-Mental State Examination; MoCA – Montreal Cognitive Assessment; ptau181 – Phosphorylated tau 181; SMC – Significant Memory Concern.

Table 3
Divided patient information and assessment tests.

Group	# patients	Age	MMSE	CDR	ADAS-Cog	MoCA	$A\beta42$	ptau181
$EMCI^+$	42	73.25 ± 0.72	26.46 ± 0.96	0.65 ± 0.21	9.20 ± 0.47	22.20 ± 1.93	717.04 ± 25.34	33.69 ± 1.61
$EMCI^-$	109	69.17 ± 0.64	27.65 ± 0.47	0.32 ± 0.07	7.13 ± 0.32	23.60 ± 0.65	1516.19 ± 48.30	17.78 ± 0.51
p-value		2.60e−05	2.82e−01	1.36e−01	1.80e−04	5.05e−01	2.05e−32	1.26e−15
MCI^+	113	73.70 ± 0.54	26.17 ± 0.71	0.64 ± 0.11	12.09 ± 0.39	21.61 ± 0.54	580.42 ± 13.68	34.98 ± 1.12
MCI^-	116	72.44 ± 0.78	28.71 ± 0.64	0.43 ± 0.13	8.80 ± 0.36	23.77 ± 0.49	1516.58 ± 59.74	19.35 ± 0.69
p-value		1.57e−01	1.54e−02	2.30e−01	2.06e−08	2.71e−03	9.28e−30	2.83e−26
$LMCI^+$	40	72.44 ± 0.81	19.18 ± 2.63	1.14 ± 0.26	13.45 ± 0.72	17.89 ± 2.35	681.96 ± 23.27	36.68 ± 1.44
$LMCI^-$	77	71.17 ± 1.50	27.67 ± 0.42	0.42 ± 0.08	9.29 ± 0.61	22.67 ± 1.23	1253.45 ± 71.40	16.91 ± 0.93
p-value		3.61e−01	5.28e−03	2.26e−02	2.40e−05	9.81e−02	1.00e−10	7.98e−21

Note: ADAS-Cog – Alzheimer's Disease Assessment Scale–Cognitive Subscale; $A\beta42$ – Amyloid beta 42; CDR – Clinical Dementia Rating; MMSE – Mini-Mental State Examination; MoCA – Montreal Cognitive Assessment; ptau181 – Phosphorylated tau 181.

as Aβ-negative. These thresholds were selected based on prior studies [19,57,58]. Table 3 summarizes the information about the patients after the stratification of MCI groups. The *p*-value, derived from t-test analysis ($p < 0.05$), across the divided subgroups, showed that $A\beta42$ and ptau181 levels differed significantly between $A\beta^+$ and $A\beta^-$ groups, confirming the validity of the stratification criteria. Regarding cognitive tests, MMSE, CDR and MoCA exhibited stage-dependent differences, with statistical significance observed only in some comparisons. For example, MMSE differed significantly between $A\beta^+$ and $A\beta^-$ individuals in the MCI and LMCI groups but not in EMCI. MoCA showed a significant difference only in the MCI group and CDR scores differed significantly between amyloid groups in LMCI but not in EMCI or MCI. In contrast, ADAS-Cog scores were significantly higher in $A\beta^+$ individuals across all stages, suggesting greater cognitive impairment in $A\beta^+$ participants regardless of disease stage, and was the only test that showed statistically significant differences between amyloid groups across all stages. Notably, ADAS-Cog score is widely adopted as the primary cognitive endpoint in AD clinical trials and a core assessment in ADNI [59,60].

3.2. Image pre-processing

To ensure data quality and consistency across images, a pipeline was implemented to prepare the data for feature extraction and model training.

3.2.1. Bias field correction

The first pre-processing step was bias field correction, which addresses low-frequency intensity nonuniformities in sMRI that cause gradual intensity variations across the image that typically arise from radio-frequency coil imperfections, magnetic field inhomogeneities and patient-related factors [61]. The observed image $I(x)$ was modeled as:

$$I(x) = T(x) \cdot B(x) + \eta(x) \tag{1}$$

where $T(x)$ is the true intensity of the tissue, $B(x)$ the bias field, and $\eta(x)$ the additive noise. The bias field is estimated iteratively using a B-spline surface and the corrected image is recovered as:

$$T(x) = I(x)/B(x) \tag{2}$$

3.2.2. Image registration

Image registration was performed to spatially align all brain images to the Montreal Neurological Institute (MNI) space, correcting for motion, positioning differences, and inter-subject anatomical variability [6,62,63]. A multi-stage registration strategy was applied: an initial affine transformation (12-parameter linear model accounting for translation, rotation, scaling and shearing) provided coarse alignment, followed by the SyN diffeomorphic registration, which captured local anatomical deformations and fine-scale structural variations [64,65]. This hierarchical approach ensured both computational efficiency and high anatomical precision, preserving topological consistency and enabling reliable feature extraction across subjects [64,65]. Registration quality was assessed using the Pearson correlation coefficient, the Dice Similarity Coefficient (DSC), and the Jaccard index between each registered image and the template to indicate successful alignment [66,67].

3.2.3. Skull stripping

Another pre-processing step in this pipeline was skull stripping, which removed non-brain tissues (e.g., skull, scalp, dura, eyes and neck) from the images, retaining only brain regions such as gray matter, white matter and subcortical structures, to avoid interfere of non-brain tissues during tissue segmentation [37,68,69]. Skull stripping was performed using the ANTs template-based brain extraction method, which nonlinearly registers each T1-weighted image to a standard anatomical template and applies the corresponding probabilistic brain mask to generate a subject-specific brain-only volume [70].

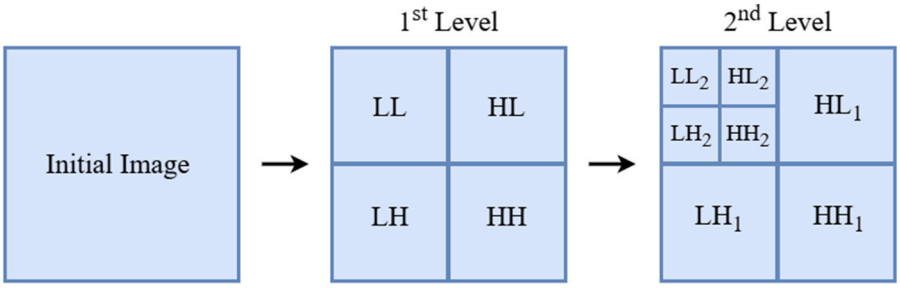

Fig. 2. Principle of 2D Discrete Wavelet Decomposition.
Note: LL – Low–Low; HL – High–Low; LH – Low–High; HH – High–High; LL_2 – Second Level Low–Low; LH_2 – Second Level Low–High; HL_2 – Second Level High–Low; HH_2 – Second Level High–High.

3.2.4. Image normalization and slicing

Finally, to improve intensity consistency across scans, white-stripe normalization was applied, in order to normalize images based on normal-appearing white matter, providing more robust results [71]. Additionally, histogram equalization was performed to minimize intensity variations and reduce noise across slices [72–74]. Following normalization, each volume was registered to the ICBM 152 non-linear atlas (2009a), an isotropic 1 × 1 × 1 mm template (197 × 233 ×189 voxels) with an accompanying brain mask [75]. The provided mask was applied to obtain a brain-only volume prior to slicing. The masked 3D NIfTI volumes were then decomposed along the axial (189 slices), coronal (233 slices), and sagittal (197 slices) planes [76]. Each 2D slice was resized to 224 × 224 pixels and saved in PNG format to ensure consistency and compatibility with downstream feature extraction and ML models.

3.3. Feature extraction

The image slices produced during the pre-processing stage were analyzed to extract distinguishing features based on quantitative metrics. MATLAB 2024b was used to compute the Histogram and GLCM features used in this study.

3.3.1. 2D discrete wavelet transform

DWT2 offers multiple advantages for disease discrimination, including simultaneous multi-resolution analysis in spatial and frequency domains, enhanced detection of fine pathological details and global structural changes in neurodegenerative diseases and natural separation of signal components from noise [77–79].

To make use of these advantages, the obtained slices of each image were subjected to a two-level DWT, which decomposes each 2D image into multiple frequency-spatial components through multi-resolution analysis [80]. The DWT2 applies separable one-dimensional wavelet transforms along both rows and columns, producing four sub-bands at each level: LL (approximation), LH (horizontal detail), HL (vertical detail) and HH (diagonal detail) [81]. Each sub-band captures distinct spatial-frequency characteristics: LL preserves the overall structure, LH emphasizes horizontal edges, HL highlights vertical edges, and HH captures diagonal features [80].

In this study, the two-level decomposition was implemented using the Biorthogonal 1.1 (bior1.1) wavelet. This wavelet was selected based on our previous work in AD MRI analysis, where it provided a good performance with relatively low computational cost [82,83]. Level 1 decomposes the original image $I_{ws}(x, y)$ into LL_1, LH_1, HL_1 and HH_1, while Level 2 further decomposes LL_1 into LL_2, LH_2, HL_2 and HH_2 (Fig. 2). The image was decomposed only up to level 2 via DWT2 because this depth provides an optimal balance between capturing meaningful multiscale details and avoiding excessive loss of spatial resolution or unnecessary complexity. This results in a wavelet collection given by:

$$\mathcal{W} = \{I_{ws}(x, y), LH_1, HL_1, HH_1, LL_2, LH_2, HL_2, HH_2\} \quad (3)$$

3.3.2. Histogram features

Histogram-based (first-order) features are derived directly from an image's pixel intensity distribution, ignoring spatial relationships between pixels [84,85]. They capture basic statistical properties of intensity by computing the histogram $h(i)$, which counts occurrences of each gray level i (0–255) [72]. From this, the probabilities $p(i) = h(i)/N$ (with N as the total number of pixels) are calculated to derive various statistical measures characterizing the intensity distribution [86]. In the context of AD, these features can capture global changes in tissue composition associated with disease progression, including gray matter atrophy and white matter degradation, which directly alter the intensity distribution of the image [86,87] In this study, these features are extracted for each image in the DWT collection $\mathcal{W}$, providing a detailed statistical characterization across different frequency sub-bands, and are described in Table 4. The total number of features per plane is calculated as,

$$\text{Features} = \text{slices} \times \text{features per slice} \times \text{wavelet sub-bands} \quad (4)$$

For the different anatomical planes, this gives:

- Axial: $189 \times 18 \times 9 = 30{,}616$ features
- Coronal: $233 \times 18 \times 9 = 37{,}746$ features
- Sagittal: $197 \times 18 \times 9 = 31{,}914$ features

3.3.3. Gray-level co-occurrence matrix features

The GLCM is a statistical method for analyzing texture in digital images, capturing spatial relationships between pixel intensities. It is constructed by examining pixel pairs separated by a distance d and offset θ, recording the frequency of each gray-level pair (i, j), and normalizing to obtain probabilities $p(i, j)$ [88]. From the GLCM, statistical texture features (gray-level co-occurrence properties) are computed. Unlike histogram features, GLCM features account for spatial relationships and use different offsets to characterize gray-level intensities [84]. Regarding AD, these features can identify tissue organization patterns, such as local intensity variations, since they are highly sensitive to microstructural alterations that go beyond visible atrophy. They can detect changes in spatial pixel arrangements even when overall intensity values remain relatively stable [89,90]. In this study, the GLCM has been computed from each scaled images using four offsets ([1,0], [−1,1], [−1,0], [-1,-1]) with distance $d = 1$, to extract texture metrics. Table 5 presents the features used in this analysis. Each feature was then averaged across the four offsets, and Eq. (4) was used to obtain the total number of features for the different anatomical planes:

- Axial: $189 \times 22 \times 9 = 37{,}422$ features
- Coronal: $233 \times 22 \times 9 = 46{,}134$ features
- Sagittal: $197 \times 22 \times 9 = 39{,}006$ features

Table 4
Extracted histogram features with corresponding equations and definitions.

Feature	Equation	Definition
Mean (mean)	$\mu = \sum_{i=0}^{L-1} i \cdot p(i)$	Indicates the overall brightness of the image.
Median (median)	$\begin{cases} x_{(\frac{N+1}{2})} & \text{if } N \text{ is odd} \\ \frac{x_{(\frac{N}{2})}+x_{(\frac{N}{2}+1)}}{2} & \text{if } N \text{ is even} \end{cases}$	Middle intensity when all pixel values are sorted. $x_{(i)}$ is the ith ordered value.
Standard Deviation (std)	$\sigma = \sqrt{\sum_{i=0}^{L-1}(i-\mu)^2 \cdot p(i)}$	Quantifies how much intensity values deviate from the mean.
Variance (var)	$\sigma^2 = \sum_{i=0}^{L-1}(i-\mu)^2 \cdot p(i)$	Measures intensity dispersion around the mean.
Kurtosis (Kurt)	$\text{kurt} = \frac{\sum_{i=0}^{L-1}(i-\mu)^4 \cdot p(i)}{\sigma^4} - 3$	Indicates whether the distribution has heavy or light tails compared to a normal distribution.
Skewness (Skew)	$\text{skew} = \frac{\sum_{i=0}^{L-1}(i-\mu)^3 \cdot p(i)}{\sigma^3}$	Measures histogram asymmetry.
Maximum (max)	$\max(x) = x_{\max}$	Represents the brightest region in the image.
Minimum (min)	$\min(x) = x_{\min}$	Represents the darkest region in the image.
Range (range)	$R = x_{\max} - x_{\min}$	Indicates the total intensity spread.
Entropy (Entrh)	$H = -\sum_{i=0}^{L-1} p(i) \log_2 p(i)$	Measures the randomness of intensity distribution.
Energy (Enerh)	$E = \sum_{j=1}^{N} x_j^2$	Reflects the overall magnitude of pixel intensities.
Root Mean Square (RMS)	$\text{RMS} = \sqrt{\frac{1}{N}\sum_{j=1}^{N} x_j^2}$	Indicates the effective or average signal strength.
Smoothness (smooth)	$S = 1 - \frac{1}{1+\sigma^2}$	Indicates local intensity variation.
Uniformity (unif)	$U = \sum_{i=0}^{L-1} p(i)^2$	Also known as Angular Second Moment. Measures the uniformity of intensity distribution.
25th Percentile (per25)	$P_{25} = x_{(\lfloor k \rfloor)} + (k - \lfloor k \rfloor)(x_{(\lceil k \rceil)} - x_{(\lfloor k \rfloor)})$	Intensity value below which 25% of pixels fall, where $k = 0.25(N+1)$.
75th Percentile (per75)	$P_{75} = x_{(\lfloor k \rfloor)} + (k - \lfloor k \rfloor)(x_{(\lceil k \rceil)} - x_{(\lfloor k \rfloor)})$	Intensity value below which 75% of pixels fall, where $k = 0.75(N+1)$.
Interquartile Range (IQR)	$\text{IQR} = P_{75} - P_{25}$	Measures central spread while ignoring outliers.
Mode (mode)	$\text{Mode} = \frac{b_m + b_{m+1}}{2}$	Represents the dominant gray level, where $m = \arg\max_i h(i)$ and b_i is the bin boundary.

3.3.4. Data compression

Eight descriptive statistics were computed for each measure series across slices of each plane: mean, median, standard deviation (std), variance (var), 5th percentile (Q_{05}), 95th percentile (Q_{95}), skewness and kurtosis. This procedure yielded 1584 features per anatomical plane for GLCM-based descriptors (9 subbands × 22 measures × 8 statistics), resulting in a total of 4752 features across the three planes. Similarly, 1296 features per plane were generated for Histogram-based descriptors (9 subbands × 18 measures × 8 statistics), resulting in a total of 3888 features. After this step, each feature is denoted by its name, with the DWT level indicated as a superscript and the compressor type as a subscript:

$$featureName_{compressor}^{DWT\,subband} \tag{5}$$

3.3.5. Data normalization

Feature normalization was applied at the subband level per plane using the z-score method. For each classification task, the ith instance of the jth feature, x_{ij} (where $1 \le j \le N_F$, corresponding to $N_F = 176$ for GLCM and $N_F = 144$ for Histogram features per subband), was standardized according to:

$$\tilde{x}_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j}, \tag{6}$$

where μ_j and σ_j denote the mean and standard deviation derived exclusively from the training partition. These parameters were subsequently applied to transform both the training and test sets, strictly preventing data leakage.

3.3.6. Feature selection

In an effort to enhance performance, seven feature selection techniques were employed to identify the most discriminative features available across all subbands: F-Score, Select Percentile, Mutual Information (MI), Chi-squared (χ^2), False Discovery Rate (FDR), L1 regularization (Lasso Fixed) and Tree-Based methods [91]. This process was conducted independently for each pairwise comparison, anatomical plane (single or combined) and feature extraction method (applied individually or in combination). Crucially, to ensure methodological rigor and prevent data leakage, selection was strictly restricted to training data. Each method was utilized to determine the optimal feature subsets, evaluating the set cardinalities ranging from 1 to X, where X represents the total number of features available for a given configuration ($X = 1584$ for GLCM features and $X = 1296$ Histogram features, for each anatomical plane).

The univariate filter methods prioritized features with the strongest statistical relevance, selecting those with the highest scores (e.g., F-value, χ^2 statistic, MI score) or the lowest FDR-adjusted p-values. Conversely, the model-based approaches (Lasso and Tree-Based) identified features based on non-zero coefficients or intrinsic feature importance [91]. By systematically evaluating different selector-classifier combinations, this approach ensured that each model was optimized for its specific dataset characteristics.

3.4. Classification framework

3.4.1. Machine learning

Normalized features were analyzed in Python (version 3.9.12) using various ML models to perform classification and generate the corresponding performance reports. The performance of the model in discriminating a series of pairwise comparisons in all diagnostic categories was evaluated using a set of 16 ML algorithms implemented in Scikit-learn, following a hold-out validation approach. In total, 36 binary classification tasks were performed, which included all possible clinically relevant combinations. Each model processed either features when combining data from all three anatomical planes or features

Table 5
Extracted GLCM features with corresponding equations and definitions.

Feature	Equation	Definition
Autocorrelation (Auto)	$\sum_{i,j} i\,j\,p(i,j)$	Quantifies the linear dependency between neighboring pixel intensities.
Contrast (Cont)	$\sum_{i,j} \lvert i-j\rvert^2\,p(i,j)$	Measures local intensity variation between pixel pairs.
Correlation 1 (Corr1)	$\frac{\sum_{i,j}(i-\mu_i)(j-\mu_j)p(i,j)}{\sigma_i\sigma_j}$	Evaluates the linear correlation between gray levels of neighboring pixels.
Correlation 2 (Corr2)	$\frac{\sum_{i,j} i\,j\,p(i,j)-\mu_x\mu_y}{\sigma_x\sigma_y}$	Alternative formulation for correlation between neighboring gray levels.
Cluster Prominence (ClusP)	$\sum_{i,j}(i+j-\mu_x-\mu_y)^4 p(i,j)$	Describes the asymmetry and peakedness of the GLCM.
Cluster Shade (ClusS)	$\sum_{i,j}(i+j-\mu_x-\mu_y)^3 p(i,j)$	Measures the skewness of the GLCM.
Dissimilarity (Diss)	$\sum_{i,j} \lvert i-j\rvert\,p(i,j)$	Represents the degree of difference between neighboring pixel pairs.
Energy (EnerG)	$\sum_{i,j} p(i,j)^2$	Also known as Angular Second Moment. Indicates texture uniformity.
Entropy (Entrg)	$-\sum_{i,j} p(i,j)\log p(i,j)$	Measures texture randomness.
Homogeneity 1 (Home1)	$\sum_{i,j} \frac{p(i,j)}{1+\lvert i-j\rvert}$	Reflects how close the GLCM elements are to the diagonal.
Homogeneity 2 (Home2)	$\sum_{i,j} \frac{p(i,j)}{1+(i-j)^2}$	Similar to Homogeneity 1 but penalizes larger gray-level differences more strongly.
Maximum Probability (Max_P)	$\max_{i,j} p(i,j)$	Represents the most dominant gray-level pair occurrence.
Sum of Squares (Variance)	$\sum_{i,j}(i-\mu)^2 p(i,j)$	Measures the spread of intensity values around the mean.
Sum Average (Sa)	$\sum_{k=2}^{2N_g} k\,p_{x+y}(k)$	Average of the sum probabilities, where $p_{x+y}(k)=\sum_{i,j:i+j=k} p(i,j)$.
Sum Variance (Sv)	$\sum_{k=2}^{2N_g}(k-SA)^2 p_{x+y}(k)$	Quantifies deviation from the sum average (SA).
Sum Entropy (Se)	$-\sum_{k=2}^{2N_g} p_{x+y}(k)\log p_{x+y}(k)$	Reflects randomness in the summed gray-level distribution.
Difference Variance (Dv)	$\sum_{k=0}^{N_g-1} k^2 p_{x-y}(k)$	Indicates intensity contrast diversity, where $p_{x-y}(k)=\sum_{i,j:\lvert i-j\rvert=k} p(i,j)$.
Difference Entropy (De)	$-\sum_{k=0}^{N_g-1} p_{x-y}(k)\log p_{x-y}(k)$	Measures unpredictability in gray-level differences.
Information Measure of Correlation 1 (Imof1)	$\frac{HXY-HXY1}{\max(HX,HY)}$	Mutual information-based correlation measure using joint and conditional entropies.
Information Measure of Correlation 2 (Imof2)	$\sqrt{1-e^{-2(HXY2-HXY)}}$	Alternative entropy-based correlation measure reflecting nonlinear dependencies.
Inverse Difference Normalized (Idm)	$\sum_{i,j} \frac{p(i,j)}{1+\frac{\lvert i-j\rvert}{N_g}}$	Normalized version of inverse difference; compensates for the number of gray levels N_g.
Inverse Difference Moment Normalized (Idmn)	$\sum_{i,j} \frac{p(i,j)}{1+\frac{(i-j)^2}{N_g^2}}$	Normalized variant of homogeneity emphasizing small intensity differences.

when analyzing each plane individually. These features were derived from extracted per image, each computed across offsets and wavelet sub-bands for each anatomical plane and group comparison. In the hold-out method, the dataset was randomly divided into training and testing subsets, with 80% of the data used for model training and 20% reserved for testing. To address class imbalance in the dataset, Random Under Sampling (RUS) was applied exclusively to training set, in order to randomly remove samples from the majority class until both classes get the same representation [92,93]. To ensure optimal model configuration, a grid search with 3-fold cross-validation was conducted exclusively on the training set. This procedure fine-tuned the critical hyperparameters (e.g., regularization parameter C, kernel type, or tree depth) based on *accuracy* maximization within the training data, ensuring that model selection remained independent of the test data to prevent data leakage.

To ensure a rigorous evaluation, 16 ML classifiers were used to represent different methodological families. The first family of classifiers used were the linear models, containing the Logistic Regression (LR), Logistic Regression with Cross-Validation (LRCV), Elastic Net Logistic Regression (eNLR), Linear Discriminant Analysis (LDA), Passive Aggressive Classifier (PA) and Stochastic Gradient Descent (SGD). Another category of classification method is the support vector machines (SVC), that encompasses both linear and non-linear variants, such as Linear SVC (LSVC), SVC with optimized hyperparameters (OptSVC) and Calibrated SVC (cSVC). A third category of classifiers are the ensemble methods, which include the Random Forest (RF), gradient boosting frameworks including eXtreme Gradient Boosting (XGB), Gradient Boosting (GBC), Light Gradient-Boosting Machine (LGBM) and Categorical Boosting (CB), as well as a Soft Voting Classifier (VCs).

To further assess the stability, robustness, and generalizability of the proposed framework, additional validation and resampling strategies were employed. Based on classification performance, the best-performing classifier for each disease stage was selected using the hold-out strategy. This classifier was then evaluated using stratified k-fold cross-validation (CV) ($k = 5$), enabling comparison with the initial train/test split. Importantly, the hyperparameter tuning and feature selection performed during the hold-out phase were not reused in the CV setting. Instead, the classifier was evaluated using default hyperparameters and the full feature set in each fold. This design choice was adopted to avoid information leakage between model selection and evaluation stages and to reduce the substantial computational burden associated with repeated optimization procedures. Within the CV framework, two resampling strategies (SMOTE and RUS) were evaluated, applied exclusively to the training folds to prevent data leakage. While not constituting a fully nested CV scheme, this approach provides a complementary assessment of model stability, robustness and generalizability.

3.4.2. Deep learning

In parallel with the ML approach, a Deep Learning (DL) framework was implemented using PyTorch and TorchVision to perform pairwise classification directly from MRI slices, without handcrafted features, under a patient-level evaluation scheme.

A ResNet-18 CNN was adopted via transfer learning, with the convolutional backbone frozen to mitigate overfitting given the dataset size, and only the classification head fine-tuned. The final layer was replaced by a dropout layer ($p = 0.5$) followed by a linear classifier. Batch-normalization layers were kept in evaluation mode, and inputs were normalized using ImageNet statistics. Training employed data augmentation (random rotations up to 10° and brightness/contrast jitter), avoiding anatomically implausible transformations. The evaluation pipeline used normalization only.

Datasets were split at the patient level into stratified training (80%) and testing (20%) sets. Class imbalance was addressed using RUS and SMOTE, applied exclusively to the training set. Note that, when SMOTE was used, it operated on intermediate embeddings rather than raw images. The model was trained using Adam, with cross-entropy loss and cosine annealing, for up to 30 epochs with early stopping based on validation loss.

Slice-level predictions were aggregated at the patient level by averaging class probabilities across slices, with the final label obtained via thresholding. Additionally, stratified k-fold CV ($k = 5$) was performed, with internal validation splits and results reported as mean ± standard deviation.

3.4.3. Classification metrics

ML and DL Models' performance was assessed using seven evaluation metrics, including Acc, $Spec$, $Prec$, Rec, $F1$, AUC and $Gmean$.

Acc represents the proportion of correct predictions out of all predictions made. It measures the overall performance of the model [94] and is given by:

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

where TP are the True Positive, (TN) the True Negative, (FP) the False Positive and (FN) the False Negative [95].

Rec, also known as sensitivity or true positive rate, corresponds to the proportion of actual positive cases that were correctly identified by the model [94] and can be defined as:

$$Rec = \frac{TP}{TP + FN} \tag{8}$$

$Prec$, which can also be called as positive predictive value, is the proportion of predicted positive cases that are actually positive [94]. It is given by:

$$Prec = \frac{TP}{TP + FP} \tag{9}$$

$Spec$, also known as true negative rate, is the proportion of actual negative cases that were correctly identified by the model [94], being defined as:

$$Spec = \frac{TN}{TN + FP} \tag{10}$$

$F1$ is the harmonic mean of $Prec$ and Rec, providing a single metric that balances both measures [94]. This metric can be defined as:

$$F1 = 2 \times \frac{Prec \times Rec}{Prec + Rec} \tag{11}$$

AUC represents the area under the Receiver Operating Characteristic (ROC) curve, measuring the model's ability to distinguish between classes across all classification thresholds. This can be given by the following equation:

$$AUC = \int_0^1 TPR(FPR^{-1}(x))dx \tag{12}$$

where TPR is the True Positive Rate and FPR is the False Positive Rate. AUC reflects model performance across all possible decision thresholds rather than relying on a single reference point. This property is especially valuable when distinguishing between adjacent cognitive stages (e.g., EMCI⁻ vs. MCI⁻), where the optimal classification boundary may vary depending on clinical considerations.

Finally, *G-mean*, defined below, is the geometric mean of Rec and $Spec$, providing a balanced measure that is particularly useful for imbalanced datasets [94].

$$Gmean = \sqrt{Rec \times Spec} \tag{13}$$

To determine the optimal classification performance, a composite evaluation criterion was applied, assigning equal weight to Acc and AUC (50% each). This balanced threshold was used to identify the best-performing configuration for each anatomical plane, classifier, and feature selection method in all pairwise comparisons, ensuring greater consistency in model evaluation.

4. Results and discussion

4.1. Hold-out ML classification performance

The best classification performances obtained across all stage comparisons are summarized in Supplementary Table A1. Each entry reports the optimal values of Acc, AUC, $Prec$, Rec, $Spec$, $F1$ and $Gmean$, along with the corresponding anatomical plane and the feature type that led to the best result (Histogram, GLCM or Combination).

The combination feature type proves to be the most robust strategy, consistently securing AUC values exceeding 0.99 in complex stratifications such as AD vs. EMCI⁻ (AUC 0.998) and SMC vs. MCI⁺ (AUC 0.996) [96,97]. However, the standalone Histogram features demonstrated a peak performance in the AD vs. EMCI⁺ comparison (AUC 1.0), suggesting that intensity-based biomarkers alone are sufficient for perfect separability in advanced pathology [86]. In terms of classifiers, linear models (LR and LRCV) were among the most frequently selected high-performing methods, although other classifiers, including VC, OptSVC, RF and LDA, also contributed substantially to the top results. This trend indicates that for the highest-performing feature subsets, the decision boundaries between neurodegenerative stages are largely linear [98].

Regarding feature selection, the MI and F-Score selectors were pivotal in constructing the models with the highest AUCs, often retaining large feature sets (≈ 900–1000 features) to maximize information retention, as seen in the EMCI⁻ vs. MCI⁺ comparison (AUC 0.996). Conversely, selectors like FDR provided efficient models with fewer features (e.g., < 10) while maintaining high AUCs (0.883 for SMC vs. EMCI⁺).

Anatomically, the Coronal plane was frequently associated with the highest AUC values in AD classifications, consistent with its sensitivity to medial temporal lobe atrophy reported in previous studies [82,99]. However, high performance was observed across all anatomical planes. Notably, the Axial plane proved particularly effective for early-stage and SMC cohorts, such as SMC vs. CN (AUC 0.935).

The AUC values also show a strong ability of the proposed framework to distinguish between stages related to Aβ. The system achieved near-perfect discrimination between AD and early-stage cohorts regardless of amyloid status (AUC 1.0 for AD vs. EMCI⁺ and 0.998 for AD vs. EMCI⁻). These results indicate that the extracted features capture structural alterations associated with AD, enabling robust discrimination from early-stage cohorts regardless of the amyloid status. However, the model found it significantly more challenging to distinguish AD from MCI⁺, yielding a maximum AUC of 0.788. Conversely, the model successfully distinguished MCI⁺ vs. LMCI⁻ with a high AUC of 1.0, suggesting that the divergence in amyloid status may facilitates radiomic separation even when clinical symptoms are comparable. When comparing the same clinical stage, EMCI⁻ vs. EMCI⁺ achieved a moderate AUC of 0.843. Although lower than the inter-stage comparisons, this performance remains clinically relevant, indicating that the radiomic

signature detects subtle changes associated with amyloid deposition before visible atrophy appears.

Finally, the SMC cohort showed high distinctiveness from advanced pathological stages, with an *AUC* of 0.996 for both SMC vs. AD and SMC vs. MCI$^+$. Interestingly, the discrimination between SMC vs. CN also reached a high *AUC* of 0.935. This high *AUC* suggests that despite the lack of objective cognitive deficits in SMC patients, there are detectable neuroimaging alterations that distinguish them from truly normal controls, suggesting that the approach has potential to identify early neurodegenerative changes. The discriminative power was less pronounced in SMC vs EMCI$^-$ (*AUC* 0.868), suggesting a potential structural similarity between subjects with subjective complaints and those with early impairment.

Regarding *Gmean* values, the models achieve quite high values (> 0.95) when distinguishing distinct stages, such as AD vs. EMCI$^-$ or LMCI$^+$, EMCI$^+$ vs. MCI$^+$ and SMC vs. AD. Moderate performance (0.8–0.95) is observed for comparisons involving more subtle transitions, such as SMC vs. CN and EMCI$^-$ vs. LMCI$^+$, suggesting that early cognitive changes are detectable but with some limitations. Conversely, lower *Gmean* values (< 0.8) occur mainly in within-stage comparisons, including AD vs. MCI$^+$, EMCI$^-$ vs. EMCI$^+$, EMCI$^-$ vs. LMCI$^-$ and MCI$^-$ vs. MCI$^+$, reflecting the inherent difficulty in discriminating subtle differences within closely related disease stages. Higher values of *Gmean* are generally obtained when positive groups or later disease stages are present (e.g., EMCI$^+$ vs. MCI$^+$, AD vs. EMCI$^-$ or SMC vs. MCI$^+$). In contrast, comparisons involving only negative groups or early stages (e.g., EMCI$^-$ vs. LMCI$^-$, EMCI$^-$ vs. EMCI$^+$, or SMC vs. EMCI$^-$) generally have lower *Gmean* (often < 0.8–0.85), indicating that the model struggles to balance sensitivity and *Spec* when class differences are subtle, highlighting the challenge of early detection and progression monitoring.

4.2. Hold-out ML feature importance analyses

4.2.1. Analysis of the ten top-tier important features per pairwise best model

The permutation scores for each pairwise comparison are presented in Supplementary Table A2, where only positive values were considered. This table reveals the extent to which the classifier's predictive performance is dependent on specific radiomic features. For classifications involving distinct clinical extremes, the model exhibits a heavy reliance on individual biomarkers. Notably, in the comparison of SMC vs. EMCI$^+$, shuffling the top feature ($Auto^{HH2}_{avg}$) results in a massive performance drop of 16.84%. Similarly, distinguishing AD from CN and from MCI$^-$ relies critically on single features such as max^{LH1}_{q95} and $Imo1^{LH1}_{med}$, where removing the signal from these variables causes accuracy penalties exceeding 12%. This magnitude indicates that these specific first-order and correlation-based features are "load-bearing" components of the decision boundary: without them, the model's diagnostic capability degrades significantly.

Conversely, comparisons between clinically adjacent or subtle stages are characterized by much lower permutation scores, implying a distributed feature importance. For tasks such as CN vs. EMCI$^+$ and SMC vs. AD, the top features yield performance drops of only 4.31% and 2.50%, respectively. This low impact suggests that no single feature acts as a dominant discriminator. Instead, the model likely exploits high feature redundancy, where the information lost by shuffling one variable is compensated by others. This pattern highlights the complexity of early-stage detection, where the classifier must aggregate marginal gains from multiple weak learners rather than relying on a definitive radiomic signature.

Furthermore, the stability of these dependencies varies considerably. High standard deviations relative to the mean scores (e.g., AD vs. MCI$^-$ at $12.93 \pm 4.84\%$) indicate that the importance of these features fluctuates across CV folds, suggesting that the model's reliance on them is unstable and dependent on specific data splits. Additionally, a distinction in feature types is evident, while high-dependency tasks (involving AD) are driven by intensity changes (*max*, *var*), the subtler MCI subgroup distinctions rely more on textural complexity measures.

Across the 36 pairwise comparisons analyzed, PromClus proves to be the most pervasive feature, appearing in the top 10 rankings for 15 separate tasks, particularly dominating the distinctions involving MCI subgroups. Close behind, Imo1 demonstrates remarkable consistency, being selected as a key biomarker in 14 tasks. Similarly, Imo2 and the first-order statistic Max are highly frequent, each appearing in 13 tasks. This frequency distribution highlights that while intensity extremes (Max) are essential for detecting severe disease, the structural complexity of texture ($PromClus$, $Imo1$ and $Imo2$) is the primary tool the model uses to navigate the majority of the clinical spectrum.

4.2.2. Analysis of the importance power of GLCM and histogram individual features for the best pairwised models

Looking at supplementary Table A3, related with the cumulative permutation scores of different extracted features across the 36 pairwise comparisons, histogram-derived features show the widest coverage. Skew appears in 27 of 36 comparisons, Range in roughly 28 of 36, Min in about 30 of 36, and Std in about 24 of 36. The strongest texture GLCM features, Imof1 and Imof2, appear in about 26 of 36 comparisons. Very sparse or irrelevant features include Sv (0 appearances) and central tendencies such as Median or Mode (about 5 appearances), indicating that simple central-value metrics contribute little for the best models.

Analyzing the "Top 5" per pairwise comparisons (which features most often land among the five highest scores), the recurring leaders are Imof1 (16 Top-5 hits), range (15 Top-5), Corr2 (14 Top-5), Max (13 Top-5), ClusP (11) and two histogram shape measures, Kurt and skew (each 10 Top-5). Imof2 follows with 9 Top-5 hits. These eight features constitute the principal "top-tier" set. Low-impact features have few Top-5 appearances despite moderate coverage: Auto (3 Top-5), var (3), and a cluster at 2 hits (Enerh 2, Max_p 2, Cont 2, RMS 2, De 2, Diss 2, Entrg 2, IQR 2). Several features contribute essentially with no discriminative value (0 Top-5), including Enerh, Idm, Idmn, median, per25, mode and Sv.

In terms of discriminative strength (largest cumulative permutation scores), texture dominates in amyloid-negative or texture-driven contrasts: Imof1 reaches 83.04% in AD vs. MCI$^-$ and 74.12% in SMC vs. MCI$^-$; Imof2 reaches 61.41% in AD vs. MCI$^-$; ClusS reaches 59.67% and Corr2 reaches 52.61% in the same comparison. For amyloid-positive contrasts, intensity features dominate: Std peaks at 76.72% in AD vs. EMCI$^+$ and Max at 59.38% in the same column. Other notable histogram highs include Skew at 49.55% (AD vs.CN) and Kurt at 42.65% (SMC vs. LMCI$^-$).

SMC contrasts exhibit a hybrid profile. SMC vs. AD behaves like an intensity-driven separation (Skew = 48.64%, Std = 43.03%), whereas SMC vs. MCI$^-$ behaves like a texture-driven separation (Imof1 = 74.12%).

The amyloid split is clear: in "negative" comparisons (-), texture features crowd the highest ranks (e.g., AD vs. MCI$^-$ is effectively a texture sweep: Imof1, Imof2, ClusS, Corr2, Dv). In "positive" comparisons (+), intensity spread and peaks (Std, Max, Range) shoulder aside texture; AD vs. EMCI$^+$ is the strongest example (Std 76.72%, Max 59.38%). This pattern suggests amyloid positivity manifests as global intensity shifts, whereas amyloid negativity or earlier stages are better captured by spatial co-occurrence structure.

4.2.3. DWT2 subbands' importance for the best pairwise models

Supplementary Table A4 presents the cumulative permutation scores for features extracted from images across different stages of cognitive impairment. Negative values were excluded, and only positive contributions were considered.

A comparative analysis reveals that image decomposition via DWT2 is far more effective than analyzing the original image (Level 0). In nearly all pairwise comparisons, the original image consistently

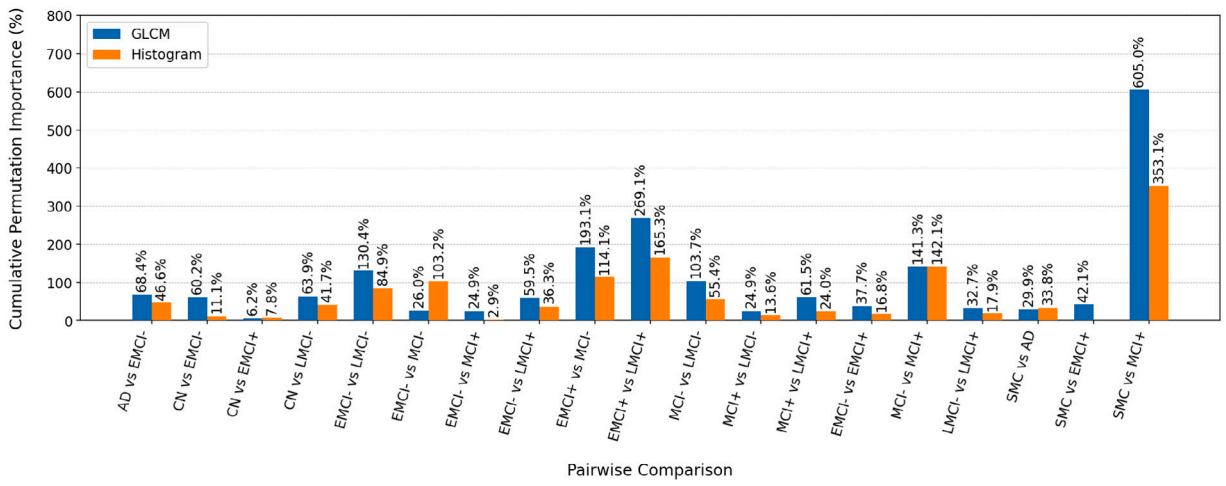

Fig. 3. Feature Type Permutation Comparison.
Note: AD – Alzheimer's Disease; EMCI - Early Mild Cognitive Impairment; EMCI$^+$ – Early MCI, amyloid-positive; EMCI$^-$ – Early MCI, amyloid-negative; LMCI – Late Mild Cognitive Impairment; LMCI$^+$ – Late MCI, amyloid-positive; LMCI$^-$ – Late MCI, amyloid-negative; LR – Logistic Regression; MCI – Mild Cognitive Impairment; MCI$^+$ – MCI, amyloid-positive; MCI$^-$ – MCI, amyloid-negative; SMC – Subjective Memory Complaints.

receives the lowest rankings (typically 7th to 9th), suggesting that raw pixel intensities alone lack the discriminative power necessary for accurate discrimination. In contrast, the sub-bands at Level 2 consistently achieve the top rankings, indicating that the most relevant diagnostic information is encoded within specific spatial frequencies and textural details.

Among the decomposed image components, the Level 2 sub-bands — specifically *HL2* and *LH2* — emerge as the most significant, often securing the Rank 1 position. Other studies from Lahmiri also prove that the usage of HL2 and LH2 improve brain sMRI classification [100]. In the context of image processing, these bands correspond to specific directional variations (horizontal and vertical details). The fact that these specific orientations yield the highest cumulative permutation scores implies that the pathological changes associated with these conditions are best detected through these specific textural patterns rather than coarse approximation coefficients (*LL*) or the original image. The values exceeding 100% confirm the cumulative nature of the extracted metrics within this subbands, representing the aggregated importance of multiple significant features identified during the permutation testing.

4.2.4. Analysis of the importance power of GLCM and histogram methods within the best combined pairwised models

Fig. 3 compares the cumulative permutation scores, calculated from positive contributions, of two distinct feature extraction techniques — GLCM and Histogram analysis — within the context of combined models that utilize both to achieve optimal performance.

A clear pattern emerges showing that GLCM-based features generally contribute more significantly to the discrimination tasks than Histogram-based features. In the majority of pairwise comparisons (15 out of 19), the GLCM contribution is higher, often by a substantial margin. In particular, for the SMC vs. EMCI$^+$ comparison, although the model combines both feature types, the Histogram features were irrelevant for this discrimination and only GLCM features added meaningful predictive value. Moreover, in the SMC vs. MCI$^+$ comparison, GLCM features accumulate a massive score of 605.00%, nearly double that of the Histogram features (353.09%). This suggests that for most diagnostic boundaries, the spatial relationships and complex textural patterns captured by GLCM provide the primary source of discriminative information.

Nevertheless, Histogram features scores are superior in specific contexts, including specific contexts: EMCI$^-$ vs. MCI$^-$ (103.22% vs. 26.00%) and SMC vs. AD (33.82% vs. 29.85%). This indicates that for certain clinical differentiations, global intensity variations (e.g., overall brightness or contrast shifts) are more informative than local textural details. Additionally, in the MCI$^-$ vs. MCI$^+$ comparison, the contributions are remarkably balanced (141.30% vs. 142.07%), demonstrating that for specific subgroups, both textural complexity and global intensity distribution are equally vital.

4.3. Generalization and model quality assessment via cross-validation

The ML results in Fig. 4 and in Supplementary Tables A1 and A5 reveal a clear discrepancy between hold-out and CV performance, indicating a substantial generalization gap. This effect was quantified by computing the relative reduction in *AUC* between hold-out (NoCV) and CV across all 36 pairwise comparisons. For CV-Oversampling (CV-Ov), 8 comparisons (22%) showed reductions $\leqslant$ 5%, 12 (33%) between 5%–10%, 14 (39%) between 10%–20%, and 2 (6%) $>$ 20%. For CV-Undersampling (CV-Un), the corresponding values were 6 (17%), 11 (31%), 17 (47%) and 2 (6%), respectively.

These results confirm that hold-out evaluation systematically overestimates performance, with approximately 80% of comparisons showing *AUC* reductions $>$ 5% and nearly half $>$ 10%. While both resampling strategies reveal this bias, CV-Ov produces a wider dispersion, including more extreme drops, whereas CV-Un yields more consistent, predominantly moderate degradation.

AUC was adopted as the primary metric due to its discriminative capability and relative robustness to class imbalance. However, the observed variability under CV indicates that it must be interpreted alongside class-specific metrics to assess model stability. Across all comparisons, NoCV consistently yields higher *AUC* values. For example, in AD vs EMCI$^-$, *AUC* decreases from 0.998 (NoCV) to 0.9578 ± 0.0202 (CV-Ov) and 0.9519 ± 0.0309 (CV-Un), illustrating the optimistic bias even in well-separated groups.

A stage- and biomarker-dependent pattern is evident. Comparisons involving early disease stages, particularly EMCI, show lower *AUC* and higher variability, reflecting the difficulty of detecting subtle pathological differences. This is most pronounced in within-stage amyloid stratification (e.g., EMCI$^+$ vs. EMCI$^-$), where performance approaches chance level, indicating substantial overlap between Aβ^+ and Aβ^- individuals at early stages.

As the disease progresses, discriminative performance improves. In MCI and LMCI, the separation between Aβ^+ and Aβ^- groups becomes more evident, suggesting that the pathological signal associated with amyloid accumulation strengthens over time. In particular, LMCI$^+$

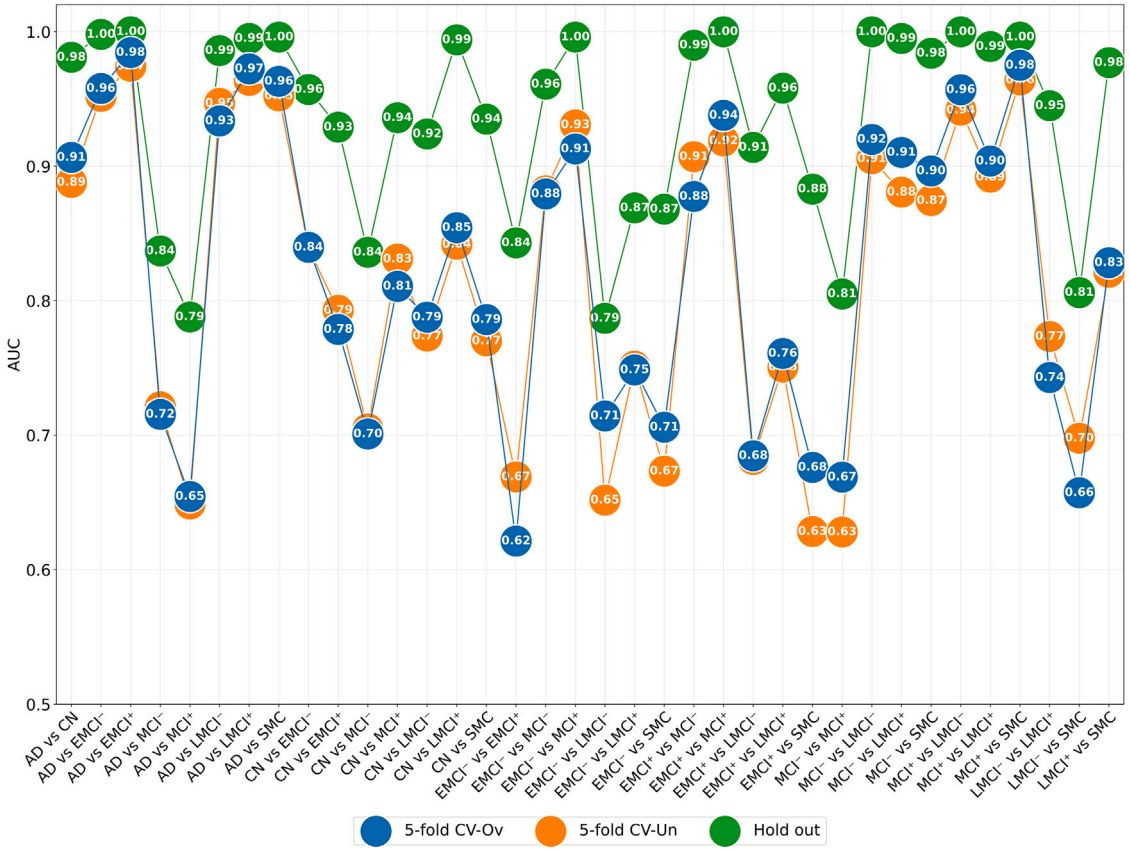

Fig. 4. ML - Hold out vs 5-fold Un vs 5-fold Ov.
Note: AD – Alzheimer's Disease; EMCI - Early Mild Cognitive Impairment; EMCI⁺ – Early MCI, amyloid-positive; EMCI⁻ – Early MCI, amyloid-negative; LMCI – Late Mild Cognitive Impairment; LMCI⁺ – Late MCI, amyloid-positive; LMCI⁻ – Late MCI, amyloid-negative; LR – Logistic Regression; MCI – Mild Cognitive Impairment; MCI⁺ – MCI, amyloid-positive; MCI⁻ – MCI, amyloid-negative; SMC – Subjective Memory Complaints.

vs. LMCI⁻ comparisons generally yield higher *AUC* than their EMCI counterparts, indicating enhanced separability in later stages.

Performance further increases with clinical distance between groups. Comparisons involving distant stages (e.g., EMCI vs. AD or MCI vs. AD) consistently achieve higher *AUC* values, whereas adjacent-stage or within-stage comparisons (e.g., EMCI⁺ vs. EMCI⁻ or MCI⁻ vs. MCI⁺) remain challenging. This behavior reflects the continuous nature of disease progression and the gradual emergence of discriminative pathological features.

Overall, these findings indicate that model performance is intrinsically linked to both disease stage and amyloid status, with early-stage and within-stage distinctions remaining inherently challenging due to substantial biological overlap, while later stages and cross-stage comparisons provide clearer separability as pathological changes accumulate.

4.4. Machine learning versus deep learning

The comparison between ML and DL approaches (Fig. 5 and Supplementary Table A6) indicates that increased model complexity did not translate into systematic performance gains. Across the evaluated pairwise classifications, ML achieved equal or higher *AUC* values in most cases, while also demonstrating more consistent behavior across *Acc*, *Rec*, *Prec*, *Spec*, *F1* and *Gmean*. This suggests that performance is primarily limited by the intrinsic separability of the clinical groups and the stability of the learning process under limited and imbalanced data conditions, rather than model capacity.

Despite we performed DL in hold-out (see results at Supplementary Table A7), to ensure a fair comparison, only CV results were considered, as the ML hold-out (NoCV) setting produced overly optimistic estimates (with *AUC* reaching 100% in some cases), whereas CV provides more robust and generalizable performance evaluation.

In clinically well-separated comparisons, both ML and DL (Supplementary Table A6) achieved high discriminative performance, although ML was generally superior or comparable, outperforming DL in most tasks (34 out of 36). For instance, in AD vs. CN, ML achieved *AUC* values of 0.906 and 0.888 (CV-Ov/CV-Un), compared to 0.865 and 0.859 for DL. Similarly, in AD vs. EMCI⁻, both approaches performed strongly, with ML (0.958/0.952) and DL (0.956/0.958), indicating that clear clinical differences can be captured by both methods.

The advantage of ML becomes more evident in more challenging and clinically relevant comparisons. In AD vs. EMCI⁺, ML achieved *AUC* values of 0.985 and 0.974, compared to 0.940 and 0.942 for DL. Despite relatively high *AUC*, DL exhibited imbalanced metric profiles, with oversampling yielding *Rec* of 52.36%, *F1* of 46.76% and *Spec* of 100%, indicating suboptimal decision behavior. A similar pattern was observed in AD vs. LMCI⁺, where DL again showed high *Spec* (100%) but low *Rec* (51.25%) and *F1* (44.92%).

In moderate-difficulty comparisons, DL performance degraded more substantially than ML. For example, in CN vs. EMCI⁻, ML achieved *AUC* values of 0.840 and 0.839, whereas DL decreased to 0.707 and 0.683. Likewise, in CN vs. EMCI⁺, ML obtained 0.779 and 0.793, compared to 0.623 and 0.689 for DL. This suggests that DL is less robust when class differences are subtle and sample size is limited, whereas ML benefits from lower model complexity and better adaptation to tabular clinical and biomarker data.

Fine-grained comparisons between neighboring disease stages were challenging for both approaches, but ML remained more consistent. For EMCI⁺ vs. LMCI⁻, ML achieved *AUC* values of 0.685 and

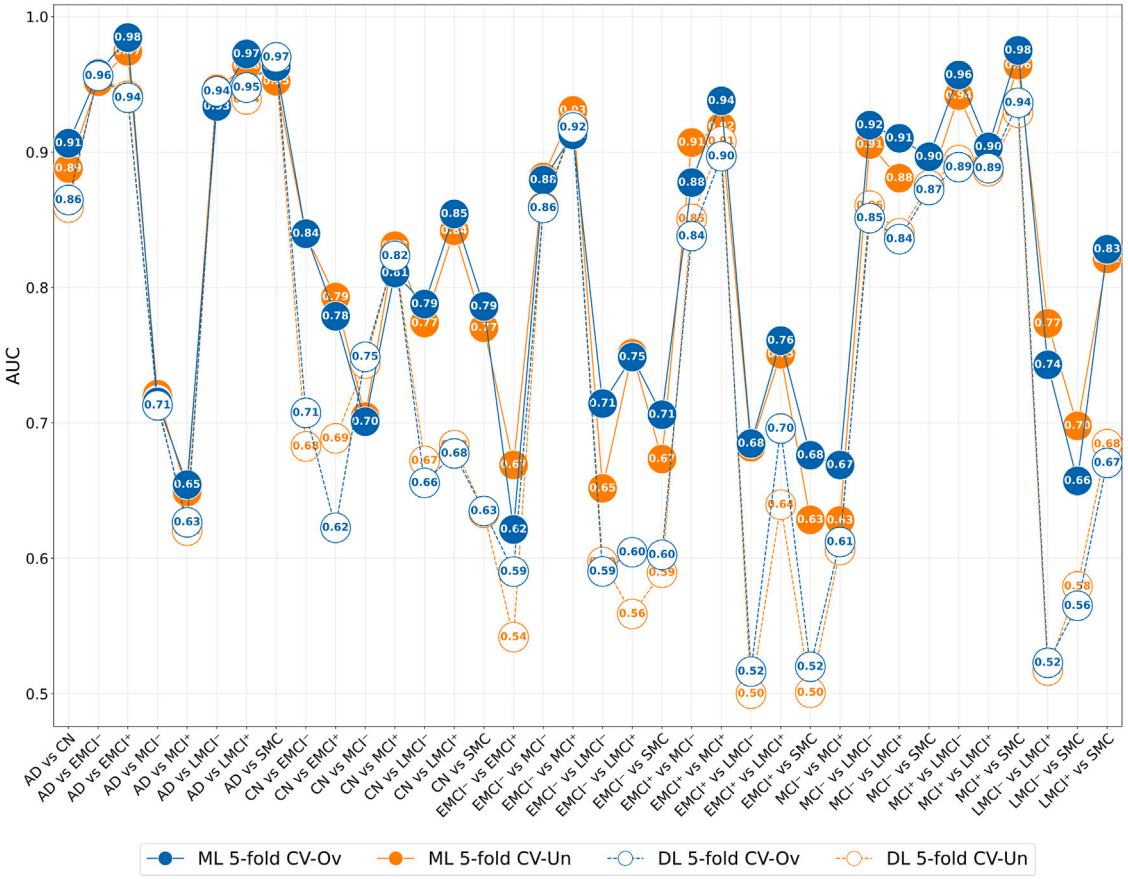

Fig. 5. ML vs DL *AUC* per cross validation method comparison.
Note: AD – Alzheimer's Disease; EMCI - Early Mild Cognitive Impairment; EMCI$^+$ – Early MCI, amyloid-positive; EMCI$^-$ – Early MCI, amyloid-negative; LMCI – Late Mild Cognitive Impairment; LMCI$^+$ – Late MCI, amyloid-positive; LMCI$^-$ – Late MCI, amyloid-negative; LR – Logistic Regression; MCI – Mild Cognitive Impairment; MCI$^+$ – MCI, amyloid-positive; MCI$^-$ – MCI, amyloid-negative; SMC – Subjective Memory Complaints.

0.682, while DL dropped to 0.516 and 0.423. Similarly, for LMCI$^-$ vs. LMCI$^+$, ML achieved 0.743 and 0.774, whereas DL remained close to chance level (0.523 and 0.517). These results reflect the substantial overlap between adjacent clinical stages.

A key observation is the instability of DL across secondary metrics. In several comparisons, DL with oversampling produced extreme *Spec* values (0% or 100%), often combined with *Rec* around 50% and low *F1*. This behavior indicates unstable or degenerate decision boundaries, likely driven by limited sample size and the sensitivity of DL models to class-balancing strategies. In contrast, ML maintained more balanced trade-offs across metrics, even in difficult tasks.

The effect of the sampling strategy also differed between approaches. In ML, oversampling and undersampling produced relatively small *AUC* variations, indicating stable behavior. In DL, oversampling frequently led to highly asymmetric predictions, while undersampling improved metric balance but did not consistently outperform ML. This suggests that DL models are more sensitive to altered training distributions introduced by resampling.

4.5. Performance comparison with the state-of-the-art

A direct comparison between the studies reported in Supplementary Table A5 with the state-of-the-art results at Table 1 must be interpreted with caution, since the methodological setting of the present work differs from most prior contributions in three important aspects. First, previous studies use heterogeneous input modalities, ranging from single-modality structural imaging to multimodal combinations of MRI, PET, genetic, fluid, plasma, clinical and cognitive variables. Second, validation protocols vary substantially, including single hold-out splits, repeated hold-out strategies and k-fold CV, which can strongly affect reported performance and limit the comparability of absolute values. Third, and most importantly, our experimental design explicitly incorporates Aβ stratification across AD-related stages. This stratification is rarely considered in diagnostic classification studies, but it is biologically relevant because amyloid status captures an underlying dimension of AD pathophysiology that is not reflected by clinical stage alone. As a result, several of the tasks evaluated here have no direct counterpart in the literature, and comparisons should therefore be understood as contextual benchmarks rather than strict head-to-head evaluations.

To ensure a conservative and robust assessment, all comparisons in this section are based on the 5-fold CV protocol. Although the model achieved higher scores under a hold-out scheme, CV provides a more stable and less partition-dependent estimate of generalization. This choice is particularly important in AD classification, where cohort composition, disease-stage imbalance, and biomarker-defined heterogeneity can markedly influence performance. Studies based on hold-out validation are therefore discussed only as contextual references, since such protocols may be more sensitive to favorable data partitions and can yield optimistically biased estimates. In this sense, the values reported here should be viewed as conservative estimates of model performance, particularly when contrasted with studies using less stringent validation schemes.

The AD vs. CN classification is the only task consistently reported across prior studies and therefore serves as the most natural reference point. Under 5-fold CV, our model achieves 82.02% *Acc* and 0.906 *AUC*. Although higher absolute accuracies have been reported, including 98.4% *Acc* [44], 97.8% [30], 97.37% *Acc* [34], and 97.01% *Acc* [31],

these results were obtained under substantially different methodological conditions. Several studies rely on multimodal data combining MRI, PET, genetics and fluid biomarkers [39,44], whereas others include clinical and cognitive scores [27,30,31], which are highly informative for discriminating clinically established AD from cognitively normal individuals. In addition, some of these reports use validation schemes that are less conservative than the CV protocol adopted here [30,31,34,44]. A more balanced comparison is therefore with single-modality imaging-based studies. In this setting, our model is highly competitive, with an *AUC* of 0.906 exceeding both Alickovic et al. (0.885, 10-fold CV) [41] and Oliveira et al. (0.839, single hold-out) [29]. The comparison with Oliveira et al. should be interpreted particularly cautiously, since their result was obtained under a hold-out protocol, whereas the present value reflects a stricter CV estimate.

The results are particularly strong for AD versus prodromal-stage comparisons, where the model consistently achieves high discriminative performance. AD is separated from EMCI with 90.13% *Acc* and 0.958 *AUC* in the Aβ^- subgroup, and with 95.48% *Acc* and 0.985 *AUC* in the Aβ^+ subgroup. For AD vs. LMCI, *Acc* ranges from 85.83% to 90.88%, while *AUC* ranges from 0.934 to 0.972. These values are at least comparable to, and in several cases exceed, established CV benchmarks. For example, Tang et al. reported 90.15% *Acc* and 0.85 *AUC* for EMCI vs. AD, and 84.54% *Acc* and 0.89 *AUC* for LMCI vs. AD [36]. Wang and Liu reported 84.90% *Acc* for LMCI vs. AD [38]. The present model therefore compares favorably with directly relevant CV studies, particularly in terms of *AUC*, which is especially informative in settings where class distributions and decision thresholds may differ across cohorts. The advantage also extends to CN vs. EMCI, where the Aβ^- comparison reaches 78.76% *Acc* and 0.840 *AUC*, substantially exceeding the 0.59 *AUC* reported by Tang et al. [36] at a comparable *Acc* level.

The EMCI vs. LMCI classification is widely recognized as one of the most challenging problems in AD-related modeling, because both groups occupy neighboring positions along the clinical continuum and often show overlapping structural and cognitive profiles. In our study, amyloid stratification further increases the granularity of this task, producing four EMCI vs. LMCI comparisons rather than a single pooled contrast. Across these comparisons, *Acc* ranges from 67.21% to 75.17%, while *AUC* ranges from 0.685 to 0.761. These values are within the range reported by established studies, such as Jitsuishi and Yamaguchi, who obtained 70% *Acc* and 0.79 *AUC* using 10-fold CV [33], and Jin et al., who reported 83.78% *Acc* and 0.89 *AUC* [35]. Importantly, the present results were obtained in a more fine-grained amyloid-stratified setting, which is intrinsically more demanding than a conventional pooled EMCI vs. LMCI comparison. Thus, rather than indicating weak performance, these results reflect the difficulty of separating biologically and clinically adjacent subgroups when additional pathophysiological structure is explicitly preserved.

Within-stage amyloid classification represents the most distinctive component of the present evaluation and is also the least directly comparable with conventional stage-based classification studies. Using structural imaging alone, the model achieves *AUC* values from 0.621 to 0.669 within EMCI, from 0.743 to 0.773 within LMCI, and from 0.628 to 0.669 within MCI. Dedicated amyloid-classification studies report higher values, typically from 0.94 to 0.97 *AUC* [43,45,49–51]. However, these studies rely on biomarkers such as plasma p-tau217, Aβ42/Aβ40, GFAP, and APOE, which are direct molecular or genetic proxies of amyloid pathology. The methodological objective is therefore fundamentally different. While those approaches detect amyloid status using variables closely linked to the underlying molecular process, our model attempts to infer amyloid status indirectly from structural imaging patterns. A fairer reference is provided by imaging-based studies, which report *AUC* values between 0.76 and 0.86 [47,48]. Against this background, the present findings are encouraging, especially because they demonstrate that structural alterations contain measurable, although indirect, information about amyloid-defined disease heterogeneity.

The AD vs. MCI$^-$ and AD vs. MCI$^+$ tasks show comparatively lower scores, but this outcome is expected given the biological structure imposed by the study design. At first glance, these values appear lower than pooled AD vs. MCI results reported by Oliveira et al. (98.5%, single hold-out) [29] and Altaf et al. (85.3%) [30]. However, these studies treat MCI as a single homogeneous category, whereas the present work deliberately separates amyloid-defined MCI subgroups. This distinction is critical. Amyloid-positive MCI is expected to be biologically closer to AD than amyloid-negative MCI, making its separation from AD intrinsically more difficult. The reduced separability of MCI$^+$ from AD is therefore consistent with the AD continuum and should not be interpreted as a methodological limitation. On the contrary, it supports the biological validity of the stratified design, since the model reflects the expected continuum between prodromal and clinically established disease. By preserving amyloid-defined heterogeneity, the analysis reveals patterns that would be obscured in conventional non-stratified AD vs. MCI classifications.

Finally, a substantial proportion of the evaluated tasks, particularly those involving SMC and amyloid-stratified pairs, have no direct counterpart in the state-of-the-art studies listed in Table 1. This lack of direct comparators highlights one of the main contributions of the present work. Rather than simply reproducing standard diagnostic contrasts, the study evaluates clinically and biologically more refined classification scenarios that better reflect the heterogeneity of the AD continuum. A particularly notable result is the discrimination between SMC and AD, where the model achieves 97.1% *Acc* and 0.996 *AUC*. The SMC stage has been only marginally addressed in previous work, such as Pan et al. [45], and only in the context of amyloid-positivity prediction from biomarkers rather than direct diagnostic classification. To our knowledge, this is the first report of direct SMC vs. AD classification performance in this setting using imaging-based features. Taken together, these results support not only the competitive performance of the model, but also the added value of evaluating AD-related classification under an amyloid-stratified framework.

5. Conclusions

This study proposes an sMRI-based radiomics framework for the discrimination of AD across six clinical stages (CN, SMC, EMCI, MCI, LMCI, AD), including stratification of MCI subgroups according to amyloid status (Aβ^+/Aβ^-). By combining histogram- and texture-based features with ML models, the proposed approach achieves robust and consistent performance across a wide range of pairwise classification tasks.

The results demonstrate reliable discrimination not only between well-separated clinical groups, such as AD vs. CN, but also across more subtle disease transitions, including early MCI stages and amyloid-status stratification. High *AUC* values were observed in most inter-stage comparisons, particularly in AD vs. EMCI$^+$ and AD vs. EMCI$^-$, indicating that sMRI-derived radiomic features capture structural alterations associated with disease progression. However, more challenging comparisons involving adjacent stages or within-stage amyloid stratification, such as EMCI$^-$ vs EMCI$^+$ and MCI$^-$ vs MCI$^+$, remain difficult, reflecting intrinsic overlap between these groups.

Feature analysis revealed complementary contributions of intensity- and texture-based descriptors, with GLCM features consistently providing the most discriminative information, particularly at multi-scale representations such as HL2 and LH2 sub-bands. Linear classifiers, including LR and LRCV, frequently achieved the highest *AUC* values, suggesting that the most informative feature spaces are largely linearly separable. Anatomically, coronal plane features were more informative for AD-related distinctions, whereas axial features were more relevant for early-stage and SMC-related comparisons.

The inclusion of Aβ stratification enabled a more biologically grounded analysis of disease progression. The results indicate that Aβ status contributes to improved discriminability in later stages,

where pathological differences are more pronounced, while remaining challenging in early-stage and within-stage comparisons. This suggests that the influence of amyloid burden on structural MRI patterns is progressive and stage-dependent.

Despite these promising findings, several limitations should be considered. The study relies exclusively on the ADNI cohort and T1-weighted sMRI, which may limit generalizability across populations and imaging protocols. In addition, distinguishing amyloid-positive and negative subjects within adjacent stages remains challenging, suggesting that subtle pathological differences may not be fully captured by sMRI alone. The relatively limited size of some subgroups and the restricted exploration of DL approaches may also have influenced performance and model stability. Overall, the results indicate that the proposed approach is competitive with the single-modality state of the art where direct comparison is possible. Its main contribution lies in a unified, amyloid-stratified, fully CV evaluation framework spanning all stages and pairwise comparisons, including many scenarios not addressed in previous studies. Notably, all performance estimates were obtained using CV, whereas hold-out strategies may lead to overly optimistic performance estimates at first step, reinforcing the robustness and reliability of the reported results.

Future work should focus on multimodal biomarker integration, including PET and CSF, validation in independent cohorts and evaluation in clinically realistic settings, such as decision-support systems and longitudinal disease monitoring.

CRediT authorship contribution statement

José Menezes: Writing – review & editing, Writing – original draft, Investigation, Conceptualization. **Maria Inês Barbosa:** Writing – review & editing, Writing – original draft, Validation, Supervision, Investigation, Funding acquisition, Conceptualization. **Pedro Miguel Rodrigues:** Writing – review & editing, Writing – original draft, Validation, Supervision, Investigation, Funding acquisition, Conceptualization.

Funding statement

This work was supported by Fundação para a Ciência e a Tecnologia (FCT), Portugal, through the project UID/50016/2025.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by National Funds from FCT - Fundação para a Ciência e a Tecnologia through project UID/50016/2025. Maria Inês Barbosa also thanks FCT and the Recovery and Resilience Plan (PRR), Portuguese Republic, for funding through contract number 2023.15056.TENURE.059.

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense, award number W81XWH-12-2-0012). The ADNI is funded by the National Institute on Aging and the National Institute of Biomedical Imaging and Bioengineering and through generous contributions from the following: AbbVie; Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai, Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche, Ltd., and its affiliated company, Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development, LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer, Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institute of Health Research provides funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org, accessed on 25 January 2026). The grantee organization is the Northern California Institute for Research and Education, and the study was coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. The ADNI data are disseminated by the Laboratory for Neuro-Imaging at the University of Southern California.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.bspc.2026.111143.

Data availability

The data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (http://adni.loni.usc.edu, accessed on 25 january 2026). As such, the investigators within the ADNI contributed to the design and implementation of the work and/or provided data but did not participate in the analysis or writing of this report. A complete list of the ADNI investigators can be found at http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf, (accessed on 25 january 2026)

References

[1] P. Scheltens, B. De Strooper, M. Kivipelto, H. Holstege, G. Chételat, et al., Alzheimer's disease, Lancet 397 (10284) (2021) 1577–1590, http://dx.doi.org/10.1016/s0140-6736(20)32205-4.

[2] J.L. Cummings, Alzheimer disease, JAMA 287 (18) (2002) 2335, http://dx.doi.org/10.1001/jama.287.18.2335.

[3] A. Shukla, R. Tiwari, S. Tiwari, Review on Alzheimer disease detection methods: Automatic pipelines and machine learning techniques, Sci 5 (1) (2023) 13, http://dx.doi.org/10.3390/sci5010013.

[4] D.S. Knopman, H. Amieva, R.C. Petersen, G. Chételat, D.M. Holtzman, et al., Alzheimer disease, Nat. Rev. Dis. Prim. 7 (1) (2021) http://dx.doi.org/10.1038/s41572-021-00269-y.

[5] S.J. Teipel, M. Grothe, S. Lista, N. Toschi, F.G. Garaci, et al., Relevance of magnetic resonance imaging for early detection and diagnosis of Alzheimer disease, Med. Clin. North Am. 97 (3) (2013) 399–424, http://dx.doi.org/10.1016/j.mcna.2012.12.013.

[6] Z. Wang, J. Liu, F. Chen, K. Wang, Y. Xu, et al., Investigating the relationship between brain age and Alzheimer's disease: A deep learning approach with multimodal MRI, Biomed. Signal Process. Control. 109 (2025) 107926, http://dx.doi.org/10.1016/j.bspc.2025.107926.

[7] A. Mora-Rubio, M.A. Bravo-Ortíz, S. Quiñones Arredondo, J.M. Saborit Torres, G.A. Ruz, et al., Classification of Alzheimer's disease stages from magnetic resonance images using deep learning, PeerJ Comput. Sci. 9 (2023) e1490, http://dx.doi.org/10.7717/peerj-cs.1490.

[8] C.R. Jack, D.A. Bennett, K. Blennow, M.C. Carrillo, B. Dunn, et al., NIA-AA research framework: Toward a biological definition of Alzheimer's disease, Alzheimer's Dement. 14 (4) (2018) 535–562, http://dx.doi.org/10.1016/j.jalz.2018.02.018.

[9] A.M. Kiselica, Empirically defining the preclinical stages of the Alzheimer's continuum in the Alzheimer's disease neuroimaging initiative, Psychogeriatrics 21 (4) (2021) 491–502, http://dx.doi.org/10.1111/psyg.12697.

[10] R.C. Petersen, P. Aisen, B.F. Boeve, Y.E. Geda, R.J. Ivnik, et al., Mild cognitive impairment due to Alzheimer disease in the community, Ann. Neurol. 74 (2) (2013) 199–208, http://dx.doi.org/10.1002/ana.23931.

[11] A.A. Tahami Monfared, M.J. Byrnes, L.A. White, Q. Zhang, Alzheimer's disease: Epidemiology and clinical progression, Neurol. Ther. 11 (2) (2022) 553–569, http://dx.doi.org/10.1007/s40120-022-00338-8.

[12] P.M. Rodrigues, D. Freitas, J.P. Teixeira, B. Bispo, D. Alves, et al., Electroencephalogram hybrid method for Alzheimer early detection, Procedia Comput. Sci. 138 (2018) 209–214, http://dx.doi.org/10.1016/j.procs.2018.10.030.

[13] P.M. Rodrigues, D. Freitas, J.P. Teixeira, Alzheimer electroencephalogram temporal events detection by K-means, Procedia Technol. 5 (2012) 859–864, http://dx.doi.org/10.1016/j.protcy.2012.09.095.
[14] L.G. Apostolova, Alzheimer disease, Continuum 22 (2) (2016) 419–434, http://dx.doi.org/10.1212/con.0000000000000307.
[15] S. Balsis, J.F. Benge, D.A. Lowe, L. Geraci, R.S. Doody, How do scores on the ADAS-Cog, MMSE, and CDR-SOB correspond? Clin. Neuropsychol. 29 (7) (2015) 1002–1009, http://dx.doi.org/10.1080/13854046.2015.1119312.
[16] I. Arevalo-Rodriguez, N. Smailagic, M. Roqué-Figuls, A. Ciapponi, E. Sanchez-Perez, et al., Mini-mental state examination (MMSE) for the early detection of dementia in people with mild cognitive impairment (MCI), Cochrane Database Syst. Rev. 2021 (7) (2021) http://dx.doi.org/10.1002/14651858.cd010783.pub3.
[17] S.E. O'Bryant, Staging dementia using clinical dementia rating scale sum of boxes scores: A Texas Alzheimer's research consortium study, Arch. Neurol. 65 (8) (2008) 1091, http://dx.doi.org/10.1001/archneur.65.8.1091.
[18] X. Wang, F. Li, J. Tian, Q. Gao, H. Zhu, Bayesian estimation for the accuracy of three neuropsychological tests in detecting Alzheimer's disease and mild cognitive impairment: a retrospective analysis of the ADNI database, Eur. J. Med. Res. 28 (1) (2023) http://dx.doi.org/10.1186/s40001-023-01265-6.
[19] A.C. van Harten, H.J. Wiste, S.D. Weigand, M.M. Mielke, W.K. Kremers, et al., Detection of Alzheimer's disease amyloid beta 1-42, p-tau, and t-tau assays, Alzheimer's Dement. 18 (4) (2021) 635–644, http://dx.doi.org/10.1002/alz.12406.
[20] P.M. Rodrigues, J.P. Teixeira, D.R.S. Freitas, Alzheimer's electroencephalogram event scalp and source localization, in: HandBook of Research on Trends in the Diagnosis and Treatment of Chronic Conditions, IGI Global, 2016, pp. 33–49, http://dx.doi.org/10.4018/978-1-4666-8828-5.ch002.
[21] G.B. Frisoni, N.C. Fox, C.R. Jack, P. Scheltens, P.M. Thompson, The clinical use of structural MRI in Alzheimer disease, Nat. Rev. Neurol. 6 (2) (2010) 67–77, http://dx.doi.org/10.1038/nrneurol.2009.215.
[22] A. Nordberg, Towards early diagnosis in Alzheimer disease, Nat. Rev. Neurol. 11 (2) (2015) 69–70, http://dx.doi.org/10.1038/nrneurol.2014.257.
[23] J. Xin, A. Wang, R. Guo, W. Liu, X. Tang, CNN and swin-transformer based efficient model for Alzheimer's disease diagnosis with sMRI, Biomed. Signal Process. Control. 86 (2023) 105189, http://dx.doi.org/10.1016/j.bspc.2023.105189.
[24] G. Castellano, A. Esposito, E. Lella, G. Montanaro, G. Vessio, Automated detection of Alzheimer's disease: a multi-modal approach with 3D MRI and amyloid PET, Sci. Rep. 14 (1) (2024) http://dx.doi.org/10.1038/s41598-024-56001-9.
[25] H. Okazawa, M. Nogami, S. Ishida, A. Makino, T. Mori, et al., PET/MRI multimodality imaging to evaluate changes in glymphatic system function and biomarkers of Alzheimer's disease, Sci. Rep. 14 (1) (2024) http://dx.doi.org/10.1038/s41598-024-62806-5.
[26] P.M. Rodrigues, J.P. Madeiro, J.A.L. Marques, Enhancing health and public health through machine learning: Decision support for smarter choices, Bioengineering 10 (7) (2023) 792, http://dx.doi.org/10.3390/bioengineering10070792.
[27] A.S. Alatrany, W. Khan, A. Hussain, H. Kolivand, D. Al-Jumeily, An explainable machine learning approach for Alzheimer's disease classification, Sci. Rep. 14 (1) (2024) http://dx.doi.org/10.1038/s41598-024-51985-w.
[28] S. Salunkhe, M. Bachute, S. Gite, N. Vyas, S. Khanna, et al., Classification of Alzheimer's disease patients using texture analysis and machine learning, Appl. Syst. Innov. 4 (3) (2021) 49, http://dx.doi.org/10.3390/asi4030049.
[29] M.J. Oliveira, P. Ribeiro, P.M. Rodrigues, Machine learning-driven GLCM analysis of structural MRI for Alzheimer's disease diagnosis, Bioengineering 11 (11) (2024) 1153, http://dx.doi.org/10.3390/bioengineering11111153.
[30] T. Altaf, S.M. Anwar, N. Gul, M.N. Majeed, M. Majid, Multi-class Alzheimer's disease classification using image and clinical features, Biomed. Signal Process. Control. 43 (2018) 64–74, http://dx.doi.org/10.1016/j.bspc.2018.02.019.
[31] I. Beheshti, N. Maikusa, H. Matsuda, H. Demirel, G. Anbarjafari, Histogram-based feature extraction from individual gray matter similarity-matrix for Alzheimer's disease classification, J. Alzheimer's Dis. 55 (4) (2016) 1571–1582, http://dx.doi.org/10.3233/jad-160850.
[32] S.R. R.S., B. M.C., N. V.N., V. Thomas, Classifying Alzheimer's disease using machine learning: Insights from default mode network alterations, Biomed. Signal Process. Control. 112 (2026) 108526, http://dx.doi.org/10.1016/j.bspc.2025.108526.
[33] T. Jitsuishi, A. Yamaguchi, Searching for optimal machine learning model to classify mild cognitive impairment (MCI) subtypes using multimodal MRI data, Sci. Rep. 12 (1) (2022) http://dx.doi.org/10.1038/s41598-022-08231-y.
[34] S. Toshkhujaev, K.H. Lee, K.Y. Choi, J.J. Lee, G.-R. Kwon, et al., Classification of Alzheimer's disease and mild cognitive impairment based on cortical and subcortical features from MRI T1 brain images utilizing four different types of datasets, J. Heal. Eng. 2020 (2020) 1–14, http://dx.doi.org/10.1155/2020/3743171.
[35] L. Jin, K. Zhao, Y. Zhao, T. Che, S. Li, A hybrid deep learning method for early and late mild cognitive impairment diagnosis with incomplete multimodal data, Front. Neuroinformatics 16 (2022) http://dx.doi.org/10.3389/fninf.2022.843566.
[36] X. Tang, J. Liu, Comparing different algorithms for the course of Alzheimer's disease using machine learning, Ann. Palliat. Med. 10 (9) (2021) 9715–9724, http://dx.doi.org/10.21037/apm-21-2013.
[37] M.S.K. Inan, N.S. Sworna, A.M. Islam, S. Islam, Z. Alom, et al., A slice selection guided deep integrated pipeline for Alzheimer's prediction from structural brain MRI, Biomed. Signal Process. Control. 89 (2024) 105773, http://dx.doi.org/10.1016/j.bspc.2023.105773.
[38] L. Wang, J. Sheng, Q. Zhang, R. Zhou, Z. Li, et al., Functional brain network measures for Alzheimer's disease classification, IEEE Access 11 (2023) 111832–111845, http://dx.doi.org/10.1109/access.2023.3323250.
[39] J. Venugopalan, L. Tong, H.R. Hassanzadeh, M.D. Wang, Multimodal deep learning models for early detection of Alzheimer's disease stage, Sci. Rep. 11 (1) (2021) http://dx.doi.org/10.1038/s41598-020-74399-w.
[40] Q. Zhang, H. Cai, Y. Long, S. Yu, Multicentric prototype and pseudo-labeling based source-free domain adaptation for Alzheimer's disease classification, Biomed. Signal Process. Control. 103 (2025) 107483, http://dx.doi.org/10.1016/j.bspc.2024.107483.
[41] E. Alickovic, A. Subasi, Automatic detection of Alzheimer disease based on histogram and random forest, in: CMBEBIH 2019, Springer International Publishing, 2019, pp. 91–96, http://dx.doi.org/10.1007/978-3-030-17971-7_14.
[42] S. Palmqvist, P. Tideman, N. Cullen, H. Zetterberg, K. Blennow, et al., Prediction of future Alzheimer's disease dementia using plasma phospho-tau combined with other accessible measures, Nature Med. 27 (6) (2021) 1034–1042, http://dx.doi.org/10.1038/s41591-021-01348-z.
[43] S. Palmqvist, P. Tideman, N. Mattsson-Carlgren, S.E. Schindler, R. Smith, et al., Blood biomarkers to detect Alzheimer disease in primary care and secondary care, JAMA 332 (15) (2024) 1245, http://dx.doi.org/10.1001/jama.2024.13855.
[44] Y. Gupta, R.K. Lama, G.-R. Kwon, Prediction and classification of Alzheimer's disease based on combined features from apolipoprotein-E genotype, cerebrospinal fluid, MR, and FDG-PET imaging biomarkers, Front. Comput. Neurosci. 13 (2019) http://dx.doi.org/10.3389/fncom.2019.00072.
[45] F. Pan, Y. Huang, X. Cai, Y. Wang, Y. Guan, et al., Integrated algorithm combining plasma biomarkers and cognitive assessments accurately predicts brain β-amyloid pathology, Commun. Med. 3 (1) (2023) http://dx.doi.org/10.1038/s43856-023-00295-9.
[46] B.T. Nallapu, K.K. Petersen, R.B. Lipton, C. Davatzikos, A. Ezzati, Plasma biomarkers as predictors of progression to dementia in individuals with mild cognitive impairment, J. Alzheimer's Dis. 98 (1) (2024) 231–246, http://dx.doi.org/10.3233/jad-230620.
[47] J.P. Kim, J. Kim, H. Jang, J. Kim, S.H. Kang, et al., Predicting amyloid positivity in patients with mild cognitive impairment using a radiomics approach, Sci. Rep. 11 (1) (2021) http://dx.doi.org/10.1038/s41598-021-86114-4.
[48] S. Kim, P. Lee, K.T. Oh, M.S. Byun, D. Yi, et al., Deep learning-based amyloid PET positivity classification model in the Alzheimer's disease continuum by using 2-[18F]FDG PET, EJNMMI Res. 11 (1) (2021) http://dx.doi.org/10.1186/s13550-021-00798-3.
[49] Y. Lin, H.S. Kwon, W. Lee, M. Hwang, J.H. Jeong, et al., Cross-cultural validation of plasma p-tau217 and p-tau181 as precision biomarkers for amyloid PET positivity: An East Asian study in Taiwan and Korea, Alzheimer's Dement. 21 (1) (2025) http://dx.doi.org/10.1002/alz.14565.
[50] M.R. Meyer, K.M. Kirmess, S. Eastwood, T.L. Wente-Roth, F. Irvin, et al., Clinical validation of the precivityad2 blood test: A mass spectrometry-based test with algorithm combining p-tau217 and aβ42/40 ratio to identify presence of brain amyloid, Alzheimer's & Dement. 20 (5) (2024) 3179–3192, http://dx.doi.org/10.1002/alz.13764.
[51] R.A. Rissman, O. Langford, R. Raman, M.C. Donohue, S. Abdel-Latif, et al., Plasma aβ42/aβ40 and phospho-tau217 concentration ratios increase the accuracy of amyloid PET classification in preclinical Alzheimer's disease, Alzheimer's Dement. 20 (2) (2023) 1214–1224, http://dx.doi.org/10.1002/alz.13542.
[52] S. Dutt, Y. Li, M. Mather, D.A. Nation, Brainstem volumetric integrity in preclinical and prodromal Alzheimer's disease, J. Alzheimer's Dis. 77 (4) (2020) 1579–1594, http://dx.doi.org/10.3233/jad-200187.
[53] A. Marchewka, F. Kherif, G. Krueger, A. Grabowska, R. Frackowiak, et al., Influence of magnetic field strength and image registration strategy on voxel-based morphometry in a study of Alzheimer's disease, Hum. Brain Mapp. 35 (5) (2013) 1865–1874, http://dx.doi.org/10.1002/hbm.22297.
[54] C.R. Jack, J. Barnes, M.A. Bernstein, B.J. Borowski, J. Brewer, et al., Magnetic resonance imaging in Alzheimer's disease neuroimaging initiative 2, Alzheimer's Dement. 11 (7) (2015) 740–756, http://dx.doi.org/10.1016/j.jalz.2015.05.002.
[55] ADNI, Alzheimer's disease neuroimaging initiative data sharing and publication policy, 2024, available online: https://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_DSP_Policy.pdf. (Accessed on 12 March 2025).
[56] W. Haynes, Bonferroni correction, in: Encyclopedia of Systems Biology, Springer New York, 2013, http://dx.doi.org/10.1007/978-1-4419-9863-7_1213, 154–154.
[57] M.R. Campbell, S. Ashrafzadeh-Kian, R.C. Petersen, M.M. Mielke, J.A. Syrjanen, et al., P-tau/aβ42 and aβ42/40 ratios in CSF are equally predictive of amyloid PET status, Alzheimer's Dement.: Diagn. Assess. Dis. Monit. 13 (1) (2021) http://dx.doi.org/10.1002/dad2.12190.

[58] O. Hansson, J. Seibyl, E. Stomrud, H. Zetterberg, J.Q. Trojanowski, et al., Csf biomarkers of Alzheimer's disease concord with amyloid-β PET and predict clinical progression: A study of fully automated immunoassays in BioFINDER and ADNI cohorts, Alzheimer's Dement. 14 (11) (2018) 1470–1481, http://dx.doi.org/10.1016/j.jalz.2018.01.010.

[59] J. Skinner, J.O. Carvalho, G.G. Potter, A. Thames, E. Zelinski, et al., The Alzheimer's disease assessment scale-cognitive-plus (ADAS-Cog-Plus): an expansion of the ADAS-Cog to improve responsiveness in MCI, Brain Imaging Behav. 6 (4) (2012) 489–501, http://dx.doi.org/10.1007/s11682-012-9166-3.

[60] J.K. Kueper, M. Speechley, M. Montero-Odasso, The Alzheimer's disease assessment scale–cognitive subscale (adas-cog): Modifications and responsiveness in pre-dementia populations. a narrative review, J. Alzheimer's Dis. 63 (2) (2018) 423–444, http://dx.doi.org/10.3233/jad-170991.

[61] N.J. Tustison, P.A. Cook, A.J. Holbrook, H.J. Johnson, J. Muschelli, et al., The ANTsX ecosystem for quantitative biological and medical imaging, Sci. Rep. 11 (1) (2021) http://dx.doi.org/10.1038/s41598-021-87564-6.

[62] B. Avants, C. Epstein, M. Grossman, J. Gee, Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain, Med. Image Anal. 12 (1) (2008) 26–41, http://dx.doi.org/10.1016/j.media.2007.06.004.

[63] M. Jenkinson, C.F. Beckmann, T.E. Behrens, M.W. Woolrich, S.M. Smith, FSL, NeuroImage 62 (2) (2012) 782–790, http://dx.doi.org/10.1016/j.neuroimage.2011.09.015.

[64] B.B. Avants, N.J. Tustison, G. Song, P.A. Cook, A. Klein, et al., A reproducible evaluation of ANTs similarity metric performance in brain image registration, NeuroImage 54 (3) (2011) 2033–2044, http://dx.doi.org/10.1016/j.neuroimage.2010.09.025.

[65] A. Klein, J. Andersson, B.A. Ardekani, J. Ashburner, B. Avants, et al., Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration, NeuroImage 46 (3) (2009) 786–802, http://dx.doi.org/10.1016/j.neuroimage.2008.12.037.

[66] K.H. Zou, S.K. Warfield, A. Bharatha, C.M.C. Tempany, M.R. Kaus, et al., Statistical validation of image segmentation quality based on a spatial overlap index, Academic Radiol. 11 (2) (2004) 178–189, http://dx.doi.org/10.1016/S1076-6332(03)00671-8.

[67] A.P. Zijdenbos, B.M. Dawant, R.A. Margolin, A.C. Palmer, Morphometric analysis of white matter lesions in MR images: Method and validation, IEEE Trans. Med. Imaging 13 (4) (1994) 716–724, http://dx.doi.org/10.1109/42.363096.

[68] D.W. Shattuck, S.R. Sandor-Leahy, K.A. Schaper, D.A. Rottenberg, R.M. Leahy, Magnetic resonance image tissue classification using a partial volume model, NeuroImage 13 (5) (2001) 856–876, http://dx.doi.org/10.1006/nimg.2000.0730.

[69] S. Vega-Guzman, G. Ramirez-Nava, I. Chairez, M. Alfaro-Ponce, Optimizing Alzheimer's diagnosis: Multimodal FDG-PET and MRI feature fusion with explainable deep learning, Biomed. Signal Process. Control. 117 (2026) 109659, http://dx.doi.org/10.1016/j.bspc.2026.109659.

[70] B.B. Avants, N. Tustison, G. Song, Advanced normalization tools (ANTS), Insight J. 2 (2009) 1–35, http://dx.doi.org/10.54294/uvnhin.

[71] R.T. Shinohara, E.M. Sweeney, J. Goldsmith, N. Shiee, F.J. Mateen, et al., Statistical normalization techniques for magnetic resonance imaging, NeuroImage: Clin. 6 (2014) 9–19, http://dx.doi.org/10.1016/j.nicl.2014.08.008.

[72] R.C. Gonzalez, R.E. Woods, Digital Image Processing, fourth ed., Pearson, New York, 2018.

[73] S. Van der Walt, J.L. Schönberger, J. Nunez-Iglesias, F. Boulogne, J.D. Warner, et al., Scikit-image: image processing in Python, PeerJ 2 (2014) e453, http://dx.doi.org/10.7717/peerj.453.

[74] S. Venkat, T. Ghodeswar, P. Chavan, S.K. Narayanasamy, K. Srinivasan, MRI-based automated diagnosis of Alzheimer's disease using Alzh-Net deep learning model, Biomed. Signal Process. Control. 102 (2025) 107367, http://dx.doi.org/10.1016/j.bspc.2024.107367.

[75] J.C. Mazziotta, A.W. Toga, A. Evans, P. Fox, J. Lancaster, A probabilistic atlas of the human brain: Theory and rationale for its development, NeuroImage 2 (2) (1995) 89–101, http://dx.doi.org/10.1006/nimg.1995.1012.

[76] X. Han, R. Kwitt, S. Aylward, S. Bakas, B. Menze, et al., Brain extraction from normal and pathological images: A joint PCA/image-reconstruction approach, NeuroImage 176 (2018) 431–445, http://dx.doi.org/10.1016/j.neuroimage.2018.04.073.

[77] E. Bullmore, J. Fadili, V. Maxim, L. Şendur, B. Whitcher, et al., Wavelets and functional magnetic resonance imaging of the human brain, NeuroImage 23 (2004) S234–S249, http://dx.doi.org/10.1016/j.neuroimage.2004.07.012.

[78] K. Hackmack, F. Paul, M. Weygandt, C. Allefeld, J.-D. Haynes, Multi-scale classification of disease using structural MRI and wavelet transform, NeuroImage 62 (1) (2012) 48–58, http://dx.doi.org/10.1016/j.neuroimage.2012.05.022.

[79] M. Fayaz, N. Torokeldiev, S. Turdumamatov, M.S. Qureshi, M.B. Qureshi, et al., An efficient methodology for brain MRI classification based on DWT and convolutional neural network, Sensors 21 (22) (2021) 7480, http://dx.doi.org/10.3390/s21227480.

[80] G. Othman, D.Q. Zeebaree, The applications of discrete wavelet transform in image processing: A review, J. Soft Comput. Data Min. 1 (2) (2020) 31–43, http://dx.doi.org/10.30880/jscdm.2020.01.02.004.

[81] G.P. Nason, B.W. Silverman, The discrete wavelet transform in S, J. Comput. Graph. Statist. 3 (2) (1994) 163–191, http://dx.doi.org/10.2307/1390667.

[82] J. Silva, B.C. Bispo, P.M. Rodrigues, Structural MRI texture analysis for detecting Alzheimer's disease, J. Med. Biological Eng. 43 (3) (2023) 227–238, http://dx.doi.org/10.1007/s40846-023-00787-y.

[83] A. Dede, H. Nunoo-Mensah, E.K. Akowuah, K.O. Boateng, P.E. Adjei, et al., Wavelet-based feature extraction for efficient high-resolution image classification, Eng. Rep. 7 (2) (2025) http://dx.doi.org/10.1002/eng2.70027.

[84] A. Ramola, A.K. Shakya, D. Van Pham, Study of statistical methods for texture analysis and their modern evolutions, Eng. Rep. 2 (4) (2020) http://dx.doi.org/10.1002/eng2.12149.

[85] T. Ojala, M. Pietikäinen, Texture classification, CVonline: The evolving, distributed, non-proprietary, on-line compendium of computer vision, 2001.

[86] N. Pasnoori, T. Flores-Garcia, B.D. Barkana, Histogram-based features track Alzheimer's progression in brain MRI, Sci. Rep. 14 (1) (2024) http://dx.doi.org/10.1038/s41598-023-50631-1.

[87] Q. Feng, Z. Ding, MRI radiomics classification and prediction in Alzheimer's disease and mild cognitive impairment: A review, Curr. Alzheimer Res. 17 (3) (2020) 297–309, http://dx.doi.org/10.2174/1567205017666200303105016.

[88] P. Mohanaiah, P. Sathyanarayana, L. GuruKumar, Image texture feature extraction using GLCM approach, Int. J. Sci. Res. Publ. 3 (5) (2013) 1–5.

[89] S. Zheng, H. Wang, F. Han, J. Chu, F. Zhang, et al., Detection of microstructural medial prefrontal cortex changes using magnetic resonance imaging texture analysis in a post-traumatic stress disorder rat model, Front. Psychiatry 13 (2022) http://dx.doi.org/10.3389/fpsyt.2022.805851.

[90] S.M. Vijithananda, M.L. Jayatilake, T.C. Gonçalves, L.M. Rato, B.S. Weerakoon, et al., Texture feature analysis of MRI-ADC images to differentiate glioma grades using machine learning techniques, Sci. Rep. 13 (1) (2023) http://dx.doi.org/10.1038/s41598-023-41353-5.

[91] M. Kuhn, K. Johnson, Feature Engineering and Selection: A Practical Approach for Predictive Models, Chapman and Hall/CRC, 2019, http://dx.doi.org/10.1201/9781315108230.

[92] A. Fernández, S. García, M. Galar, R.C. Prati, B. Krawczyk, et al., Learning from imbalanced data sets, vol. 10, Springer, 2018, http://dx.doi.org/10.1007/978-3-319-98074-4.

[93] A. Demircioğlu, Applying oversampling before cross-validation will lead to high bias in radiomics, Sci. Rep. 14 (1) (2024) 11563, http://dx.doi.org/10.1038/s41598-024-62585-z.

[94] P. Ribeiro, J. Sá, D. Paiva, P.M. Rodrigues, Cardiovascular diseases diagnosis using an ECG multi-band non-linear machine learning framework analysis, Bioengineering 11 (1) (2024) 58, http://dx.doi.org/10.3390/bioengineering11010058.

[95] Z.D. Vujovic, Classification model evaluation metrics, Int. J. Adv. Comput. Sci. Appl. 12 (6) (2021) http://dx.doi.org/10.14569/IJACSA.2021.0120670.

[96] L.S. Gill, J. Kaur, N. Goel, Machine learning and texture features based approach for classifying Alzheimer's disease, Procedia Comput. Sci. 235 (2024) 2741–2748, http://dx.doi.org/10.1016/j.procs.2024.04.258.

[97] N. Aggarwal, R.K. Agrawal, First and second order statistics features for classification of magnetic resonance brain images, J. Signal Inf. Process. 03 (02) (2012) 146–153, http://dx.doi.org/10.4236/jsip.2012.32019.

[98] Y. Varatharajah, V.K. Ramanan, R. Iyer, P. Vemuri, M.W. Weiner, et al., Predicting short-term MCI-to-AD progression using imaging, CSF, genetic factors, cognitive resilience, and demographics, Sci. Rep. 9 (1) (2019) http://dx.doi.org/10.1038/s41598-019-38793-3.

[99] B.A.C. Ramalho, L.R. Bortolato, N.D. Gomes, L. Wichert-Ana, F.E. Padovan-Neto, et al., The impact of the orientation of MRI slices on the accuracy of Alzheimer's disease classification using convolutional neural networks (CNNs), J. Med. Artif. Intell. 7 (2024) http://dx.doi.org/10.21037/jmai-24-51, 35–35.

[100] S. Lahmiri, M. Boukadoum, Classification of brain MRI using the LH and HL wavelet transform sub-bands, in: 2011 IEEE International Symposium of Circuits and Systems, ISCAS, IEEE, 2011, pp. 1025–1028, http://dx.doi.org/10.1109/iscas.2011.5937743.