



Multimodal prediction of MCI-to-dementia conversion over a two-year window

Bipul Simkhada · Thony Yan Liang · Xueting Cui · Micheal Adeyosoye · Bibhu Simkhada · Mercedes Cabrerizo · Rosie Curiel Cid · Shanna L. Burke · Armando Barreto · Naphtali Rishe · David A. Loewenstein · Malek Adjouadi

Received: 8 June 2026 / Accepted: 11 September 2026
© The Author(s), under exclusive licence to American Aging Association 2026

Abstract Mild cognitive impairment (MCI) represents a heterogeneous clinical state where some individuals remain stable while others progress to dementia. Identification of patients who are at high risk for near-term conversion remains a critical challenge for timely intervention. We developed an explainable multimodal machine-learning framework to predict progression from MCI to dementia within a clinically meaningful two-year period. We analyzed 2008 samples from 828 unique MCI subjects from the Alzheimer's Disease Neuroimaging Initiative (ADNI),

including cognitive/functional assessments, demographic/genetic variables, and structural MRI features. Subjects were classified as stable MCI (sMCI, $n = 1306$) or progressive MCI (pMCI, $n = 702$) based on two-year outcomes. Machine learning algorithms were evaluated on unimodal and multimodal using nested cross-validation, and explainability was assessed using SHapley Additive exPlanations (SHAP) and permutation feature importance. The combination of cognitive assessments, demographics/risk factors and structural MRI features outperformed individual modalities, highlighting the complementary relationship between cognitive-functional decline and neuroanatomical degeneration.

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1007/s11357-026-02551-x>.

B. Simkhada (✉) · T. Y. Liang · X. Cui · M. Adeyosoye · M. Cabrerizo · A. Barreto · N. Rishe · M. Adjouadi
Center for Advanced Technology and Education,
Department of Electrical and Computer Engineering,
Florida International University, Miami, FL, USA
e-mail: bipulsimkhada7@gmail.com

B. Simkhada
Institute for Human Genetics, Clemson University,
Greenwood, SC, USA

R. C. Cid · S. L. Burke · D. A. Loewenstein · M. Adjouadi
1Florida Alzheimer's Disease Research Center, University
of Florida, Gainesville, FL, USA

R. C. Cid · D. A. Loewenstein
Departments of Psychiatry and Behavioral Sciences,
University of Miami Miller School of Medicine, Miami,
FL, USA

S. L. Burke
Robert Stempel College of Public Health and Social Work,
School of Social Work, Florida International University,
Miami, FL, USA

N. Rishe
Knight Foundation School of Computing and Information
Sciences, Florida International University, Miami, FL,
USA

D. A. Loewenstein
Center for Cognitive Neuroscience and Aging, University
of Miami Miller School of Medicine, Miami, FL, USA

The best-performing model, a calibrated XGBoost, achieved a balanced accuracy of $81.28\% \pm 2.56\%$ during cross-validation and 81.99% (bootstrap mean 81.97%, 95% CI: [78.22%–85.51%]) on an independent held-out test set. Explainable AI analyses identified functional impairment measures (Functional Activities Questionnaire, Clinical Dementia Rating – Sum of Boxes), cognitive performance scores (Alzheimer’s Disease Assessment Scale – Cognitive Subscale 11-item version, Alzheimer’s Disease Assessment Scale – Cognitive Subscale 13-item version, Mental State Examination), *APOE* $\epsilon 4$ status, and structural abnormalities within the hippocampus, lateral ventricle, parietal and temporal regions, and amygdala, as the most influential predictors of conversion. Short-term progression from MCI to dementia can be predicted with high accuracy using multimodal information obtained from a single clinical visit. The convergence of cognitive impairment and region-specific neurodegeneration emerged as the strongest indicator of conversion risk.

Keywords Mild cognitive impairment · Alzheimer’s disease · Brain aging · Neurodegeneration · Structural MRI · Machine learning · Explainable AI · Dementia conversion

Introduction

Alzheimer’s disease and related dementias (ADRD) collectively refer to neurodegenerative disorders affecting older adults that manifest progressively over a period of 3 to 20 years. ADRD affects approximately 1 in 9 adults over the age of 65, impacting over 7 million people in the United States alone [1]. Early pathological alterations include the accumulation of amyloid-beta plaques and neurofibrillary tangles composed of hyperphosphorylated tau protein, alongside synaptic dysfunction and localized neuroinflammation [2–4]. These pathological changes progress to neurodegeneration and severe cognitive impairments in later stages. Understanding why some individuals remain cognitively resilient despite underlying neuropathology whereas others experience accelerated decline is a central question in the biology of aging and neurodegeneration. Consequently, identifying factors associated with progression from prodromal cognitive

impairment to dementia has become increasingly important for precision medicine, therapeutic intervention, and research on healthy brain aging.

The intermediate clinical stage between normal cognitive aging and dementia is often referred to as mild cognitive impairment (MCI). It is characterized by measurable cognitive decline—particularly in memory—that exceeds what is expected from normal aging but does not yet meet the diagnostic criteria for dementia [5]. Numerous longitudinal studies have shown that individuals diagnosed with MCI have a significantly elevated risk for developing dementia. A systematic review of 41 cohort studies, conducted in a specialist clinical settings, reported progression of approximately 40% of MCI patients to dementia or related conditions [6]. Furthermore, the annual conversion rate from MCI to dementia is estimated to be approximately 10% [7, 8]. Other long-term studies indicate that 80–90% of patients diagnosed with MCI progress to dementia within six years [7, 9].

Given the substantially higher risk of progression among individuals with MCI, accurately predicting whether individuals with MCI will progress to dementia has become an increasingly active area of investigation. However, MCI is highly heterogeneous with highly variable clinical trajectories, i.e., patients may progress to dementia, remain stable, or even revert to normal cognition [5, 10–12]. This heterogeneity likely reflects differences in underlying biological resilience, rates of neurodegeneration, and compensatory mechanisms that influence disease progression. As a result, reliable prediction of conversion risk remains one of the most challenging problems in dementia research. Therefore, advanced machine learning (ML) and deep learning (DL) techniques with high dimensional multi-modal feature spaces have become increasingly important to capture complex and nonlinear patterns [13, 14].

Existing ML and DL approaches to MCI progression prediction can be broadly categorized into two groups: (1) longitudinal modeling methods and (2) baseline-driven methods. Longitudinal approaches rely on data collected across multiple clinical visits to predict disease progression over time horizons typically ranging from one to five years. Recurrent neural networks (RNNs), particularly Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU) networks [15–18], and more recently,

transformer-based architectures [19–21], have been used to model temporal dependencies in patient trajectories.

In contrast, baseline-driven approaches aim to predict disease progression using baseline data from a single time point. Traditional ML methods (logistic regression, random forest, support vector machine, XGBoost) remain widely used in this context [22–26]. Simultaneously, other studies have explored neuroimaging-based deep learning techniques. For example, Abrol et al. proposed a deep residual learning framework utilizing gray matter features while Xiao et al. employed sparse logistic regression with elastic net regularization on region-of-interest (ROI) features extracted from MRI data [27, 28]. More recently, transformer-based and attention-driven models have been investigated. These include a class-activation attention approach by Luo et al., a vision transformer approach by Hoang et al., and a multi-scale attention convolution framework combined with an aging transformer by Gao et al. [29–31].

Despite these advances, an important methodological and clinical gap remains. Longitudinal models require multi-visit historical data and therefore are not directly applicable when only a single clinical visit is available. Conversely, baseline-driven approaches can provide predictions from a single assessment but typically define progression outcomes based on variable interindividual follow-up durations. While useful for long-term prognosis, this introduces inconsistencies in outcome definitions, reduces data comparability, biases model development, and limits generalizability across cohorts.

To address these gaps, we define progression using a fixed two-year prediction window and perform prediction using information available at a single clinical visit. A two-year horizon enables the identification of individuals at elevated risk sufficiently early to facilitate patient counseling, care planning, timely therapeutic intervention, and recruitment into disease-modifying clinical trials [10, 32]. Given the annual conversion rate of approximately 10%, a two-year window captures a substantial proportion of individuals that are likely to progress.

The proposed framework integrates structural MRI-derived features, cognitive and functional assessments, demographic variables, and genetic risk factors to classify individuals into two categories, stable MCI (sMCI) and progressive MCI (pMCI).

We employ multiple ML strategies, including logistic regression (LR), random forest (RF), support vector classifier (SVC), and extreme gradient boosting (XGBoost; XGB) across different sampling strategies and modality configurations. Beyond predictive performance, we incorporate explainable artificial intelligence (XAI) techniques to evaluate the drivers of disease progression and identify key predictive biomarkers.

The key contributions of this work can be summarized as follows: (1) a clinically grounded multimodal predictive framework for MCI progression within a fixed two-year window; (2) a dynamic single-visit prediction strategy that allows risk estimation from any clinical visit, thereby improving real-world applicability; (3) a nested cross-validation training strategy with explainable AI methods to identify the most suitable model and predictive features associated with MCI conversion.

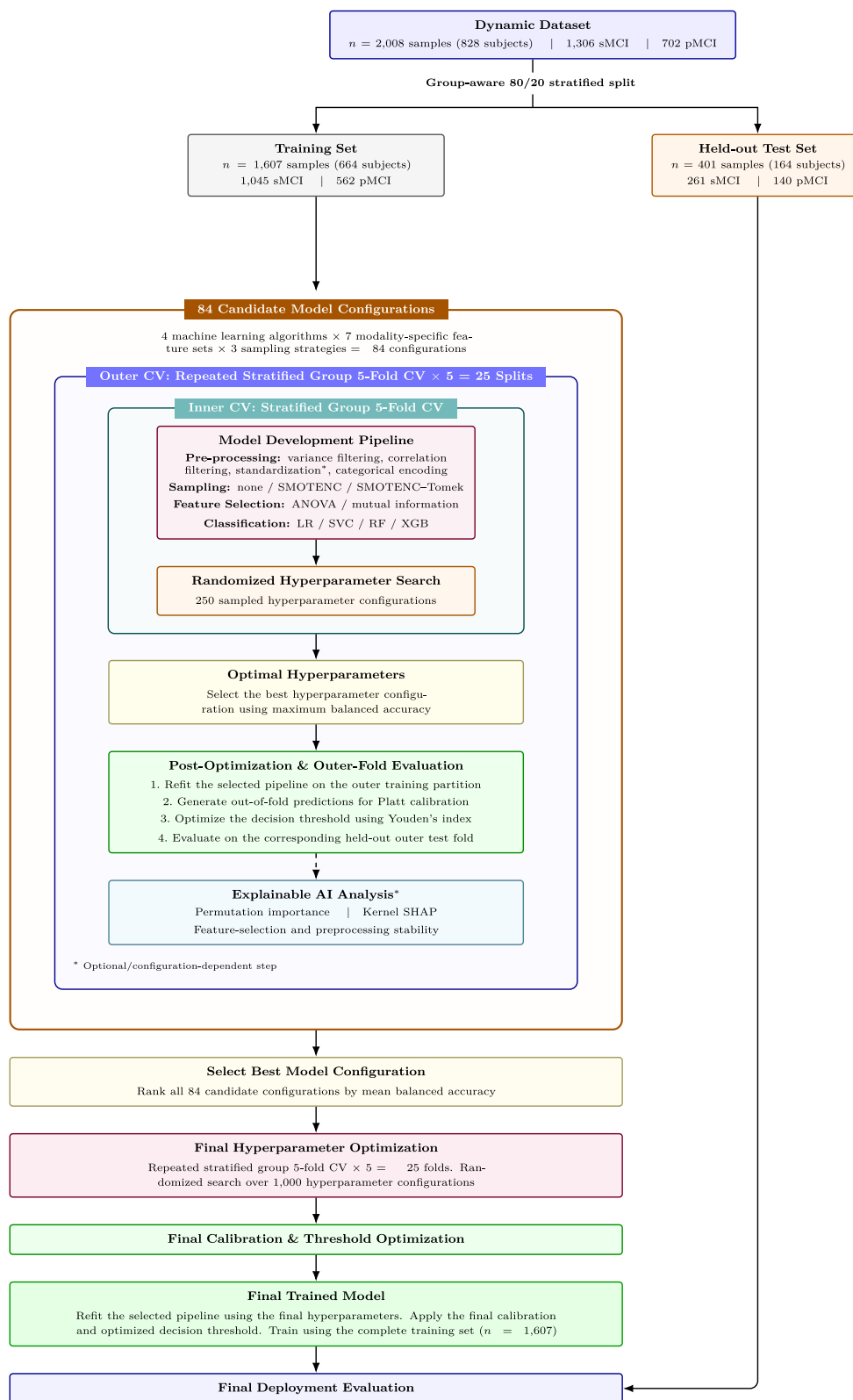
Method

Dataset

A multimodal dataset was curated using neuropsychological assessments, demographic and genetic information, and structural MRI-derived features from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database [<https://adni.loni.usc.edu/>]. Each sample in the ADNI database corresponds to a visit-level observation, and therefore, subjects were allowed to contribute to multiple samples provided that the visits were at least six months apart.

The neuropsychological modality (COG) consisted of six cognitive and functional assessments: Mini-Mental State Examination (MMSE), Clinical Dementia Rating – Global Score (CDR Global), Clinical Dementia Rating – Sum of Boxes (CDR-SB), Alzheimer's Disease Assessment Scale – Cognitive Subscale 11-item version (ADAS11), Alzheimer's Disease Assessment Scale – Cognitive Subscale 13-item version (ADAS13), and Functional Activities Questionnaire (FAQ). The demographic and genetic modality (DEM) included five features: gender, age, handedness, years of education (PTEDUCAT), and *APOE* ϵ 4 (*APOE*4).

The MRI modality (MRI) consisted of 323 structural features pre-extracted and provided directly by



◀**Fig. 1** Overview of the proposed experimental pipeline for mild cognitive impairment (MCI) progression prediction. The workflow details subject-level stratified data splitting (80/20), nested cross-validation setup (inner and outer) across 84 candidate model configurations, hyperparameter search, post-optimization calibration, decision-threshold tuning, and explainable AI (XAI) analysis. The final model is trained on the full training set ($n=1607$) and evaluated on the held-out test set ($n=401$). Asterisks (*) indicate optional/configuration-dependent pipeline steps

ADNI (Table S1). These features were grouped into five morphometric categories: subcortical volume (SV; 48 features), cortical volume (CV; 69 features), cortical surface area (SA; 70 features), cortical thickness average (CT; 68 features), and cortical thickness standard deviation (CTS; 68 features). In total, the constructed dataset consisted of 334 features across the three modalities (COG, DEM and MRI), comprising 332 quantitative features and 2 categorical features (gender and handedness; both encoded in binary form).

Samples with an MCI diagnosis at the observation visit were selected and labeled according to diagnostic progression within a strict two-year follow-up window. As such, samples that progressed to dementia (AD) within the two-year period were labeled progressive MCI (pMCI). Samples that remained MCI throughout the 2-year period were labeled stable MCI (sMCI), regardless of whether they subsequently converted to AD at later time points (> 24 months). To ensure label fidelity, fluctuating or uncertain diagnostic trajectories—including reversion from MCI to cognitively normal status—were excluded. For subjects with multiple visits spaced more than 6 months apart, each visit was evaluated independently. Additionally, all modalities and their features were required to be available at the observation visit; samples with any missing features were excluded.

The final dataset (referred to as the *dynamic dataset*) consisted of 2008 samples from 828 unique subjects; 1306 sMCI samples (577 subjects) and 702 pMCI samples (347 subjects) (Table S2, S3). This resulted in a moderately imbalanced class distribution. Notably, 96 subjects were represented in both classes, as they had at least one visit labeled sMCI and another visit that was subsequently labeled pMCI.

For comparison with conventional baseline-driven approaches, a baseline-only subset (referred to as the *baseline dataset*) was generated by selecting the

initial (baseline) visit for each subject. This comprised 690 samples, including 473 sMCI and 217 pMCI cases (Table S4, S5).

Data partitioning strategy

The dynamic dataset was partitioned into training and testing sets using an 80/20 split with group-aware stratified sampling. Stratification was performed based on the target labels (sMCI and pMCI) to preserve the class distribution. Grouping based on subject identifiers ensured that samples from the same subject were not distributed across splits, thereby preventing data leakage. This partition resulted in 1607 samples from 664 subjects in the training set (1045 sMCI and 562 pMCI) and 401 samples from 164 subjects in the testing set (261 sMCI and 140 pMCI). The training set was used exclusively for model development, whereas the testing set was held out for an unbiased estimate of predictive performance of the final model.

Conversely, the baseline dataset was not partitioned into separate training and testing subsets. Given the relatively smaller sample size and its role as a comparative benchmark, the entire baseline dataset was utilized within a group-aware nested cross-validation framework.

Experimental pipeline

The experimental pipeline (Fig. 1) was designed to enable systematic and reproducible evaluation of multiple modal configurations (feature sets, sampling strategy, ML model). The pipeline was implemented using the scikit-learn library to ensure consistency throughout the model development.

Given input features $X \in \mathbb{R}^{n \times 323}$, preprocessing was first applied to reduce redundancy and normalize feature scales. For quantitative features, a variance threshold (set as a tunable parameter) was applied to remove low-variance features. In addition, correlation-based filtering was performed using a pairwise Pearson correlation matrix, with one feature from each highly correlated pair removed when the correlation coefficient exceeded a predefined threshold. Feature standardization (zero mean and unit variance) was optionally applied to non-tree-based models and was consistently performed within training folds [33, 34].

To address class imbalance (approximately a 2:1 ratio of sMCI to pMCI), data resampling was incorporated as an optional step within the pipeline. The configurations included either no resampling (None), Synthetic Minority Over-sampling Technique for Nominal and Continuous features (SMOTENC), or SMOTENC followed by Tomek Links (SMOTENC-Tomek). SMOTENC accounts for both numerical and categorical features to generate synthetic samples for the minority class [35]. Therefore, it is better suited to heterogeneous clinical datasets because it preserves the integrity of categorical variables during synthetic sample generation. Tomek Links were subsequently applied to remove borderline or ambiguous samples near class boundaries [36].

Following resampling, feature selection was performed for dimensionality reduction. Specifically, *F*-statistics from analysis of variance (ANOVA) and Mutual Information (MI) were used to rank features based on their relevance to the target variable. The number of selected features *K* and the selection method were treated as tunable parameters within the configuration space (Table S6).

Finally, a diverse set of ML algorithms (LR, RF, SVC and XGB) were considered to capture both linear and nonlinear relationships in the data.

Nested cross validation for model selection

We evaluated four ML algorithms, seven feature sets (individual modalities, pairwise combinations, and all modalities), and three sampling strategies (None, SMOTENC, and SMOTENC-Tomek), resulting in 84 unique model configurations (Fig. 1). To prevent potential data leakage among repeated visits from the same subject, all cross-validation partitioning strictly adhered to the group-aware stratified sampling strategy introduced in the “[Data partitioning strategy](#)” section. For each configuration, the outer loop used fivefold cross-validation repeated five times, resulting in 25 outer train-test splits within the training set. Hyperparameter optimization was conducted exclusively within each outer training partition, whereas the held-out outer test folds were reserved for unbiased performance evaluation.

Within each outer training fold, hyperparameter optimization was performed using fivefold cross-validation combined with randomized search. The search space included preprocessing parameters,

feature-selection parameters, and model-specific hyperparameters (Table S6). Hyperparameters were sampled from predefined distributions, including uniform and log-uniform ranges where appropriate. Randomized search was adopted to evaluate 250 configurations for each outer fold due to its computational efficiency for exploring high-dimensional parameter spaces.

The unified experimental pipeline described in the “[Experimental pipeline](#)” section was applied consistently within the inner cross-validation process. Hyperparameter selection was based on balanced accuracy. Balanced accuracy account for class imbalance irrespective of class prevalence by assigning equal weight to sensitivity and specificity. From a clinical perspective, balanced accuracy prevents model selection from being dominated by the majority of sMCI class; thereby helping to ensure reliable identification of individuals who are likely to progress to dementia while maintaining predictive performance for stable cases.

After the optimal hyperparameters were identified, the selected pipeline was refitted within the outer training partition. Probability calibration is particularly critical in medical applications, where well-calibrated outputs are essential for reliable risk stratification and informed clinical decision-making [37, 38]. Therefore, post hoc calibration was performed using Platt scaling (sigmoid calibration). Calibration was conducted outside the inner hyperparameter optimization loop to reduce computational overhead while maintaining strict separation between training and evaluation data.

The decision threshold was selected using calibrated out-of-fold calibrated predictions generated exclusively within the outer training partition. The optimal decision threshold was determined by maximizing Youden’s index ($J = \text{Sensitivity} + \text{Specificity} - 1$). Maximizing *J* identifies the decision threshold that jointly optimizes both sensitivity and specificity and provides the best trade-off between correctly identifying sMCI and pMCI [39, 40]. Following threshold optimization, the calibrated model was refitted on the complete outer training fold and evaluated on the held-out outer test fold using the optimized threshold.

For each outer split, multiple evaluation metrics were computed; balanced accuracy, macro-averaged F1-score, sensitivity, specificity, precision–recall area

under the curve (PR-AUC), receiver operating characteristic area under the curve (ROC-AUC), and Brier score. Threshold-dependent metrics were computed using the optimized decision threshold, whereas PR-AUC and ROC-AUC were computed using calibrated probability estimates.

Model configurations were primarily ranked by mean balanced accuracy, with fold-level variability considered as a secondary criterion. Finally, the best-performing configuration for each ML algorithm was separately evaluated on the baseline dataset. This enabled a controlled comparison between the proposed dynamic single-visit framework and conventional baseline-only approaches.

Explainable AI (XAI)

To improve interpretability, explainable artificial intelligence (XAI) analyses were performed on the selected model configuration within the same nested cross-validation framework used for model evaluation. Feature importance estimates were computed directly on the final calibrated model using the optimized decision threshold to reflect the model's operational performance. This was achieved via permutation importance with 50 repetitions per feature. To provide a multi-dimensional assessment of feature importance, importance was evaluated against balanced accuracy, and macro-averaged F1-score.

Global feature contributions were further analyzed using SHapley Additive exPlanations (SHAP). A model-agnostic kernel SHAP explainer was employed for all models to ensure consistency across heterogeneous learning algorithms and externally calibrated prediction pipelines. SHAP values were computed using calibrated model outputs with a background reference set consisting of 100 randomly sampled normalized training instances. Due to class imbalance, SHAP analyses were conducted separately for sMCI and pMCI classes. Within each fold, SHAP values were averaged across samples within each class and subsequently aggregated across outer folds to obtain more stable global importance estimates. Features excluded during pre-processing or feature selection in each fold were assigned zero contribution to maintain consistency across folds. Feature stability was assessed by tracking the frequency with which features were either removed during variance

or correlation filtering or retained after feature selection across cross-validation folds.

Final model development and evaluation

For the final deployable framework, hyperparameter optimization was performed on the best-performing configuration identified in the “[Nested cross validation for model selection](#)” section. To handle potential missing data during deployment, an imputation module was integrated into the inference pipeline after evaluating mean, median, and k-nearest neighbors (KNN) imputation strategies (Section S1). Optimization was restricted to the training split (80% of the dynamic dataset) described in the “[Data partitioning strategy](#)” section. The optimal hyperparameter configuration was selected based on balanced accuracy from randomized search over 1000 hyperparameter configurations.

Following hyperparameter optimization, the selected pipeline was refitted on the complete training split and subsequently calibrated using Platt scaling. Decision-threshold optimization was then performed using out-of-fold calibrated probabilities generated from the training data, following the same calibration and threshold-tuning procedure described in the “[Nested cross validation for model selection](#)” section. The final calibrated model with the optimized decision threshold was evaluated on the test set (20% of the dynamic dataset). Beyond the primary evaluation metrics outlined in the “[Nested cross validation for model selection](#)” section, test set performance was further characterized using the Positive Predictive Value (PPV), Negative Predictive Value (NPV), Matthews correlation coefficient (MCC), and log loss.

Results

Top-performing model configuration

Evaluating 84 distinct pipeline configurations based on balanced accuracy from nested cross-validation identified XGB with the DEM-COG-MRI feature set and SMOTENC sampling as the best-performing configuration. This final selected configuration achieved a balanced accuracy of $81.28\% \pm 2.56\%$, sensitivity of $83.17\% \pm 5.01\%$, and specificity of $79.39\% \pm 5.54\%$

(Table 1). The mean optimized decision threshold of 0.31, with an average brier score of 0.131.

The integration of all three modalities resulted in higher balanced accuracy than any individual modality, yielding absolute balanced accuracy gains of +2.38% over COG modality, +10.35% over MRI modality, and +19.33% over DEM modality (Table 1). Furthermore, sensitivity decreased as the prediction horizon increased from 95.01% [95% CI: 93.51%–96.50%] at 6 months to 87.06% [84.80%–89.32%] at 12 months, and 73.59% [70.57–76.61%] at 24 months (Fig. 2).

Addressing class imbalance with SMOTENC resulted in a modest improvement in balanced accuracy compared with the unsampled configuration. Balanced accuracy increased from $80.80\% \pm 2.95\%$ to $81.28\% \pm 2.56\%$. This overall gain was driven by a substantial increase in specificity from $77.31\% \pm 4.30\%$ to $79.39\% \pm 5.54\%$ (+2.08%), accompanied by a slight decrease in sensitivity from $84.30\% \pm 5.57\%$ to $83.17\% \pm 5.01\%$ (−1.13%) (Table 1).

Finally, the selected dynamic single-visit configuration outperformed the baseline-only strategy across the reported evaluation metrics. Balanced accuracy increased from 76.49% to 81.28% ($\Delta = +4.79\%$), sensitivity from 80.18% to 83.17% ($\Delta = +2.99\%$), and specificity from 72.8% to 79.39% ($\Delta = +6.59\%$) (Fig. 3).

Explainable AI (XAI)

We performed explainable AI (XAI) analyses on the primary configuration identified in the “[Top-performing model configuration](#)” section. The analyses primarily focused on evaluating consistency in feature-selection stability, and patterns of feature contribution across different data partitions. Detailed feature-level attribution analyses are provided in the Supplementary Material.

Feature stability analysis

From the initial set of 334 features, 11 features were removed through correlation-based filtering across the 25-fold outer cross-validation folds, and no features were excluded based on variance filtering. The removed features were predominantly derived from the MRI modality. These included one surface area (SA) feature and nine subcortical volume (SV) measures. Only a single cognitive feature, ADAS13, was removed in 2 of the 25 folds.

Four MRI-derived features, comprising one SA feature and three SV measures, were consistently removed across all 25 folds, indicating persistent redundancy throughout the evaluation process. An additional two SV features were excluded in two folds, while the remaining four MRI variables were removed in only one-fold each.

Table 1 Performance evaluation of the selected model configurations and ablation study. Evaluation across resampling techniques and feature modality combinations ranked by balanced

accuracy across 25-fold outer cross-validation. Values are presented as mean $\pm$ standard deviation

Model	Feature Set	Sampling	F1 Macro	PR AUC	Balanced Acc. (%)	Sensitivity (%)	Specificity (%)
XGB	DEM-COG-MRI	SMOTENC	0.797 $\pm$ 0.030	0.791 $\pm$ 0.045	81.28 $\pm$ 2.56	83.17 $\pm$ 5.01	79.39 $\pm$ 5.54
XGB	DEM-COG-MRI	SMOTENC-Tomek	0.794 $\pm$ 0.025	0.784 $\pm$ 0.046	81.10 $\pm$ 2.32	83.31 $\pm$ 5.36	78.89 $\pm$ 4.80
XGB	DEM-COG-MRI	None	0.788 $\pm$ 0.029	0.790 $\pm$ 0.046	80.80 $\pm$ 2.95	84.30 $\pm$ 5.57	77.31 $\pm$ 4.30
XGB	COG-MRI	SMOTENC	0.789 $\pm$ 0.029	0.776 $\pm$ 0.048	80.62 $\pm$ 2.72	83.02 $\pm$ 4.79	78.22 $\pm$ 4.79
XGB	DEM-COG	SMOTENC	0.788 $\pm$ 0.021	0.791 $\pm$ 0.039	79.78 $\pm$ 2.12	78.54 $\pm$ 5.28	81.01 $\pm$ 4.07
XGB	COG	SMOTENC	0.775 $\pm$ 0.029	0.779 $\pm$ 0.051	78.90 $\pm$ 2.53	79.47 $\pm$ 4.84	78.34 $\pm$ 5.66
XGB	DEM-MRI	SMOTENC	0.721 $\pm$ 0.043	0.689 $\pm$ 0.056	73.40 $\pm$ 4.28	72.59 $\pm$ 8.17	74.20 $\pm$ 6.71
XGB	MRI	SMOTENC	0.693 $\pm$ 0.040	0.655 $\pm$ 0.055	70.93 $\pm$ 3.67	71.74 $\pm$ 8.61	70.11 $\pm$ 8.48
XGB	DEM	SMOTENC	0.601 $\pm$ 0.042	0.464 $\pm$ 0.048	61.95 $\pm$ 4.48	64.35 $\pm$ 8.63	59.56 $\pm$ 6.12

Model: *XGB* Extreme Gradient Boosting. Feature sets combine: *DEM* demographics and clinical risk factors; *COG* cognitive and functional assessments; *MRI* magnetic resonance imaging-derived structural features. Sampling methods include: *SMOTENC* Synthetic Minority Over-sampling Technique for Nominal and Continuous features; *SMOTENC-Tomek* SMOTENC combined with Tomek links; *None* no resampling applied. Metrics: *F1 Macro* macro-averaged F1-score; *PR AUC* precision-recall area under the curve; *Balanced Acc* balanced accuracy

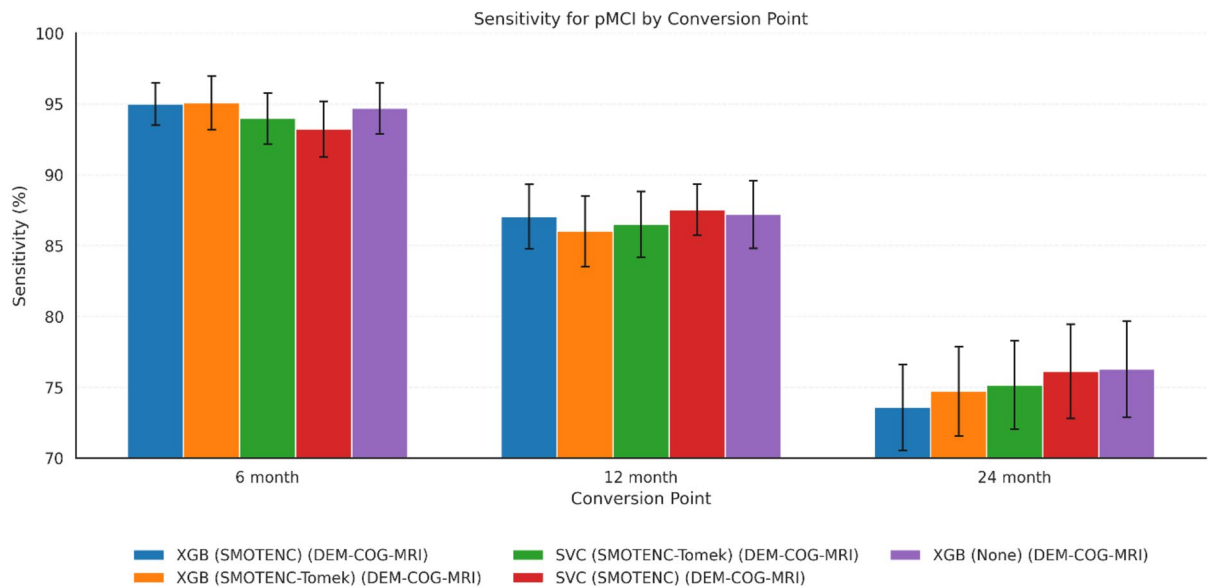


Fig. 2 Classification sensitivity on progressive mild cognitive impairment (pMCI) samples across conversion horizons. Sensitivity (%) for the top five model configurations ranked by balanced accuracy and evaluated at 6-month, 12-month, and 24-month conversion points. Error bars represent 95% confidence intervals estimated via five fold cross-validation repeated 5 times. Model configurations combine classifiers—XGB (Extreme Gradient Boosting) and SVC (Support Vector

Classifier)—with resampling techniques including SMOTENC (Synthetic Minority Over-sampling Technique for Nominal and Continuous features), and SMOTENC-Tomek (SMOTENC with Tomek links undersampling). All top models utilized a multimodal feature set consisting of DEM (demographics and clinical risk factors), COG (cognitive and functional assessments), and MRI (magnetic resonance imaging-derived structural features)

Feature stability analysis identified eight features that were consistently selected across all 25 folds. Among these stable features, APOE4 represented DEM modality, while four features originated from COG modality: FAQ, ADAS11, CDRSB, and MMSE. The remaining three stable features were MRI modality (SV measures of left and right hippocampus and CT measure of right fusiform).

Permutation importance analysis

Cognitive variables showed the highest permutation importance values across both balanced accuracy and macro-averaged F1-scores (Fig. 4). Among all features, ADAS13 demonstrated the largest contribution (balanced accuracy=4.15%, selected in 23 folds) followed by FAQ (3.41%, 25), and CDRSB (1.83%, 25 folds). APOE4 was the most influential non-cognitive feature, ranking fourth overall with permutation importance values of 1.13% for balanced accuracy and was also selected in all 25 folds. In contrast, MRI-derived biomarkers generally exhibited

lower permutation importance values. Among the neuroimaging features, ventricular volume measures, particularly the left inferior lateral ventricle, showed the highest permutation importance, followed by CT measure in the right entorhinal and SV measures in left hippocampus and right amygdala.

Class-balanced SHAP feature importance analysis

We further analyzed SHAP feature importance separately for the sMCI and pMCI groups (Fig. 5). Cognitive variables demonstrated the highest absolute SHAP values across both groups. In the sMCI group, FAQ demonstrated the highest contribution (mean |SHAP|=0.067, selected in 25 folds), followed by ADAS13 (0.066, 23 folds) and CDRSB (0.0419, 25 folds) (Fig. 5a). In the pMCI group, ADAS13 showed the largest contribution (0.090, 23 folds), followed by FAQ (0.084, 25 folds) and CDRSB (0.057, 25 folds) (Fig. 5b). Additionally, APOE4, MMSE and ADAS11 also showed comparatively high SHAP values across both groups.

Performance Comparison of Selected Models: Baseline vs Dynamic Dataset

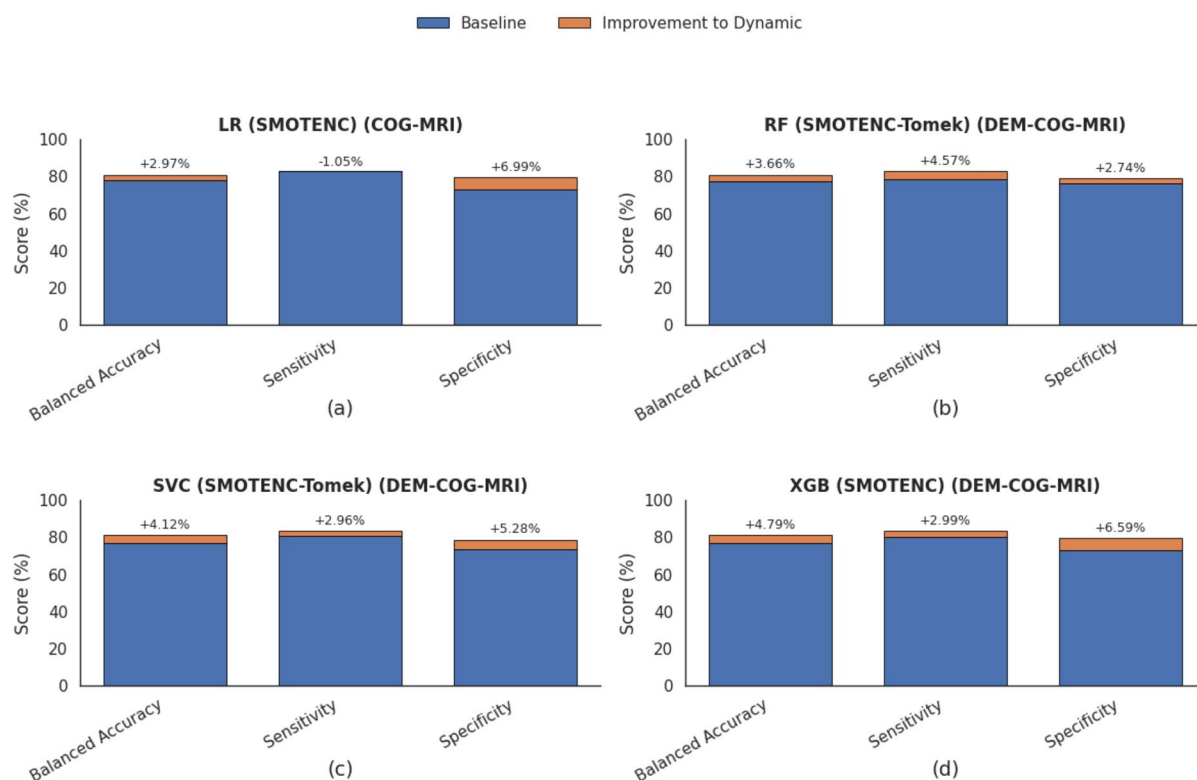


Fig. 3 Performance comparison between baseline-only and dynamic datasets across evaluation metrics. Performance across balanced accuracy, sensitivity, and specificity for the optimal configuration of each machine learning algorithm: (a) LR (logistic regression), (b) RF (random forest), (c) SVC (support vector classifier), and (d) XGB (Extreme Gradient Boosting). Blue bars represent baseline dataset performance, while stacked orange segments indicate performance changes achieved using the dynamic dataset, with exact percentage dif-

ferences labeled above each bar. Resampling methods include SMOTENC (Synthetic Minority Over-sampling Technique for Nominal and Continuous features) and SMOTENC-Tomek (SMOTENC combined with Tomek links undersampling). Feature modalities encompass DEM (demographics and clinical risk factors), COG (cognitive and functional assessments), and MRI (magnetic resonance imaging-derived structural features)

MRI-derived features exhibited lower but consistent SHAP contributions. Among the neuroimaging features for pMCI, left inferior lateral ventricle (SV), left hippocampus (SV), left inferior parietal (CT), and right inferior parietal (CV), ranked highest in both feature importance and selection frequency.

Independent test-set performance

Evaluation of imputation strategies within the final cross-validated framework described in the “[Top-performing model configuration](#)” section demonstrated that KNN imputation achieved the highest balanced accuracy ($80.31\% \pm 1.86\%$) compared to median

($79.74\% \pm 1.88\%$) and mean ($79.66\% \pm 2.05\%$) imputation. Consequently, KNN ($k=5$) imputation was integrated into the final deployment pipeline.

Following the integration of KNN imputation, we evaluated 401 samples, including 261 sMCI cases (65.1%) and 140 pMCI cases (34.9%) on the independent held-out test set (Table 2, Fig. 6). The finalized pipeline employed mutual information filtering to select 50 primary features from the training distribution (two DEM features: APOE4 and education, five COG features: MMSE, CDRSB, ADAS11, ADAS13, and FAQ, and 43 MRI features) (Table S7).

Using the default decision threshold (0.50), the uncalibrated model achieved a balanced accuracy of

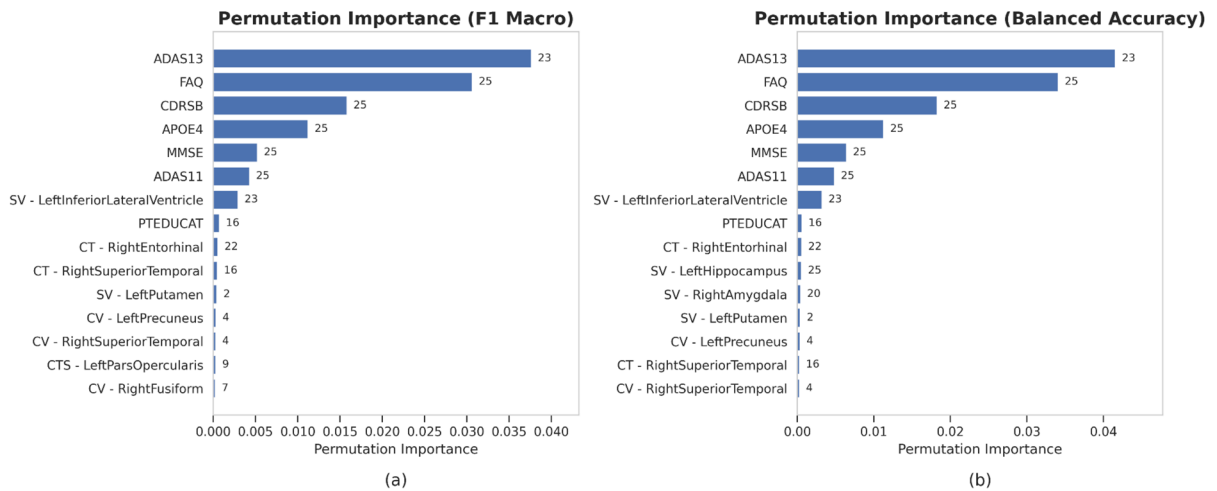


Fig. 4 Permutation importance rankings of the top 15 features. Permutation importance for the optimized XGB (Extreme Gradient Boosting) model utilizing SMOTENC (Synthetic Minority Over-sampling Technique for Nominal and Continuous features) resampling and DEM-COG-MRI multimodal features (DEM=demographics and clinical risk factors; COG=cognitive and functional assessments; MRI=magnetic resonance imaging-derived structural features). Importance

values are evaluated across 25-fold outer cross-validation using (a) macro F1-score and (b) balanced accuracy. Numerical annotations indicate the feature selection frequency across 25 folds. Structural MRI feature prefixes denote specific morphometric metrics: SV (subcortical volume), CT (cortical thickness), CV (cortical volume), and CTS (cortical thickness standard deviation)

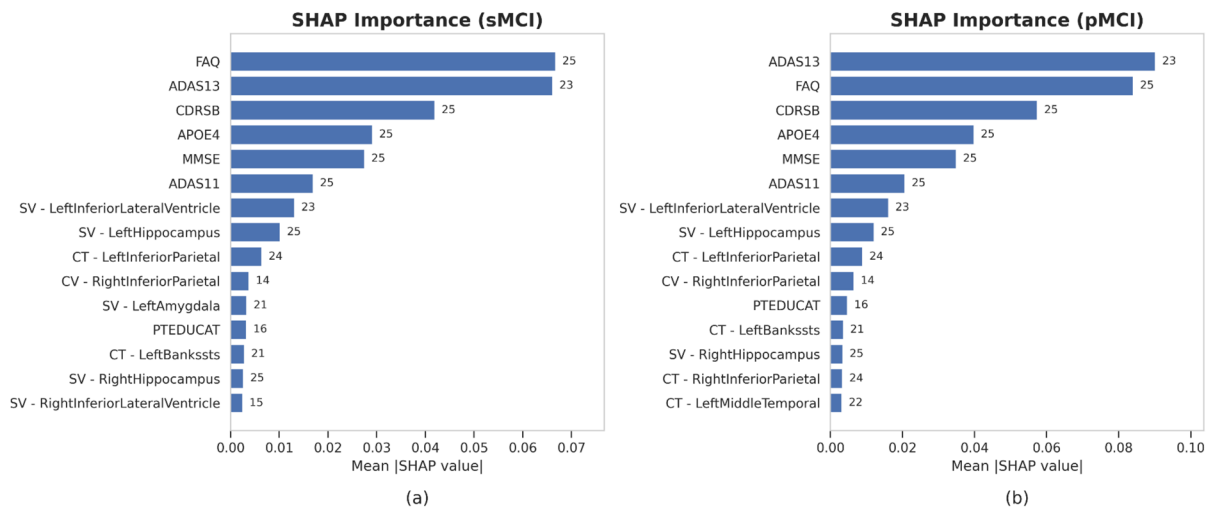


Fig. 5 SHAP (SHapley Additive exPlanations) importance rankings of the top 15 features across diagnostic groups. SHAP mean absolute feature importance values for (a) sMCI (stable mild cognitive impairment) and (b) pMCI (progressive mild cognitive impairment). Values were derived from the optimized XGB (Extreme Gradient Boosting) model using SMOTENC (Synthetic Minority Over-sampling Technique for Nominal and Continuous features) resampling with DEM-COG-MRI multimodal features (DEM=demographics and

clinical risk factors; COG=cognitive and functional assessments; MRI=magnetic resonance imaging-derived structural features). Importance rankings were evaluated across a 25-fold outer cross-validation, with numerical annotations next to each bar denoting the feature selection frequency out of the 25 folds. Structural MRI feature prefixes denote specific morphometric metrics: SV (subcortical volume), CT (cortical thickness), and CV (cortical volume)

Table 2 Independent test-set (N=401) performance of the optimized and calibrated extreme gradient boosting model using synthetic minority over-sampling technique for nominal and continuous features with demographics, cognitive, and MRI features [XGB (SMOTENC) DEM-COG-MRI]. Point estimates on the full test set alongside mean bootstrap estimates with 95% confidence intervals derived from 2000 bootstrap resamples are reported

Metric	Average	Bootstrap Average [95% CI]
F1 (sMCI)	0.847	0.847 [0.813–0.878]
F1 (pMCI)	0.759	0.758 [0.706–0.805]
Macro F1	0.803	0.802 [0.763–0.839]
Sensitivity (%)	84.29	84.28 [77.70–90.00]
Specificity (%)	79.69	79.66 [74.81–84.21]
Balanced Accuracy (%)	81.99	81.97 [78.22–85.51]
PPV (%)	69.01	68.94 [62.35–75.45]
NPV (%)	90.43	90.44 [86.50–94.02]
ROC AUC	0.886	0.886 [0.853–0.915]
PR AUC	0.791	0.793 [0.720–0.854]
MCC	0.617	0.616 [0.542–0.687]
Brier Score	0.131	0.131 [0.111–0.153]
Log Loss	0.409	0.409 [0.357–0.465]

Diagnostic targets: *sMCI* stable mild cognitive impairment; *pMCI* progressive mild cognitive impairment. Metrics: *PPV* positive predictive value; *NPV* negative predictive value; *ROC AUC* area under the receiver operating characteristic curve; *PR AUC* precision-recall area under the curve; *MCC* Matthews correlation coefficient

80.23% (sensitivity = 80.00%, specificity = 80.46%). After probability calibration, balanced accuracy was 80.31%, with a sensitivity of 77.86% and specificity of 82.76%. Optimizing the decision threshold of the calibrated model to 0.37 yielded the balanced accuracy of 81.99% (mean: 81.97%; 95% CI: 78.22%–85.51%; evaluated via 2000 bootstrap resamples). The performance was accompanied by a sensitivity of 84.29% (mean: 84.28%; 95% CI: 77.70%–90.00%), specificity of 79.69% (mean: 79.66%; 95% CI: 74.81%–84.21%), positive predictive value of 69.01% (mean: 68.94; 95% CI: 62.35%–75.45%), and negative predictive value of 90.43% (mean: 90.44%; 95% CI: 86.50%–94.02%).

Ablation and modality contribution analyses

In ablation experiments on the independent test set, restricting feature input to single modalities reduced balanced accuracy compared to the full multimodal

baseline by 4.73% (95% CI: 0.94%–8.39%) for COG modality, 5.22% (95% CI: 0.66%–9.49%) for MRI modality, and 17.62% (95% CI: 12.37%–23.05%) for DEM modality. At the individual feature level, excluding APOE4 reduced balanced accuracy by 1.11%, excluding MMSE decreased balanced accuracy by 1.17%, and removing CDRSB, ADAS13, and FAQ resulted in performance reductions of 0.02%, 0.14%, and 0.95%, respectively.

Discussion

In this study, we developed and validated a multi-modal machine-learning framework using single-visit cognitive and structural MRI features. It predicts progression from MCI to dementia within a clinically meaningful two-year horizon. The selected XGB model demonstrated consistent performance across nested cross-validation, and independent testing, indicating robust generalizability within the ADNI cohort. Importantly, the findings suggest that short-term conversion risk is primarily associated with the combined effects of cognitive-functional impairment and region-specific neurodegenerative changes. This observation supports the concept that clinical decline and structural brain degeneration may evolve in parallel during the prodromal stages of dementia.

The selection of a two-year prediction horizon balances clinical utility with prognostic reliability. While a one-year window offers a narrow timeframe that restricts therapeutic intervention, longer horizons (3–5 years) introduce significant uncertainty driven by biological heterogeneity, intercurrent health conditions, and participant attrition. This pattern is reflected in our empirical findings, where predictive performance is attenuated as the conversion horizon expanded from 6 to 24 months (Fig. 2). Consequently, these findings reinforce the importance of selecting clinically meaningful prediction windows that balance temporal precision with opportunities for intervention.

The strong contribution of cognitive variables, including FAQ, ADAS11, ADAS13, CDR-SB, and MMSE, is consistent with their established associations with functional decline and disease progression. These variables ranked among the most influential predictors in both permutation importance and SHAP analyses (Figs. 4 and 5). Although several of these

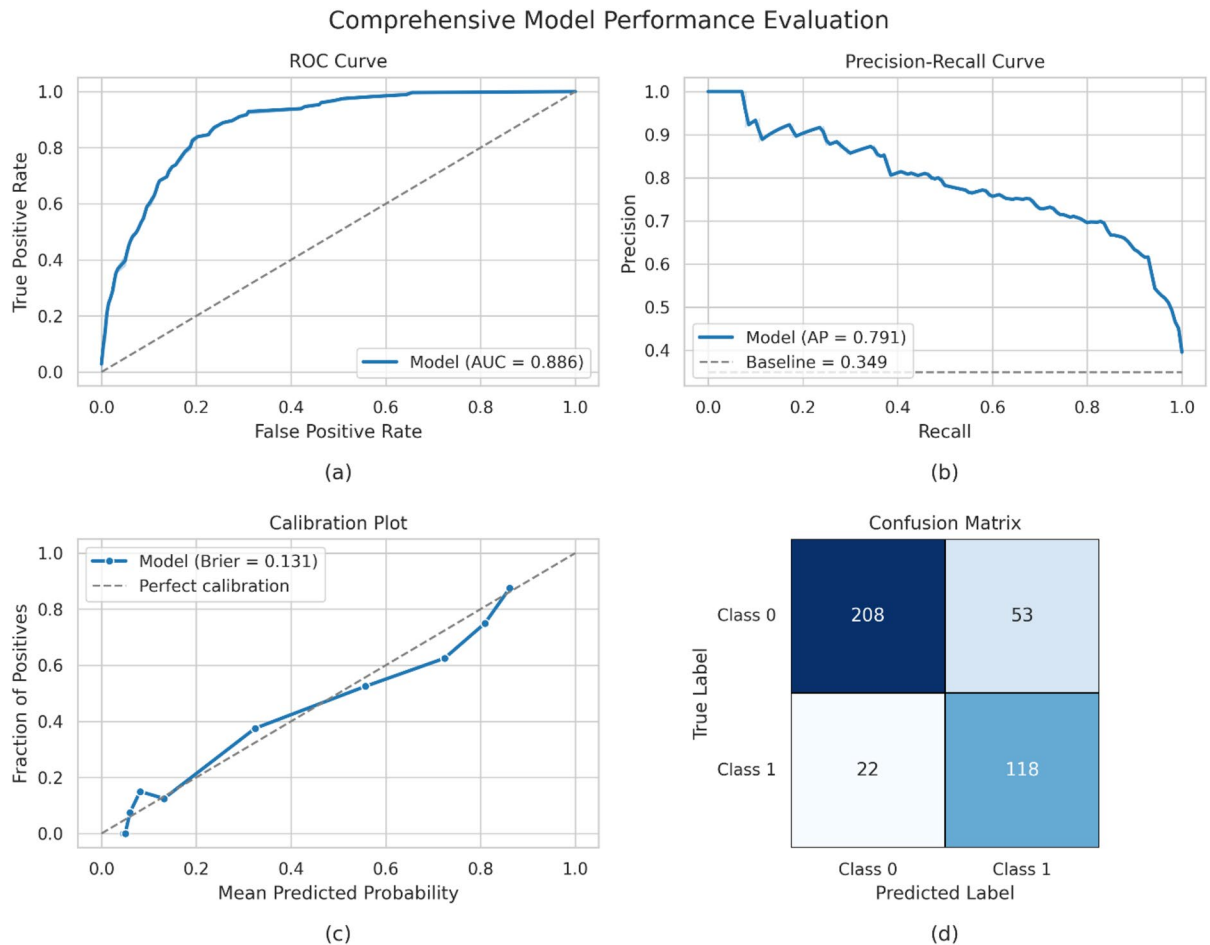


Fig. 6 Performance evaluation of the optimized model on the independent held-out test set (N=401). Metrics evaluated include: **(a)** ROC (receiver operating characteristic) curve yielding an AUC (area under the receiver operating characteristic curve) of 0.886; **(b)** Precision–recall curve showing an AP (average precision) of 0.791 compared to the random baseline

of 0.349; **(c)** Calibration curve with a Brier score of 0.131 evaluating predicted probability reliability against perfect calibration; and **(d)** Confusion matrix detailing classification outcomes for sMCI (stable mild cognitive impairment; Class 0) and pMCI (progressive mild cognitive impairment; Class 1)

measures contribute to clinical evaluation, dementia conversion in this study was determined through longitudinal diagnostic assessment rather than predefined score thresholds. Consequently, these measures functioned as prognostic indicators of future decline rather than direct determinants of the outcome. Interestingly, ADAS13 emerged as one of the most influential predictors despite being excluded during correlation filtering in two of the 25 cross-validation folds. This pattern likely reflects fold-specific redundancy with correlated predictors rather than instability in its prognostic value. Its strong importance when retained further indicates that occasional exclusion

due to multicollinearity does not diminish its overall predictive relevance. Importantly, the present findings do not suggest that these measures represent novel biomarkers; rather, they highlight their enhanced prognostic contribution when integrated with demographic, genetic, and structural MRI information within a fixed two-year prediction framework.

Structural MRI provided complementary biological information into cognitive assessment, yielding the highest overall predictive performance when combined into a multimodal framework (Table 1). Structural MRI alone demonstrated lower predictive accuracy than cognitive modality alone. However,

it offers an objective measure of neurodegeneration that is less directly influenced by factors such as educational attainment, cognitive reserve, or test-taking variability. MRI-enhanced prediction may therefore be particularly valuable in atypical clinical presentations or cases with conflicting neuropsychological findings. Overall, these results advocate for a tiered clinical framework: cognitive screening as a practical initial first-line evaluation, with MRI-enhanced multimodal assessment targeted to high-risk, atypical, or diagnostically ambiguous cases.

Rather than pre-selecting regions based on prior biological hypotheses, our feature-selection and explainability pipeline operated in a strictly data-driven manner across all 323 MRI-derived features. Despite this unbiased approach, the model consistently prioritized the hippocampus, amygdala, entorhinal cortex, inferior parietal cortex, temporal cortex, and inferior lateral ventricles (Figs. 4 and 5). These structures are known to undergo early neurodegenerative changes during Alzheimer's disease and play critical roles in episodic memory, semantic processing, and large-scale cortical network integration [41–44]. Accordingly, these findings should be interpreted primarily as confirmation that established neuroanatomical markers contribute meaningfully to short-term prediction, rather than as evidence for newly discovered imaging biomarkers.

APOE ϵ 4 status and educational attainment, both established risk factors for Alzheimer's disease, contributed additional predictive information [45–47].

An important finding was the superior performance of the dynamic single-visit strategy compared with baseline-only modeling approaches. Unlike conventional baseline approaches, our multimodal machine learning framework with post-hoc explainability estimates progression risk from any single visit instance. Furthermore, because the framework operates independently of historical observations, it bypasses the need for complete longitudinal trajectories. Instead, each qualifying visit can serve as a new prediction point for the subsequent fixed two-year horizon. This dynamic single-visit strategy enables practical applicability where clinicians can update risk assessments as new data becomes available without requiring complete longitudinal trajectories or baseline data. Such flexibility is particularly advantageous in routine clinical practice to complement

clinical expertise and support follow-up planning or clinical trial enrollment.

The relatively small performance differences among the four ML models (LR, SVC, RF, and XGB) suggest that feature representation contributed more towards predictive performance than classifier complexity. The selected XGB model likely benefited from its ability to capture nonlinear relationships within the reduced feature space while maintaining strong generalization.

At the global feature level, SHAP and permutation importance showed a weak negative rank correlation (Spearman $\rho = -0.19$), despite a strong linear correlation in importance magnitudes (Pearson $r = 0.97$). Among the top 15 features, 9 were shared between the two methods, corresponding to a Jaccard similarity index of 0.43. The two methods also showed concordance among the highest-ranked features, with the top 7 features exhibiting agreement in rank. Collectively, these findings suggest that although SHAP and permutation importance produce broadly consistent estimates of feature-importance magnitude, they may differ substantially in the relative ordering of features beyond the most influential predictors.

Probability calibration and threshold optimization further improve the potential clinical applicability of the proposed framework (Figs. 6 and S1). Although the uncalibrated model demonstrated well-balanced sensitivity and specificity, its predicted probabilities exhibited slight overconfidence, as indicated by the calibration curve. Post hoc calibration effectively produced probability estimates that more closely reflected the true likelihood of future conversion. Furthermore, threshold optimization using the Youden's index approach provided a clinically reasonable compromise for risk stratification by reducing the potential bias toward either false-positive or false-negative predictions. Nevertheless, different clinical settings may favor alternative thresholds and can be dynamically adjusted to fit specific priorities, such as maximizing early case detection versus minimizing unnecessary follow-up.

From a clinical implementation perspective, the proposed framework is primarily intended for use in specialist memory clinics, where standardized cognitive assessments and structural brain MRI are routinely obtained as part of the diagnostic workup for patients with MCI. In these settings, the model could serve as a complementary decision-support tool to

estimate an individual's short-term progression risk. Such estimates could help clinicians prioritize follow-up, support patient counseling, and identify individuals who may warrant additional biomarker evaluation or consideration of disease-modifying therapies or clinical trials. Because the dynamic single-visit framework does not require a complete longitudinal history, it may be more readily integrated into existing memory-clinic workflows. However, broader translation faces key barriers, including MRI cost and availability, inter-site variability in imaging protocols and cognitive scoring, and the required computational infrastructure. Overcoming these challenges will require multi-center protocol harmonization and electronic health record (EHR) integration. Prospective external validation is required before clinical implementation.

We evaluated the proposed framework against recent state-of-the-art approaches (2023–2025) for predicting MCI-to-dementia conversion across different models, modalities, and evaluation strategies [23, 24, 26, 31, 48]. As detailed in Table 3, the proposed

XGB model achieved higher balanced accuracy than those reported by Mieling et al. (77.00%), Agostinho et al. (78.43%), Gao et al. (70.60%), and Bucholtz et al. (76.95%). However, these numerical comparisons should be interpreted cautiously due to substantial methodological heterogeneity across studies, including differences in cohort demographics, sample sizes, input modalities, follow-up durations, and outcome definitions. Despite these methodological differences, our framework achieved competitive predictive accuracy while maintaining a balanced trade-off between sensitivity and specificity.

Limitations and future work

Several limitations should be acknowledged (Table 4). The proposed framework was developed using the ADNI cohort, which was collected under controlled research protocols. As such, it may not fully capture the heterogeneity and complexity of real-world clinical populations, potentially limiting generalizability.

Table 3 Comparison of the proposed Extreme Gradient Boosting (XGB) model with state-of-the-art mild cognitive impairment (MCI)-to-dementia conversion prediction methods across feature modalities, prediction strategies, and evaluation settings. Diagnostic targets compare stable mild cognitive

impairment (sMCI) and progressive mild cognitive impairment (pMCI). Performance metrics for the proposed framework are reported for both 25-fold cross-validation (CV) and independent hold-out test set evaluation (N=401; 261 sMCI/140 pMCI)

Study	Model	Prediction Strategy	Modalities	Test Setting	Balanced Acc	F1 (pMCI)	Sensitivity (pMCI)	Specificity (sMCI)	ROC-AUC
Gelir et al. 2024 [26]	XGB	Base-line → last visit	COG + MRI + PET + DEM	tenfold CV	–	0.757	75.23	–	0.840
Mieling et al. 2025 [23]	XGB	Base-line → last visit	MRI + CSF + COG + DEM	16/15 split	77.00	0.740	66.70	87.50	–
Agostinho et al. 2024 [24]	Ensemble SVM	Base-line → last visit	MRI + FDG	tenfold CV	78.43	0.800	77.12	79.75	0.830
Gao et al. 2023 [31]	CNN + Transformer	3-year window	MRI	112/66 split	70.60	–	59.10	82.10	0.737
Bucholtz et al. 2023 [48]	LR	Fixed 2-year window using longitudinal input	COG + COG trajectories	tenfold CV	76.95	–	81.20	72.70	0.750
Proposed	XGB	Fixed 2-year window	MRI + COG + DEM	25-fold CV	81.28	–	83.17	79.39	0.89
				Hold-out (261/140)	81.99	0.759	84.29	79.69	0.886

Models: *XGB* Extreme Gradient Boosting; *Ensemble SVM* Support Vector Machine Ensemble; *CNN* Convolutional Neural Network; *LR* Logistic Regression. Modalities: *DEM* demographics and clinical risk factors; *COG* cognitive and functional assessments; *MRI* magnetic resonance imaging; *PET* positron emission tomography; *CSF* cerebrospinal fluid biomarkers; *FDG* fluorodeoxyglucose-PET. Metrics: *Balanced Acc.* balanced accuracy; *Sensitivity* pMCI detection rate; *Specificity* sMCI detection rate; *ROC-AUC* area under the receiver operating characteristic curve

Table 4 Summary of current study limitations, their potential impact on model performance and clinical applicability, and proposed future research directions

Limitation	Potential Impact	Future Direction
Single-cohort (ADNI) development	Limited generalizability to broader clinical populations	External validation and other independent cohorts
Exclusion of missing data	Reduced applicability and potential selection bias	Imputation/missing-data strategies operating with incomplete inputs in independent clinical cohorts
Reliance on FreeSurfer-derived MRI features	Potential software and preprocessing variability	Evaluate harmonized pipelines and alternative imaging features
Cohort Bias, Clinical Logs and Biomarker Omissions	Inherent selection bias and reduced generalizability	Validate on diverse, multi-site cohorts integrating plasma/CSF biomarkers, vital signs, and EHR data
Dynamic single-visit prediction only	Does not explicitly exploit temporal disease trajectories	Hybrid models combining dynamic prediction with longitudinal modeling
Multi visit-level samples per subject	Possible residual within-subject dependency	Evaluate methods explicitly modeling repeated observations

Despite enforcing strict subject-level data splits to prevent information leakage, the use of multiple visit-level observations introduces residual within-subject correlation because repeated observations from the same participant are not statistically independent.

Samples with missing values were excluded during primary model development to maintain a consistent multimodal feature space. To improve robustness during inference, KNN imputation was subsequently incorporated into the final pipeline after comparison with mean and median imputation strategies. Nevertheless, complete-case model development may introduce selection bias, and the performance of the imputation strategy requires further evaluation in clinical populations where missing multimodal data may be more prevalent.

Although nested cross-validation, subject-level partitioning, and independent held-out testing were used to reduce optimistic bias, model development and optimization were performed within a single cohort (ADNI); therefore, some cohort-specific overfitting cannot be excluded. External validation remains essential before real-world translation or clinical implementation.

We evaluated the feasibility of external validation using existing Alzheimer's disease datasets; however, currently available datasets did not permit a methodologically equivalent evaluation of the complete framework. OASIS-3 lacks several critical top-ranked predictors (FAQ, ADAS11, and ADAS13), such that extensive imputation would have materially altered

the intended feature space and compromised the interpretability of the benchmark. Evaluation using AIBL would require additional FreeSurfer preprocessing of raw MRI data, which was beyond the scope of the present study. Accordingly, rigorous multi-site harmonization and prospective external validation represent key priorities for future work.

Individuals who reverted from MCI to normal cognition were excluded as the present study focused specifically on progression to dementia. Although this simplified the prediction task and produced a more homogeneous outcome definition, it does not capture the complete spectrum of MCI trajectories. Future studies will extend this framework toward multi-class prediction capable of simultaneously forecasting progression, stability, and cognitive reversion.

Additionally, the MRI modality, which accounted for a significant portion of the feature space, relied on features extracted with the FreeSurfer pipeline. While widely adopted, FreeSurfer-based processing is computationally intensive and may introduce variability across different versions and preprocessing configurations, potentially affecting reproducibility. Furthermore, the current framework does not incorporate other informative biomarkers, such as positron emission tomography (PET) imaging and blood-based biomarkers, which have demonstrated strong predictive value in Alzheimer's disease progression [49–51]. Lastly, secondary clinical variables—including race/ethnicity, medical history data, and vital signs—were excluded due to dataset

constraints, which may limit the broader generalizability of the predictive framework.

In this study, we established that a dynamic, single-visit input framework can reliably predict MCI-to-dementia conversion within a two-year horizon. Building upon these findings, future research will focus on developing a deep learning-based multimodal framework that integrates the most informative cognitive features identified in this study with imaging modalities (MRI and PET), blood-based biomarkers and additional genetic information. To capture temporal dependencies, longitudinal modeling should be incorporated by leveraging features from multiple time points within a flexible temporal window.

Conclusion

This study presented an explainable multimodal machine learning framework for predicting conversion from MCI to dementia within a fixed two-year window using data from a single clinical visit. Comprehensive evaluation across 84 configurations demonstrated that multimodality consistently outperformed individual modalities, highlighting the complementary relationship between cognitive-functional decline and neuroanatomical atrophy. The best-performing configuration, an XGB using DEM-COG-MRI features with SMOTENC sampling, achieved strong and consistent performance across nested cross-validation and independent held-out testing. Furthermore, the proposed dynamic single-visit strategy consistently improved performance over baseline-only prediction approaches, while explainable AI analyses confirmed that the model relied primarily on clinically meaningful features associated with Alzheimer's disease progression.

Although these findings highlight the potential of multimodal AI for short-term dementia risk stratification, the proposed framework currently serves as a research tool rather than a tool ready for routine clinical implementation. Translation toward clinical use will require validation across independent and more diverse cohorts, systematic evaluation under incomplete multimodal data conditions, and prospective assessment of feasibility and clinical utility within real-world memory-care settings.

Author contribution Bipul Simkhada: research scope, methodology and framework development, investigation, literature synthesis, and original manuscript drafting. Thony Yan Liang and Xueting Cui: dataset curation and development, XAI implementation and analysis, and interpretation of model findings. Micheal Adeyoso: literature synthesis, model calibration, and decision-threshold development. Bibhu Simkhada: exploratory data analysis (EDA), original manuscript drafting, results analysis, visualization, and manuscript editing. Mercedes Cabrerizo, Armando Barreto, and Naphtali Rishe: conceptualization, methodology development, technical validation, interpretation of findings, and critical manuscript review. Rosie Curiel Cid and Shanna L. Burke: clinical evaluation and interpretation of findings, assessment of clinical relevance, formulation of clinical applicability, and manuscript review and editing. David A. Loewenstein: dataset development, validation of cognitive findings, interpretation of results, manuscript editing, and funding acquisition. Malek Adjouadi: research scope and conceptualization, framework development, methodological guidance, funding acquisition, and critical review and editing. All authors contributed to the interpretation of the results, reviewed the manuscript and approved the submitted article.

Funding This research is supported by the National Science Foundation (NSF) under grants CNS-1920182, CNS-2018611, the National Institutes of Health under grants NIA/NIH 1P30AG066506-01 with the IFlorida ADRC, NIA/NIH 5R01AG061106-02, NIA/NIH 5R01AG047649-05, R01AG077677-04, and the Ware Foundation.

Data Availability The datasets analyzed during the current study are available in the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (<https://adni.loni.usc.edu>). All derived dataset combinations generated during this study are available from the corresponding author upon request.

Code availability The code used to perform the analyses and generate the results presented in this study is publicly available at GitHub [https://github.com/bipulsimkhada/mci_progression_ml].

Declarations

Competing interests The authors declare no competing interests.

References

1. Alzheimer's disease facts and figures. In: Alzheimer's Association. <https://www.alz.org/alzheimers-dementia/facts-figures>. Accessed 1 Aug 2026
2. Metcalfe MJ, Figueiredo-Pereira ME. Relationship between tau pathology and neuroinflammation in Alzheimer's disease. *Mt Sinai J Med*. 2010;77:50–8. <https://doi.org/10.1002/msj.20163>.
3. Ratan Y, Rajput A, Maleysm S, et al. An insight into cellular and molecular mechanisms underlying the

- pathogenesis of neurodegeneration in Alzheimer's disease. *Biomedicines*. 2023;11. <https://doi.org/10.3390/biomedicines11051398>.
4. Kamatham PT, Shukla R, Khatri DK, Vora LK. Pathogenesis, diagnostics, and therapeutics for Alzheimer's disease: breaking the memory barrier. *Ageing Res Rev*. 2024;101. <https://doi.org/10.1016/j.arr.2024.102481>.
 5. Anand S, Schoo C (2026) Mild cognitive impairment. In: StatPearls. StatPearls Publishing, Treasure Island (FL)
 6. Mitchell AJ, Shiri-Feshki M. Rate of progression of mild cognitive impairment to dementia – meta-analysis of 41 robust inception cohort studies. *Acta Psychiatr Scand*. 2009;119:252–65. <https://doi.org/10.1111/j.1600-0447.2008.01326.x>.
 7. Eshkkoor SA, Hamid TA, Mun CY, Ng CK. Mild cognitive impairment and its management in older people. *CIA*. 2015;10:687–93. <https://doi.org/10.2147/CIA.S73922>.
 8. Rossini PM, Pappalettera C. Should all MCI with Alzheimer's biological diagnosis receive anti-amyloid therapy? *Cell Death Dis*. 2026;17:212. <https://doi.org/10.1038/s41419-026-08456-z>.
 9. Petersen RC, Doody R, Kurz A, et al. Current concepts in mild cognitive impairment. *Arch Neurol*. 2001;58:1985–92. <https://doi.org/10.1001/archneur.58.12.1985>.
 10. Matthews FE, Stephan BCM, McKeith IG, et al. Two-year progression from mild cognitive impairment to dementia: to what extent do different definitions agree? *J Am Geriatr Soc*. 2008;56:1424–33. <https://doi.org/10.1111/j.1532-5415.2008.01820.x>.
 11. Ganguli M, Jia Y, Hughes TF, et al. Mild cognitive impairment that does not progress to dementia: a population-based study. *J Am Geriatr Soc*. 2019;67:232–8. <https://doi.org/10.1111/jgs.15642>.
 12. Takabatake K, Ohara T, Minohara T, et al. Short-term and long-term reversion rates to normal cognition and their contributing factors among individuals with mild cognitive impairment in a Japanese community: the Hisayama study. *BMC Geriatr*. 2025;25. <https://doi.org/10.1186/s12877-025-06750-7>.
 13. Arya AD, Verma SS, Chakarabarti P, et al. A systematic review on machine learning and deep learning techniques in the effective diagnosis of Alzheimer's disease. *Brain Inform*. 2023;10. <https://doi.org/10.1186/s40708-023-00195-7>.
 14. Ai M, Liu Y, Liu D, et al. Research progress in predicting the conversion from mild cognitive impairment to Alzheimer's disease via multimodal MRI and artificial intelligence. *Front Neurol*. 2025. <https://doi.org/10.3389/fneur.2025.1596632>.
 15. El-Sappagh S, Saleh H, Ali F, et al. Two-stage deep learning model for Alzheimer's disease detection and prediction of the mild cognitive impairment time. *Neural Comput Appl*. 2022;34:14487–509. <https://doi.org/10.1007/s00521-022-07263-9>.
 16. Aghajanian S, Mohammadifard F, Mohammadi I, et al. Longitudinal structural MRI-based deep learning and radiomics features for predicting Alzheimer's disease progression. *Alzheimers Res Ther*. 2025;17:182. <https://doi.org/10.1186/s13195-025-01827-2>.
 17. Abdelhameed A, Feng J, Hu X, et al. AI-powered model for accurate prediction of MCI-to-AD progression. *Acta Pharm Sin B*. 2025;15:4427–37. <https://doi.org/10.1016/j.apsb.2025.01.027>.
 18. C MV, G B (2025) Predicting cognitive decline: a hybrid CNN-GRU model for the prognosis of MCI-to-AD conversion. In: 2025 IEEE 17th International Conference on Computational Intelligence and Communication Networks (CICN). pp 2262–2267
 19. Zheng G, Lu Y, Chen H. Deep learning-based framework for predicting mild cognitive impairment progression in neurology using longitudinal MRI. *IEEE Access*. 2025;13:68903–19. <https://doi.org/10.1109/ACCESS.2025.3562432>.
 20. Karasu E, Baytaş İM. Conversion-aware forecasting of Alzheimer's disease via featurewise attention. *Pattern Anal Appl*. 2025;28:64. <https://doi.org/10.1007/s10044-025-01447-4>.
 21. Karaman BK, Nguyen M, Kim H, Sabuncu MR. A deep survival model for predicting Alzheimer's diagnosis based on multi-modal longitudinal data. *Big Data Min Anal*. 2026;9:465–80. <https://doi.org/10.26599/BDMA.2025.9020064>.
 22. Adelson RP, Garikipati A, Maharjan J, et al. Machine learning approach for improved longitudinal prediction of progression from mild cognitive impairment to Alzheimer's disease. *Diagnostics (Basel)*. 2024;14. <https://doi.org/10.3390/diagnostics14010013>.
 23. Mieling M, Yousuf M, Bunzeck N, et al. Predicting the progression of MCI and Alzheimer's disease on structural brain integrity and other features with machine learning. *GeroScience*. 2026;48:463–87. <https://doi.org/10.1007/s11357-025-01626-5>.
 24. Agostinho D, Simões M, Castelo-Branco M, Alzheimer's Disease Neuroimaging Initiative. Predicting conversion from mild cognitive impairment to Alzheimer's disease: a multimodal approach. *Brain Commun*. 2024;6. <https://doi.org/10.1093/braincomms/fcae208>.
 25. Valsdóttir V, Jónsdóttir MK, Magnúsdóttir BB, et al. Comparative study of machine learning methods for modeling associations between risk factors and future dementia cases. *GeroScience*. 2024;46:737–50. <https://doi.org/10.1007/s11357-023-01040-9>.
 26. Gelir F, Akan T, Alp S, et al. Machine learning approaches for predicting progression to Alzheimer's disease in patients with mild cognitive impairment. *J Med Biol Eng*. 2025;45:63–83. <https://doi.org/10.1007/s40846-024-00918-z>.
 27. Abrol A, Bhattarai M, Fedorov A, et al. Deep residual learning for neuroimaging: an application to predict progression to Alzheimer's disease. *J Neurosci Methods*. 2020;339. <https://doi.org/10.1016/j.jneumeth.2020.108701>.
 28. Xiao R, Cui X, Qiao H, et al. Early diagnosis model of Alzheimer's disease based on sparse logistic regression with the generalized elastic net. *Biomed Signal Process Control*. 2021;66. <https://doi.org/10.1016/j.bspc.2020.102362>.
 29. Luo M, He Z, Cui H, et al. Class activation attention transfer neural networks for MCI conversion prediction. *Comput Biol Med*. 2023;156. <https://doi.org/10.1016/j.combiomed.2023.106700>.

30. Hoang GM, Kim U-H, Kim JG. Vision transformers for the prediction of mild cognitive impairment to Alzheimer's disease progression using mid-sagittal sMRI. *Front Aging Neurosci.* 2023. <https://doi.org/10.3389/fnagi.2023.1102869>.
31. Gao X, Cai H, Liu M. A hybrid multi-scale attention convolution and aging transformer network for Alzheimer's disease diagnosis. *IEEE J Biomed Health Inform.* 2023;27:3292–301. <https://doi.org/10.1109/JBHI.2023.3270937>.
32. Vermeulen RJ, Andersson V, Banken J, et al. Limited generalizability and high risk of bias in multivariable models predicting conversion risk from mild cognitive impairment to dementia: a systematic review. *Alzheimers Dement.* 2025;21. <https://doi.org/10.1002/alz.70069>.
33. Ozsahin DU, Taiwo Mustapha M, Mubarak AS, et al (2022) Impact of feature scaling on machine learning models for the diagnosis of diabetes. In: 2022 International Conference on Artificial Intelligence in Everything (AIE). pp 87–94
34. Pinheiro JMH, de Oliveira SVB, Silva THS, et al. The impact of feature scaling in machine learning: effects on regression and classification tasks. *IEEE Access.* 2025;13:199903–31. <https://doi.org/10.1109/ACCESS.2025.3635541>.
35. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. *J Artif Intell Res.* 2002;16:321–57. <https://doi.org/10.1613/jair.953>.
36. Mustafa S, Li G, Iqbal Y, et al. Tomek links-based SMOTE method for class imbalance in blood cell classification with dual path sliding window attention model. *Talanta.* 2026;303. <https://doi.org/10.1016/j.talanta.2026.129473>.
37. Van Calster B, Vickers AJ. Calibration of risk prediction models: impact on decision-analytic performance. *Med Decis Making.* 2015;35:162–9. <https://doi.org/10.1177/0272989X14547233>.
38. Sahoo R, Zhao S, Chen A, Ermon S (2021) Reliable decisions with threshold calibration. In: Advances in neural information processing systems. Curran Associates, Inc., pp 1831–1844
39. Maghami-Mehr A, Torabi H, Nadeb H, Zhao Y. Linear regression models for analyzing the covariate-adjusted Youden index and associated cut-off points in three diagnostic groups. *J Biopharm Stat.* 2025;0:1–25. <https://doi.org/10.1080/10543406.2025.2558141>.
40. Sun A, Deng Z, Zhao J, et al (2025) Biomarkers selection and combination based on the weighted Youden index
41. Devanand DP, Pradhaban G, Liu X, et al. Hippocampal and entorhinal atrophy in mild cognitive impairment. *Neurology.* 2007;68:828–36. <https://doi.org/10.1212/01.wnl.0000256697.20968.d7>.
42. Persson K, Barca ML, Eldholm RS, et al. Visual evaluation of medial temporal lobe atrophy as a clinical marker of conversion from mild cognitive impairment to dementia and for predicting progression in patients with mild cognitive impairment and mild Alzheimer's disease. *Dement Geriatr Cogn Disord.* 2017;44:12–24. <https://doi.org/10.1159/000477342>.
43. Tripathi M, Tripathi M, Parida GK, Kumar R Biomarker-based prediction of progression to...: *Neurology India.* In: Ovid. <https://www.ovid.com/jnl/neur/fulltext/10.4103/0028-3886.271245~biomarker-based-prediction-of-progression-to-dementia-f-18>. Accessed 2 Aug 2026
44. Lu F, Ma Q, Shi C, et al. Changes in the parietal lobe subregion volume at various stages of Alzheimer's disease and the role in cognitively normal and mild cognitive impairment conversion. *J Integr Neurosci.* 2025. <https://doi.org/10.31083/JIN25991>.
45. Zeng Q, Wang M, Wang E, et al. Sex and APOE genotype specific brain regional vulnerability to Alzheimer's disease. *GeroScience.* 2026. <https://doi.org/10.1007/s11357-025-02089-4>.
46. Morrison C, Oliver MD, Berry V, et al. The influence of APOE status on rate of cognitive decline. *GeroScience.* 2024;46:3263–74. <https://doi.org/10.1007/s11357-024-01069-4>.
47. Godinho F, Maruta C, Borbinha C, Pavão Martins I. Effect of education on cognitive performance in patients with mild cognitive impairment. *Appl Neuropsychol Adult.* 2022;29:1440–9. <https://doi.org/10.1080/23279095.2021.1887191>.
48. Bucholc M, Titarenko S, Ding X, et al. A hybrid machine learning approach for prediction of conversion from mild cognitive impairment to dementia. *Expert Syst Appl.* 2023;217. <https://doi.org/10.1016/j.eswa.2023.119541>.
49. Caminiti SP, De Francesco S, Tondo G, et al. FDG-PET markers of heterogeneity and different risk of progression in amnesic MCI. *Alzheimer's & Dementia.* 2024;20:159–72. <https://doi.org/10.1002/alz.13385>.
50. Silva-Spinola A, Lima M, Leitão MJ, et al. Blood biomarkers in mild cognitive impairment patients: relationship between analytes and progression to Alzheimer disease dementia. *Eur J Neurol.* 2023;30:1565–73. <https://doi.org/10.1111/ene.15762>.
51. Booncharoen K, Hiransuthikul A, Likitjaroen Y, et al. PET and blood biomarkers as predictors of short-term progression in MCI and mild dementia: prospective insights from a Southeast Asian middle-income country. *Alzheimers Dement.* 2025;21. https://doi.org/10.1002/alz70856_100227.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.