International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2026.0412
e ISSN: 2584-2854
Volume: 04 Issue: 10
October 2026
Page No: 2923 - 2939

Explainable Machine Learning for Cross-Sectional Classification of Cognitive Impairment: A Leakage-Safe Study with Independent-Cohort External Validation across OASIS-1 and OASIS-2

Dipali Prakash Patil[1], *Dr. Chandrani Singh*[2]
[1,2]*Sinhgad Institute of Management, Vadgaon (Bk), Pune 411 041, Maharashtra, India*
Emails: *dipalipatil.siom@sinhgad.edu*[1]
ORCID: *D.P.P. 0000-0002-1133-6082; C.S. 0000-0003-0970-2239*

Abstract
Machine learning models built on routinely collected clinical variables have been proposed as a low-cost alternative to neuroimaging for identifying cognitive impairment, but many published studies report optimistic performance because of preprocessing leakage, participant-level leakage, or validation within a single cohort. This study developed and externally validated a leakage-safe machine learning framework for classifying concurrent cognitive impairment, and quantified the contributions of the Mini-Mental State Examination (MMSE) and MRI-derived brain-volume measures. Models were developed on the OASIS-1 cohort (n = 235) and evaluated, without refitting, on the independent baseline OASIS-2 cohort (n = 150), with impairment defined as a Clinical Dementia Rating (CDR) above zero. All preprocessing was fitted only within cross-validated training pipelines, and five classifiers were tuned by grid search. A random forest using age, sex, education, socioeconomic status and the MMSE achieved an external area under the receiver operating characteristic curve (AUC) of 0.920 (95% CI 0.868–0.964) and a Brier score of 0.111. Removing the MMSE reduced the AUC to 0.749. Compared with the conventional MMSE < 24 threshold, which identified only 32% of impaired participants, the model achieved 80.0% sensitivity at 91.8% specificity. Adding brain-volume measures raised the external AUC to 0.963. Baseline clinical and imaging variables did not distinguish participants who later converted to impairment. Under leakage-safe development and independent-cohort validation, the framework is best suited as a calibrated case-finding aid rather than an early prediction tool.

Keywords: *Cognitive impairment; Data leakage; Dementia screening; Explainable machine learning; External validation.*

1. Introduction
More than 55 million people are currently living with dementia worldwide, and this number is projected to rise to approximately 139 million by 2050, with a significant increase in low- and middle-income countries (LMICs) [1,2]. Early identification is essential because therapeutic and supportive interventions are most effective before substantial functional decline occurs. However, cognitive impairment is frequently recognized only after noticeable deterioration in daily functioning, limiting opportunities for timely intervention. Widely used cognitive screening instruments, including the Mini-Mental State Examination (MMSE) [3] and the Montreal Cognitive Assessment (MoCA) [4], are inexpensive, rapid, and extensively validated for clinical use. Nevertheless, their diagnostic sensitivity for detecting mild cognitive impairment is limited, and their performance can be influenced by educational attainment, language, and cultural background. Structural neuroimaging can substantially improve diagnostic accuracy by identifying disease-related brain changes [5,6]. However, the high cost of imaging, limited availability of scanners, and inadequate healthcare infrastructure, particularly in LMICs, restrict its use as a practical first-line

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2026.0412
e ISSN: 2584-2854
Volume: 04 Issue: 10
October 2026
Page No: 2923 - 2939

screening tool for large-scale population-based assessment.

1.1.The Reliability Problem in This Literature

Machine learning models based on routinely collected clinical data have emerged as a promising intermediate approach, offering greater diagnostic capability than conventional cognitive screening while avoiding the cost and limited accessibility of neuroimaging. Although numerous studies report excellent predictive performance, many estimates are difficult to interpret because of methodological shortcomings that artificially inflate model accuracy. Three issues are particularly common. First, preprocessing leakage occurs when data transformations, such as scaling or imputation, are performed before the training–testing split, allowing information from the evaluation set to influence model development [7]. Second, participant-level leakage arises in longitudinal datasets when multiple visits from the same individual are distributed across both training and testing sets, violating the assumption of independent evaluation. Third, and perhaps most importantly, many studies rely exclusively on random train–test splits within a single cohort, providing little evidence that a model will generalize to an independent population. An additional challenge, specific to cognitive screening, has received comparatively little attention. Clinical variables such as the Mini-Mental State Examination (MMSE) are frequently used as predictor variables, even though cognitive assessments also contribute to establishing the diagnostic reference standard. Consequently, high predictive performance may simply reflect the model learning to reproduce an existing clinical decision rule rather than identifying novel disease-related patterns. This does not necessarily diminish the model's clinical value; a well-calibrated machine learning model may still improve the consistency and scalability of screening. However, its utility should be assessed by comparing its performance directly with the established screening rule rather than solely against alternative machine learning algorithms.

1.2.Related Work

The Mini-Mental State Examination (MMSE), introduced by Folstein et al. in 1975 [3], has remained the most widely adopted instrument for cognitive screening because of its simplicity, brevity, and extensive clinical validation. The subsequent development of the Montreal Cognitive Assessment (MoCA) addressed some of the MMSE's limitations by improving sensitivity to mild cognitive impairment [4]. Both assessments generate standardized numerical scores that are easily integrated into statistical and machine learning models. However, this convenience introduces an important methodological concern that has received relatively limited attention. Cognitive test scores are frequently incorporated as predictor variables, while the reference diagnosis is simultaneously established using cognitive assessments, creating a potential source of label–predictor dependency that may artificially enhance model performance. Consequently, models may primarily reproduce established clinical decision rules rather than discover novel diagnostic information. The availability of openly accessible datasets has substantially accelerated methodological research in dementia prediction. The Open Access Series of Imaging Studies (OASIS), initially released as a cross-sectional dataset by Marcus et al. [8] and later expanded into a longitudinal cohort [9], provides structural MRI alongside comprehensive demographic and clinical assessments, making it one of the most extensively studied resources for Alzheimer's disease research. Classical machine learning algorithms have demonstrated encouraging performance on these data. Early work employing support vector machines (SVMs) achieved high accuracy for distinguishing patients with Alzheimer's disease from cognitively healthy controls; however, performance declined markedly when models were applied to the clinically more challenging task of predicting future conversion from normal cognition or mild impairment [6]. This distinction between accurately classifying current disease status and reliably forecasting future cognitive decline has remained one of the most consistent findings in

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2026.0412
e ISSN: 2584-2854
Volume: 04 Issue: 10
October 2026
Page No: 2923 - 2939

dementia prediction research and underscores the need to evaluate these objectives separately. Tree-based ensemble methods have subsequently become prominent for structured clinical datasets because of their robustness to nonlinear relationships, interactions among predictors, and mixed variable types. Random forests [10] and gradient-boosted decision trees [11] consistently demonstrate competitive discrimination while providing measures of feature importance, with the MMSE frequently emerging as the dominant predictor of cognitive impairment [12]. In parallel, deep learning approaches applied to structural neuroimaging have achieved state-of-the-art diagnostic performance by automatically learning hierarchical anatomical representations from volumetric MRI data [13, 14]. Despite these advances, neuroimaging-based models require specialized imaging infrastructure, substantial computational resources, and expert interpretation, limiting their scalability and practical deployment in primary care and resource-constrained healthcare settings. Recent methodological guidance has shifted the emphasis of clinical artificial intelligence from maximizing predictive accuracy toward demonstrating robustness, reproducibility, and clinical applicability. International reporting frameworks for prediction models advocate rigorous external validation, assessment of calibration, transparent model development, and comprehensive reporting to facilitate reproducibility and clinical translation [15–19]. At the same time, an expanding methodological literature has demonstrated that preprocessing leakage, participant-level leakage, and spectrum bias can substantially inflate reported diagnostic performance, leading to overly optimistic estimates of generalizability [7, 20]. Beyond methodological rigor, there is increasing recognition that model interpretability alone is insufficient to ensure safe clinical deployment. Contemporary frameworks advocate algorithmovigilance, encompassing continuous post-deployment performance monitoring, data-drift detection, bias surveillance, calibration assessment, and structured incident reporting to ensure that predictive systems remain reliable as clinical populations and practice environments evolve [21]. These principles provide the methodological foundation for the present study and motivate the deployment considerations discussed in Section 3.2.4.

1.3.Objectives

This study was designed to address the principal methodological limitations identified in previous machine learning studies for cognitive impairment screening. The proposition and evaluation of a framework is presented that eliminates information leakage, avoids participant-level contamination, validates performance on an independent cohort, and critically examines the contribution of cognitive screening variables to model performance. Accordingly, the study addresses the following objectives: O1. To develop a leakage-safe machine learning framework by ensuring that all preprocessing and model training are performed exclusively within the training data, thereby preventing information leakage. O2. To evaluate the generalizability of the developed models through external validation on an independent cohort without model refitting. O3. To quantify the contribution of the Mini-Mental State Examination (MMSE) by comparing model performance with and without MMSE and benchmarking the models against the conventional MMSE screening threshold. O4. To determine whether routinely available MRI-derived brain-volume measures provide additional predictive value beyond clinical and demographic variables and to examine whether baseline characteristics can identify individuals who subsequently develop cognitive impairment.

1.4.Relationship to Prior Work by the Authors

An earlier version of this research, conducted by the same authors, explored machine learning for cognitive impairment detection using a combined dataset derived from two repositories. That study employed a conventional random train–test split and preprocessing procedures that were performed prior to data partitioning. In addition, one predictor variable was included despite its absence in one of the contributing cohorts. These methodological

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2026.0412
e ISSN: 2584-2854
Volume: 04 Issue: 10
October 2026
Page No: 2923 - 2939

choices introduced the potential for information leakage and limited the validity of performance estimates. The present study substantially extends and improves upon the previous work through a rigorously designed evaluation framework. All preprocessing operations are performed exclusively within cross-validated training pipelines, thereby preventing information leakage. Model development is undertaken using the OASIS-1 cohort and externally validated, without refitting, on the independent baseline OASIS-2 cohort to assess generalizability across populations. Furthermore, only predictor variables consistently available across both cohorts are included, ensuring methodological consistency and reproducibility. As a consequence of these more stringent methodological safeguards, the predictive performance reported in this study is more conservative than that of the earlier work. However, these estimates provide a more realistic assessment of model performance under conditions that better reflect clinical deployment. The principal contribution of this study is therefore not an increase in predictive accuracy but the establishment of a transparent, leakage-resistant, externally validated framework for developing and evaluating machine learning models for cognitive impairment screening.

1. Method

1.1.Study Design and Reporting

This is a diagnostic model development and external validation study using two independently collected, openly licensed cohorts. Development and external validation were performed on separate cohorts, with no refitting on the validation data at any stage. The study is reported in accordance with the TRIPOD+AI statement, which provides reporting guidance for clinical prediction models developed using regression or machine learning methods [15] and updates the original TRIPOD statement [16,18]; external validation follows current guidance on the conduct and interpretation of such analyses [19]. The completed TRIPOD+AI checklist, indicating the page on which each item is addressed, is provided as Supplementary Table S3.

1.2.Data Sources

Two openly licensed cohorts from the Open Access Series of Imaging Studies were used. The development cohort was OASIS-1 (Cross-Sectional) [8], comprising 416 participants aged 18 to 96 across 436 imaging sessions. The external-validation cohort was OASIS-2 (Longitudinal) [9], an independent collection of 150 participants aged 60 to 96 across 373 imaging sessions. The two cohorts were collected under different protocols and sampling frames, which is what makes validation across them informative rather than circular. Both were obtained from the OASIS project in January 2026, using the tabular demographic and clinical summary files distributed with each collection. Unlike OASIS-3, these two collections are distributed as single fixed releases and carry no version identifier. The authors registered for access, agreed to the OASIS Data Use Agreement, and complied with its terms; no attempt was made to establish the identity of, or make contact with, any participant.

From each cohort, we used age, sex, years, coded level of education, socioeconomic status, MMSE, the Clinical Dementia Rating (CDR) [22], and the MRI-derived measures of estimated total intracranial volume (eTIV) and normalized whole-brain volume (nWBV). Neither cohort contains the Montreal Cognitive Assessment, so that instrument is not used anywhere in this study.

1.3.Outcome Definition

Cognitive impairment was defined as a CDR greater than zero, grouping very mild, mild, and moderate stages against cognitively normal participants (CDR of zero). The CDR, rather than the MMSE, defines the outcome. This is a deliberate design choice: the CDR is a clinician-rated staging instrument based on a structured interview covering memory, orientation, judgment, and daily function, and using it as the label so that the outcome is partially separable from the MMSE used as a predictor. We nonetheless acknowledge in Section 3.2.6 that the CDR is not fully independent of the cognitive constructs the MMSE measures, and Section 3.1.4 quantifies the resulting entanglement.

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2026.0412
e ISSN: 2584-2854
Volume: 04 Issue: 10
October 2026
Page No: 2923 - 2939

1.4.Eligibility and One Record per Participant

In OASIS-1, 201 participants without a CDR assessment, predominantly younger adults enrolled for other reasons, were excluded, leaving 235 participants each contributing a single record. In OASIS-2, only the baseline visit of each participant was retained, giving 150 participants each contributing a single record and making participant-level leakage impossible by construction. The 223 follow-up visits were reserved for the descriptive conversion analysis in Section 3.1.10. The full participant flow is shown in Figure 1.

2. Method

2.1.Study Design and Reporting

This is a diagnostic model development and external validation study using two independently collected, openly licensed cohorts. Development and external validation were performed on separate cohorts, with no refitting on the validation data at any stage. The study is reported in accordance with the TRIPOD+AI statement, which provides reporting guidance for clinical prediction models developed using regression or machine learning methods [15] and updates the original TRIPOD statement [16,18]; external validation follows current guidance on the conduct and interpretation of such analyses [19]. The completed TRIPOD+AI checklist, indicating the page on which each item is addressed, is provided as Supplementary Table S3.

2.2.Data Sources

Two openly licensed cohorts from the Open Access Series of Imaging Studies were used. The development cohort was OASIS-1 (Cross-Sectional) [8], comprising 416 participants aged 18 to 96 across 436 imaging sessions. The external-validation cohort was OASIS-2 (Longitudinal) [9], an independent collection of 150 participants aged 60 to 96 across 373 imaging sessions. The two cohorts were collected under different protocols and sampling frames, which is what makes validation across them informative rather than circular. Both were obtained from the OASIS project in January 2026, using the tabular demographic and clinical summary files distributed with each collection. Unlike OASIS-3, these two collections are distributed as single fixed releases and carry no version identifier. The authors registered for access, agreed to the OASIS Data Use Agreement, and complied with its terms; no attempt was made to establish the identity of, or make contact with, any participant.

From each cohort, we used age, sex, years, coded level of education, socioeconomic status, MMSE, the Clinical Dementia Rating (CDR) [22], and the MRI-derived measures of estimated total intracranial volume (eTIV) and normalized whole-brain volume (nWBV). Neither cohort contains the Montreal Cognitive Assessment, so that instrument is not used anywhere in this study.

2.3.Outcome Definition

Cognitive impairment was defined as a CDR greater than zero, grouping very mild, mild, and moderate stages against cognitively normal participants (CDR of zero). The CDR, rather than the MMSE, defines the outcome. This is a deliberate design choice: the CDR is a clinician-rated staging instrument based on a structured interview covering memory, orientation, judgment, and daily function, and using it as the label so that the outcome is partially separable from the MMSE used as a predictor. We nonetheless acknowledge in Section 3.2.6 that the CDR is not fully independent of the cognitive constructs the MMSE measures, and Section 3.1.4 quantifies the resulting entanglement.

2.4.Eligibility and One Record per Participant

In OASIS-1, 201 participants without a CDR assessment, predominantly younger adults enrolled for other reasons, were excluded, leaving 235 participants each contributing a single record. In OASIS-2, only the baseline visit of each participant was retained, giving 150 participants each contributing a single record and making participant-level leakage impossible by construction. The 223 follow-up visits were reserved for the descriptive conversion analysis in Section 3.1.10. The full participant flow is shown in Figure 1. As shown in Figure 1 Participant flow for the development (OASIS-1) and external-validation (OASIS-2) cohorts. Each participant

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2026.0412
e ISSN: 2584-2854
Volume: 04 Issue: 10
October 2026
Page No: 2923 - 2939

contributes exactly one record to the primary analysis, eliminating participant-level leakage by construction.

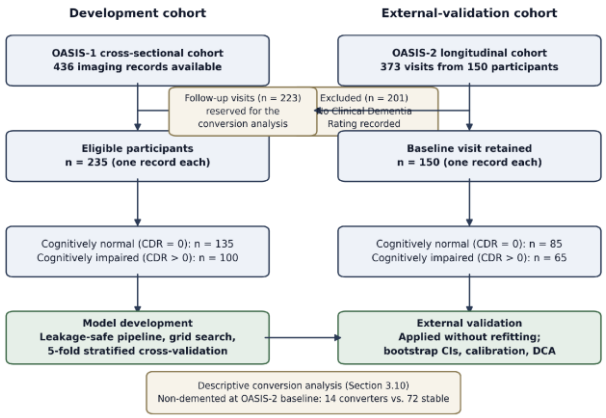


Figure 1 Participant flow for the development (OASIS-1) and external-validation (OASIS-2) cohorts. Each participant contributes exactly one record to the primary analysis, eliminating participant-level leakage by construction

2.5.Feature Harmonisation and a Documented Discrepancy

Age, sex, socioeconomic status, MMSE, and the two brain-volume measures were recorded on compatible scales across the cohorts. Education was not: OASIS-1 records education as an ordinal category from one to five, whereas OASIS-2 records completed years of schooling. Treating these as the same numeric variable would be an error. We therefore mapped OASIS-2 years onto the OASIS-1 ordinal scale using the published category definitions (below high school, high-school graduate, some college, college graduate, and beyond college). This mapping is stated explicitly so that it can be scrutinized and reproduced, and we flag education as the least reliably harmonized feature in the study.

2.6.Leakage-Safe Preprocessing Pipeline

All preprocessing was encapsulated inside a single estimator pipeline. Median imputation for numeric variables, most-frequent imputation and one-hot encoding for categorical variables, and standardization were fitted on the development cohort only and, within cross-validation, refitted from scratch on each training fold. The external cohort was transformed using parameters fitted on the development data and never contributed to them. No scaler or imputer ever saw the validation cohort during fitting.

For example, in the proposed pipeline, the median age and scaling parameters are computed only from the training data.These learned values are then applied unchanged to the validation fold during cross-validation and to the independent OASIS-2 cohort during external validation. Similarly, categorical variables such as sex or socioeconomic status are one-hot encoded using categories identified exclusively in the training data. This design ensures that the validation cohort remains completely unseen throughout preprocessing and model development, thereby providing an unbiased estimate of predictive performance.

2.7.Models and Hyperparameter Tuning

Five classifiers were evaluated: L2-regularized logistic regression, a single decision tree, a support vector machine with a radial basis function kernel, a random forest [10], and a gradient-boosted tree ensemble [11]. The decision tree and logistic regression are included as deliberate minimal-complexity comparators. Hyperparameters were selected by grid search under five-fold stratified cross-validation on the development cohort, optimizing the area under the receiver operating characteristic curve (AUC). The random seed was fixed at 42 throughout. All models were implemented in scikit-learn [23], and package versions are pinned in the released code.

2.8.Experiments

To evaluate the proposed leakage-resistant and externally validated framework, five experiments were pre-specified.

E1: Develop the clinical model using the OASIS-1 development cohort and evaluate its performance, without model refitting, on the independent OASIS-2 cohort. This experiment provides the primary assessment of external generalizability under the leakage-safe pipeline.

E2: Reverse the development and validation cohorts (OASIS-2 → OASIS-1) to examine whether the framework produces consistent results irrespective of cohort direction, thereby assessing

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2026.0412
e ISSN: 2584-2854
Volume: 04 Issue: 10
October 2026
Page No: 2923 - 2939

the robustness of external validation.

E3: Remove the MMSE from the predictor set and retain only demographic and socioeconomic variables to quantify the independent contribution of the MMSE and assess model performance without cognitive test information.

E4: Augment the clinical model with routinely available MRI-derived brain-volume measures to determine whether imaging variables provide incremental predictive value beyond clinical information.

E5: Evaluate a model using MRI-derived brain-volume measures together with demographic and socioeconomic variables, excluding the MMSE, to investigate the complementary role of structural imaging in cognitive impairment classification. Finally, the primary clinical model was benchmarked against the conventional MMSE cut-off rule applied directly to the independent external cohort to determine whether the machine learning framework provides meaningful improvement over standard clinical screening.

2.9.Evaluation and Statistical Analysis

The evaluation strategy was designed to correspond directly to the five pre-specified experiments. For all externally validated models (E1–E5), predictive performance was evaluated on the independent validation cohort using accuracy, balanced accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), F1-score, area under the receiver operating characteristic curve (AUC), area under the precision–recall curve (AUPRC), and the Brier score. Ninety-five percent confidence intervals were estimated using 2,000 bootstrap resamples of the external cohort.

2.9.1. External validation and model comparison

Experiments E1 and E2 evaluated the transportability and robustness of the proposed leakage-safe framework by developing models on one cohort and validating them, without model refitting, on the independent cohort. The consistency of predictive performance across both directions of external validation was examined to assess the robustness of the framework.

Experiments E3, E4, and E5 quantified the contribution of different predictor sets. Specifically, E3 evaluated the effect of excluding the MMSE, E4 assessed the incremental value of MRI-derived brain-volume measures when added to the clinical model, and E5 investigated the predictive ability of MRI-derived measures in the absence of MMSE. Differences in discrimination between competing models were evaluated using paired bootstrap resampling. When the 95% confidence interval for the AUC difference included zero, the models were considered statistically indistinguishable regardless of their point estimates.

2.9.2. Calibration and clinical utility

Model calibration was evaluated using calibration (reliability) curves and the Brier score [17, 26]. Clinical usefulness was assessed using Decision Curve Analysis (DCA) [25], which estimates the net clinical benefit across a range of probability thresholds. To determine whether the proposed framework provides clinically meaningful improvement over current practice, the primary clinical model (E1) was additionally benchmarked against the conventional MMSE cut-off rule applied directly to the external validation cohort. The Brier score is the mean squared difference between the predicted probability and the observed outcome:

$$BS = \frac{1}{N}\sum_{i=1}^{N}(p_i - y_i)^2 \qquad (1)$$

where N is the number of participants, pi is the predicted probability of impairment for participant i, and yi is the observed outcome (1 = impaired, 0 = unimpaired); lower values indicate better probability estimates. The net benefit used in decision-curve analysis at a threshold probability pt is

$$NB = \frac{TP}{N} - \frac{FP}{N} \times \frac{p_t}{1-p_t} \qquad (2)$$

where TP and FP are the numbers of true-positive and false-positive classifications at threshold pt.

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2026.0412
e ISSN: 2584-2854
Volume: 04 Issue: 10
October 2026
Page No: 2923 - 2939

2.9.3. Model interpretability

To improve transparency, model predictions were interpreted using SHapley Additive exPlanations (SHAP) [24]. SHAP values were computed using the development cohort as the background distribution; both global feature importance and representative patient-level SHAP explanations were generated according to a pre-specified interpretation protocol.

2.9.4. Operating points

Unless otherwise specified, sensitivity, specificity, PPV, and NPV were calculated using the default classification threshold of 0.50, which was fixed before model evaluation and was not optimized using the validation data. Because screening applications generally prioritize minimizing false negatives, an additional sensitivity-oriented operating point was explored by selecting the highest threshold, achieving at least 90% sensitivity on the external cohort. As this threshold was determined using evaluation data, the resulting performance is reported as an exploratory analysis rather than a pre-specified deployment threshold.

2.9.5. Subgroup analysis

To examine the robustness of model performance across clinically relevant populations, subgroup analyses were conducted on the external validation cohort. Participants were stratified by sex, age (<75 years versus ≥75 years), and education level (≤ category 3 versus > category 3 on the OASIS-1 ordinal education scale). Within each subgroup, AUC, sensitivity, and specificity were estimated together with bootstrap-derived 95% confidence intervals. These analyses were considered descriptive because the study was not powered to detect statistically significant subgroup differences.

Longitudinal descriptive analysis. Although the primary objective was cross-sectional classification, a descriptive longitudinal analysis was performed using participants in the external cohort who were cognitively normal at baseline but subsequently progressed to cognitive impairment. Baseline demographic, cognitive, and imaging characteristics of future converters were compared with those of participants who remained cognitively stable using appropriate descriptive statistics. This analysis was intended to determine whether the proposed cross-sectional framework also demonstrated prognostic capability and to clarify the distinction between concurrent classification and early detection.

3. Results And Discussion

3.1.Results

3.1.1. Cohort Characteristics

The two cohorts were similar in the prevalence of impairment (43% in each) and in median MMSE but differed in age structure and sex balance, consistent with their different sampling frames (Table 1). The similar class balance is convenient but incidental; the validation remains a genuine test of transport because the cohorts were recruited independently. As shown in Table 1. Characteristics of the development and external-validation cohorts.

Table 1 Characteristics of the development and external-validation cohorts

Characteristic	OASIS-1 (development)	OASIS-2 baseline (external)
Participants (n)	235	150
Cognitively normal (CDR = 0)	135	85
Cognitively impaired (CDR > 0)	100	65
Impaired (%)	42.6	43.3
Age, median [IQR]	73 [66–81]	75 [70–80]
Female (%)	66.4	58.7
MMSE, median [IQR]	29 [26–30]	29 [26–30]
Education (ordinal), median	3	3
SES missing (%)	8.1	5.3
nWBV, mean (SD)	0.749 (0.048)	0.736 (0.037)

CDR, Clinical Dementia Rating; IQR, interquartile

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2026.0412
e ISSN: 2584-2854
Volume: 04 Issue: 10
October 2026
Page No: 2923 - 2939

range; MMSE, Mini-Mental State Examination; nWBV, normalised whole-brain volume; SES, socioeconomic status.

3.1.2. Primary External Validation (E1)

Trained on OASIS-1 and applied unchanged to OASIS-2, the clinical model separated impaired from unimpaired participants well (Table 2, Figure 2). The random forest reached the highest external AUC at 0.920 (95% CI 0.868–0.964), ahead of the decision tree at 0.898 (0.841–0.949) and gradient boosting at 0.892 (0.829–0.944). A paired bootstrap separated the random forest from gradient boosting (ΔAUC +0.028, 95% CI +0.011 to +0.047), and the random forest is therefore carried forward as the reference clinical model. At the pre-fixed classification threshold of 0.5 it ran at 0.800 sensitivity and 0.918 specificity, with a Brier score of 0.111. Internal cross-validated AUC on OASIS-1 (0.89–0.92) sat close to the external figure—little sign of optimism, and a good omen for transport. Because a screening instrument is normally operated at whatever threshold delivers acceptable sensitivity rather than at the classifier default, we additionally report performance at a sensitivity-prioritising operating point, defined as the highest threshold achieving at least 0.90 sensitivity on the external cohort. At a threshold of 0.28, the highest achieving at least 0.90 sensitivity, the model ran at 0.908 sensitivity and 0.635 specificity (PPV 0.656, NPV 0.900). Both operating points, with their thresholds and full operating characteristics, are reported in Supplementary Table S4. The comparison makes the trade-off explicit: the additional cases identified at the lower threshold are purchased with a loss of specificity, and the choice between the two operating points is a clinical decision about referral capacity and the tolerable false-positive burden rather than a statistical one. As shown in Table 2. External-validation performance on OASIS-2 for models developed on OASIS-1 (95% bootstrap CIs). As shown in FIGURE 2. Receiver operating characteristic curves on the independent external cohort (OASIS-2) for models developed on OASIS-1 without refitting. Acc., accuracy; Sens., sensitivity; Spec., specificity; AUC, area under the receiver operating characteristic curve; RBF, radial basis function; SVM, support vector machine. Confusion matrices, PPV, NPV, balanced accuracy, precision–recall AUC and full confidence intervals for every metric are given in Supplementary Table S1.

Table 2 External-validation performance on OASIS-2 for models developed on OASIS-1 (95% bootstrap CIs)

Model	AUC (95% CI)	Acc.	Sens.	Spec.	Brier
Logistic regression	0.865 (0.799–0.927)	0.840	0.723	0.929	0.130
Decision tree	0.898 (0.847–0.952)	0.860	0.723	0.965	0.123
SVM (RBF)	0.875 (0.812–0.932)	0.860	0.769	0.929	0.125
Random forest	**0.920 (0.868–0.964)**	**0.867**	**0.800**	**0.918**	**0.111**
Gradient boosting	0.892 (0.829–0.944)	0.860	0.723	0.965	0.122

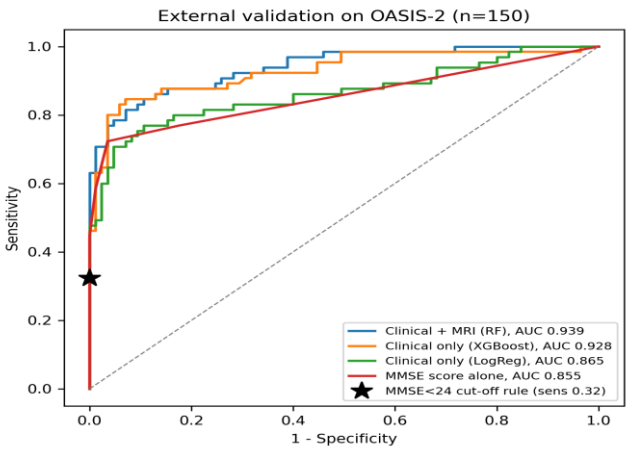


Figure 2 Receiver operating characteristic curves on the independent external cohort (OASIS-2) for models developed on OASIS-1 without refitting

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2026.0412
e ISSN: 2584-2854
Volume: 04 Issue: 10
October 2026
Page No: 2923 - 2939

3.1.3. Reverse-Direction Validation (E2)

Reversing the roles—training on OASIS-2 and testing on OASIS-1—gave an external AUC of 0.916 (95% CI 0.877–0.950) with gradient boosting, with the random forest close behind at 0.884 (0.835–0.926). Performance is therefore not an accident of one particular direction of transfer, although the smaller development cohort in this direction gives correspondingly less stable estimates. Results for all five classifiers in this direction are reported in Supplementary Table S5.

3.1.4. How Much of the Signal Is the MMSE? (E3)

Removing the MMSE, leaving age, sex, education, and socioeconomic status, dropped external AUC to 0.749 (95% CI 0.665–0.823) for the best-performing model in this configuration (support vector machine), against 0.920 (0.868–0.964) with the full feature set; the gap was large and clearly significant (ΔAUC +0.171, 95% CI +0.093 to +0.257). The MMSE therefore carries the bulk of the model. We expected this, and it needs reading with care. Because the MMSE is entangled with the CDR staging that sets the label, the correct interpretation is that the model is efficiently reusing and extending the test—not that it has found some independent signal buried in demographics.

3.1.5. Does the Model Beat the Clinical Cut-Off Rule?

For a screening tool the honest benchmark is the rule it would replace, not a lineup of other algorithms. Applied directly to the external cohort, the familiar MMSE-below-24 rule never misfired (specificity 1.00, PPV 1.00) but found only 32% of impaired participants (sensitivity 0.323, 95% CI 0.212–0.433)—it walked past two cases in three. Loosening the cut to below 27 lifted sensitivity only to 0.585. The model, by contrast, ran at 0.800 sensitivity and 0.918 specificity and out-discriminated even the raw MMSE score used as a ranker (ΔAUC +0.064, 95% CI +0.024 to +0.111). Its value is therefore easy to state: it recovers much of the sensitivity a fixed threshold throws away and pays only modestly in specificity. When a positive screen leads to further assessment rather than to treatment, that is usually a trade worth making—as the decision curves in Section 3.1.7 bear out. As shown in Table 3. The model versus conventional MMSE cut-off rules on the external cohort.

Table 3 The model versus conventional MMSE cut-off rules on the external cohort

Rule / model	Sens.	Spec.	PPV	NPV	AUC
MMSE < 24	0.323	1.000	1.000	0.659	0.855
MMSE < 26	0.446	1.000	1.000	0.702	0.855
MMSE < 27	0.585	0.988	0.974	0.757	0.855
Clinical model (random forest)	**0.800**	**0.918**	**0.881**	**0.857**	**0.920**

Sens., sensitivity; Spec., specificity; NPV, negative predictive value; PPV, positive predictive value. The AUC of 0.855 for the cut-off rows refers to the raw MMSE score used as a continuous ranker and is therefore identical across thresholds.

3.1.6. Does MRI Add Value? (E4, E5)

Adding the two MRI-derived brain-volume measures to the clinical model raised external AUC to 0.963 (95% CI 0.932–0.987) with the random forest, a small but statistically significant improvement over the clinical model (ΔAUC +0.043, 95% CI +0.014 to +0.078), and improved sensitivity to 0.892 at a specificity of 0.918. Imaging measures alone, without the MMSE, reached an AUC of only about 0.73–0.76, below the clinical questionnaire. The practical reading is that a five-minute cognitive questionnaire captures most of the available cross-sectional signal, that structural imaging adds a real but incremental amount, and that imaging is not a substitute for the clinical assessment in this setting.

3.1.7. Calibration and Clinical Utility

The clinical models were reasonably well calibrated on the external cohort: the reference random-forest clinical model had a Brier score of 0.111, and predicted probabilities tracked observed frequencies across the range. Figure 3 shows the

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2026.0412
e ISSN: 2584-2854
Volume: 04 Issue: 10
October 2026
Page No: 2923 - 2939

calibration curves for the gradient-boosting (XGBoost) clinical model (Brier score 0.101) and the random-forest clinical + MRI model (Brier score 0.109). Decision-curve analysis (Figure 4) showed positive net benefit above both default strategies across the range of threshold probabilities relevant to screening, and exceeded the net benefit of the MMSE cut-off rule across most of that range, driven by the models' recovery of sensitivity. The addition of MRI did not materially change net benefit, consistent with its small increment in discrimination.

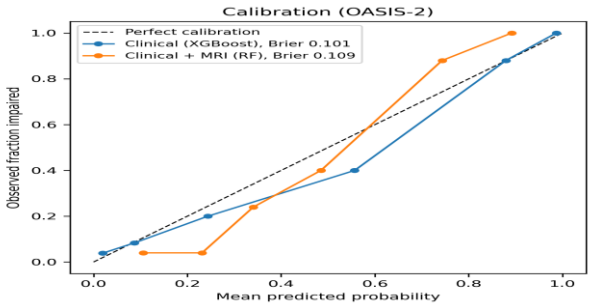


Figure 3 Calibration on the external cohort (OASIS-2) of the gradient-boosting (XGBoost) clinical model and the random-forest clinical + MRI model. Predicted probabilities are plotted against observed frequencies; the diagonal denotes perfect calibration

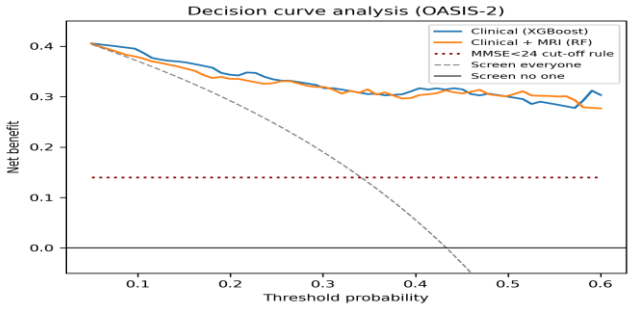


Figure 4 Decision-curve analysis on the external cohort (OASIS-2), comparing the gradient-boosting (XGBoost) clinical model and the random-forest clinical + MRI model with the MMSE < 24 cut-off rule and the screen-everyone and screen-no-one strategies

3.1.8. Model Explainability

SHAP analysis of the reference random-forest model ranked the MMSE as by far the most influential feature, followed at a distance by age, with sex, education, and socioeconomic status contributing little (Figure 5). The beeswarm representation additionally shows the direction of effect: low MMSE values push the prediction towards impairment, as expected. This ordering is consistent with the E3 result and, as discussed in Section 3.2.2, is close to a tautology rather than corroboration of clinical validity, because the MMSE is bound up with the outcome definition. The individual-case waterfall plot (Figure 6) is shown for the external participant whose predicted probability lay closest to the median, a rule fixed in advance so that the figure cannot be read as a curated example. As shown in Figure 5 Global SHAP feature importance for the reference random-forest clinical model, expressed as mean absolute SHAP value across the external cohort.

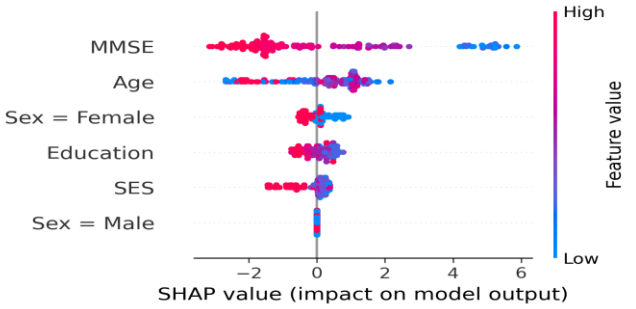


Figure 5 Global SHAP feature importance for the reference random-forest clinical model, expressed as mean absolute SHAP value across the external cohort

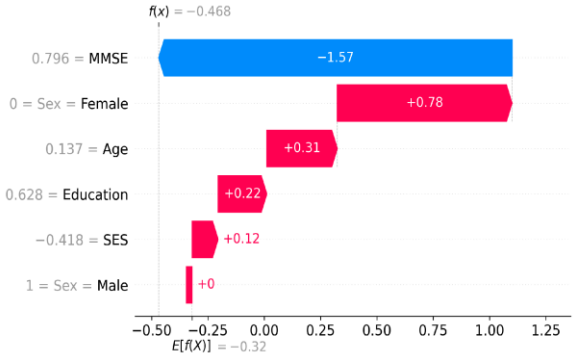


Figure 6 Individual-case SHAP explanation for the external participant whose predicted probability was closest to the median, selected by a rule fixed in advance

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2026.0412
e ISSN: 2584-2854
Volume: 04 Issue: 10
October 2026
Page No: 2923 - 2939

3.1.9. Subgroup Performance

Discrimination within pre-specified subgroups of the external cohort is shown in Table 4. Performance was broadly similar across sex and age bands, with overlapping confidence intervals throughout. Two caveats govern how Table 4 should be read. First, subgroup sample sizes are small—between roughly 50 and 100 participants, with correspondingly few impaired cases—so the confidence intervals are wide and largely overlapping. Second, the study was not powered to detect differential performance, and no formal test of interaction was pre-specified. These analyses are therefore descriptive rather than confirmatory: they are reported because differential performance across education and language strata is a foreseeable failure mode for a model dominated by an education-sensitive instrument, and because a screening tool that performs unevenly across such strata would require that fact to be established before deployment, not after. They should not be read as evidence either that performance is equitable or that it is not.

Table 4 Discrimination and operating characteristics within pre-specified subgroups of the external cohort (OASIS-2) at the fixed classification threshold of 0.5. Values are point estimates with 95% bootstrap confidence intervals; analyses are descriptive and not powered to detect differential performance

Subgroup	n	Impaired (n)	AUC (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)
All participants	150	65	0.920 (0.868–0.964)	0.800	0.918
Female	88	29	0.934 (0.869–0.983)	0.759 (0.591–0.905)	0.983 (0.942–1.000)
Male	62	36	0.896 (0.809–0.965)	0.833 (0.703–0.943)	0.769 (0.593–0.920)
Age < 75 years	70	32	0.883 (0.782–0.957)	0.719 (0.538–0.871)	0.921 (0.829–1.000)
Age ≥ 75 years	80	33	0.965 (0.922–0.994)	0.879 (0.758–0.973)	0.915 (0.824–0.981)
Education ≤ 3	83	45	0.926 (0.862–0.974)	0.800 (0.681–0.913)	0.921 (0.829–1.000)
Education > 3	67	20	0.899 (0.770–0.988)	0.800 (0.600–0.955)	0.915 (0.827–0.980)

Because both cohorts contribute only one source each to the primary analysis, performance by cohort source is given by the two directions of transfer reported in Sections 3.1.2 and 3.1.3 rather than as a subgroup of a pooled sample.

3.1.10. Cross-Sectional Classification Is Not Early Detection

Fourteen OASIS-2 participants who were cognitively normal at baseline converted to a worse category during follow-up. At baseline these future converters were indistinguishable from the 72 participants who remained stable: neither baseline MMSE ($p = 0.55$) nor baseline normalized whole-brain volume ($p = 0.42$) differed significantly between the groups (Figure 7). The strong cross-sectional performance reported above therefore does not translate into an ability to anticipate decline, and it should not be presented as early detection. The small converter group precludes building a prognostic model here; it is reported descriptively to make the boundary of the claim explicit and to motivate adequately powered longitudinal work. As shown in FIGURE 7. Baseline MMSE and normalized whole-brain volume do not separate future converters ($n = 14$) from stable participants ($n = 72$) in OASIS-2.

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2026.0412
e ISSN: 2584-2854
Volume: 04 Issue: 10
October 2026
Page No: 2923 - 2939

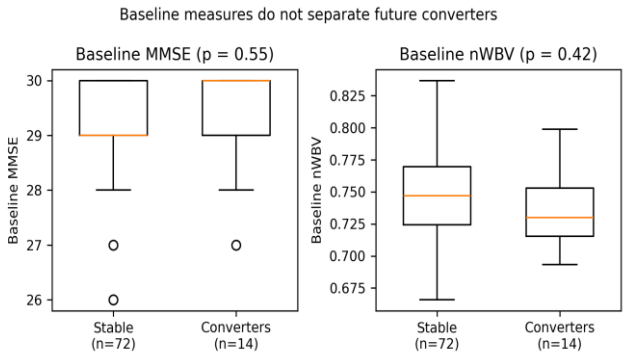


Figure 7 Baseline MMSE and normalized whole-brain volume do not separate future converters (n = 14) from stable participants (n = 72) in OASIS-2.

3.2.Discussion

3.2.1. Principal Findings

The principal finding of this study is that a transparent, leakage-resistant, and externally validated machine learning framework can achieve robust cross-sectional classification of cognitive impairment using routinely available clinical variables. Under this framework, the reference clinical model achieved an external AUC of 0.92 while maintaining good calibration and consistent performance across both directions of external validation. These findings indicate that the observed performance is reproducible across independently collected cohorts rather than being an artefact of information leakage or favourable data partitioning.

3.2.2. What the Model Is, and Is Not, Doing

Both the demographics-only experiment and the SHAP ranking point the same way: the model's power is mostly the MMSE. We see no reason to obscure this, but it does fence in what can be claimed. We have not uncovered a hidden signal in age and education. We have shown that folding the MMSE into a calibrated model, with a few routine variables alongside it, extracts considerably more usable sensitivity than a bare threshold—and that this holds up under an evaluation strict enough to believe. On that reading the model is a case-finding aid, worth having for its calibration and its recovered sensitivity, not a new biomarker. SHAP values describe model associations; they do not establish causal relevance, clinical validity, or independence from the diagnostic process.

3.2.3. Comparison with Prior Reports

The principal contribution of this work is therefore methodological rather than algorithmic. Previous studies frequently report higher predictive performance using internal random data splits, pooled cohorts, or preprocessing performed before data partitioning. By adopting a leakage-safe preprocessing pipeline together with bidirectional external validation, the present study provides performance estimates that are more likely to reflect real-world clinical deployment. The resulting AUC is lower than some previously reported values, including our own earlier analysis, but we regard this as evidence of improved methodological rigor rather than diminished model capability.

3.2.4. Responsible Deployment and Algorithmovigilance

Explainability tells a clinician what a model weighted for a given case; it does not establish that the model remains accurate after deployment, performs equitably across subgroups, or that anyone is accountable when it errs. Explainability is a transparency property and safety is a systems property; the two are not substitutes. Adopting the minimum-viable governance framing of Sirago and colleagues [21], deployment of a screening model of this kind would require, at minimum: prospective and geographically external validation in the target population with locally normed instruments; continuous post-deployment monitoring of discrimination and calibration; dataset-shift and drift detection with pre-specified review triggers; bias and fairness surveillance by age, sex, education, and language, which is essential for a model dominated by an education-sensitive instrument; defined human oversight and clinical responsibility, with the model advisory rather than determinative; incident reporting and a mechanism to withdraw or roll back a degraded model; and maintenance of calibration, since a model that drifts toward over-prediction will

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2026.0412
e ISSN: 2584-2854
Volume: 04 Issue: 10
October 2026
Page No: 2923 - 2939

generate referral volumes a constrained system cannot absorb. The absence of this infrastructure, rather than any shortfall in accuracy, is the binding constraint on deploying models of this kind. Accordingly, external validation should be viewed as a prerequisite for deployment rather than the final stage of model development, with continuous monitoring required throughout clinical use.

3.2.5. Cross-Sectional Classification Does Not Imply Prognostic Ability

Although the proposed framework achieved excellent discrimination for concurrent cognitive impairment, it did not identify individuals who subsequently progressed from normal cognition to impairment. Baseline MMSE scores and MRI-derived brain-volume measures showed no significant differences between future converters and cognitively stable participants, indicating that the predictors used in this study primarily capture contemporaneous disease status rather than preclinical neurodegenerative change. This distinction is clinically important because cross-sectional case finding and longitudinal risk prediction represent fundamentally different prediction tasks. Consequently, the proposed framework should be interpreted as a tool for screening existing cognitive impairment rather than for forecasting future decline. Development of reliable prognostic models will require larger longitudinal cohorts with sufficient numbers of incident cases and additional biomarkers sensitive to early neurodegenerative processes.

3.2.6. Limitations

First, incorporation bias: the MMSE, the dominant predictor, is entangled with the CDR-based staging that defines the outcome. The demographics-only analysis bounds this but cannot remove it. Second, demographic narrowness: both cohorts are North American, predominantly white, and relatively highly educated. No inference to South Asian, African, or low-literacy populations is warranted, and an MMSE-dominated model is specifically expected to transport poorly across educational and linguistic distributions. Third, education harmonisation: education was recorded on different scales in the two cohorts and mapped between them; it is the least reliable feature and contributes little, but the mapping remains a source of potential bias. Fourth, the outcome is cross-sectional: the study classifies concurrent status, not future decline, and the converter analysis shows baseline measures did not anticipate conversion. This is not an early-detection tool. Fifth, the external cohort is modest in size: external validation rested on 150 participants, confidence intervals are correspondingly wide, and further validation in larger, more diverse cohorts is needed.

Sixth, a single MMSE administration and clinical labels: cross-sectional scores cannot capture rate of change, and labels are clinical rather than biomarker- or autopsy-confirmed.

Conclusion

Under leakage-safe development and genuine independent-cohort external validation across OASIS-1 and OASIS-2, explainable machine learning applied to five routinely collected clinical variables classified concurrent cognitive impairment with an external AUC of 0.92 and materially outperformed the conventional MMSE cut-off, chiefly by recovering sensitivity that the fixed threshold discards. Most of the model's discrimination derived from the MMSE, structural imaging added a small but real increment, and baseline measures did not distinguish the participants who later converted. The appropriate positioning is therefore calibrated decision support for case-finding, not early detection, and not a deployable screening tool: external validation in demographically diverse and non-Western cohorts, calibration maintenance, and the governance infrastructure described above are prerequisites for any clinical use. We present the study as a rigorous, reproducible proof of concept and, equally, as a demonstration that credible performance in this area depends on removing leakage and validating across genuinely independent cohorts.

Supplementary Materials

The following supporting information is submitted with this manuscript. Table S1: complete external-validation metrics including confusion matrices, PPV, NPV, balanced accuracy, and precision–recall AUC with 95% confidence intervals; Table

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2026.0412
e ISSN: 2584-2854
Volume: 04 Issue: 10
October 2026
Page No: 2923 - 2939

S2: hyperparameter grids and selected values for all five classifiers; Table S3: TRIPOD checklist; Table S4: performance at the default and sensitivity-prioritising operating points; Table S5: reverse-direction (OASIS-2 to OASIS-1) results for all five classifiers.

Author Contributions

Conceptualisation, D.P.P.; methodology, D.P.P. and C.S.; software, D.P.P.; formal analysis, D.P.P.; data curation, D.P.P.; writing—original draft preparation, D.P.P.; writing—review and editing, C.S.; supervision, C.S. All authors have read and agreed to the published version of the manuscript.

Funding

This research received no external funding.

Institutional Review Board Statement

This study is a secondary analysis of de-identified, openly licensed data obtained from the Open Access Series of Imaging Studies. No participants were recruited, contacted, or identified by the authors, and no new data were collected. The original OASIS data collection was approved by the Institutional Review Board of Washington University School of Medicine, and the data are distributed in de-identified form under the OASIS Data Use Agreement. The present secondary analysis was reviewed by the Institutional Review Board of Sinhgad Institute of Management.

Informed Consent Statement

Patient consent was waived because the analysis used previously collected, de-identified, publicly available data in which no identifiable individual information is presented.

Data Availability Statement

OASIS-1 and OASIS-2 are openly available from the OASIS project [8,9] to researchers who agree to the OASIS Data Use Agreement; the data are not redistributed by the authors. The complete analysis code, including the harmonisation logic, the leakage-safe pipeline, hyperparameter grids, random seeds, and the scripts that regenerate every table and figure reported here, is openly available at https://github.com/dp2006/oasis-cognitive. The repository contains a documented loader implementing the eligibility and harmonisation rules of Sections 2.4 and 2.5, the leakage-safe pipeline of Section 2.6, the hyperparameter grids of Section 2.7, and a single entry point (scripts/run_all.sh), together with a requirements file pinning all package versions.

Conflicts Of Interest

The authors declare no conflicts of interest.

Abbreviations

AUC, area under the receiver operating characteristic curve; CDR, Clinical Dementia Rating; CI, confidence interval; DCA, decision-curve analysis; eTIV, estimated total intracranial volume; IQR, interquartile range; MCI, mild cognitive impairment; MMSE, Mini-Mental State Examination; NPV, negative predictive value; nWBV, normalised whole-brain volume; OASIS, Open Access Series of Imaging Studies; PPV, positive predictive value; RBF, radial basis function; SES, socioeconomic status; SHAP, SHapley Additive exPlanations; SVM, support vector machine; TRIPOD, Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis; TRIPOD+AI, the 2024 artificial-intelligence extension of TRIPOD.

Acknowledgements

Data were provided by OASIS. OASIS-1: Cross-Sectional: Principal Investigators: D. Marcus, R. Buckner, J. Csernansky, J. Morris; P50 AG05681, P01 AG03991, P01 AG026276, R01 AG021910, P20 MH071616, U24 RR021382. OASIS-2: Longitudinal: Principal Investigators: D. Marcus, R. Buckner, J. Csernansky, J. Morris; P50 AG05681, P01 AG03991, P01 AG026276, R01 AG021910, P20 MH071616, U24 RR021382.

References

[1]. World Health Organization. (2023). Dementia: Fact sheet. World Health Organization, Geneva, Switzerland.
[2]. Patterson, C. (2018). World Alzheimer report 2018: The state of the art of dementia research. Alzheimer's Disease International, London, UK.
[3]. Folstein, M. F., Folstein, S. E., & McHugh, P. R. (1975). "Mini-mental state": A practical method for grading the cognitive state of patients for the clinician. Journal of Psychiatric Research, 12, 189–198.

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2026.0412
e ISSN: 2584-2854
Volume: 04 Issue: 10
October 2026
Page No: 2923 - 2939

[4]. Nasreddine, Z. S., Phillips, N. A., Bédirian, V., Charbonneau, S., Whitehead, V., Collin, I., Cummings, J. L., & Chertkow, H. (2005). The Montreal Cognitive Assessment, MoCA: A brief screening tool for mild cognitive impairment. Journal of the American Geriatrics Society, 53, 695–699.

[5]. Petersen, R. C. (2016). Mild cognitive impairment. Continuum (Minneapolis, Minn.), 22, 404–418.

[6]. Cuingnet, R., Gerardin, E., Tessieras, J., Auzias, G., Lehéricy, S., Habert, M.-O., Chupin, M., Benali, H., & Colliot, O. (2011). Automatic classification of patients with Alzheimer's disease from structural MRI: A comparison of ten methods using the ADNI database. NeuroImage, 56, 766–781.

[7]. Kaufman, S., Rosset, S., Perlich, C., & Stitelman, O. (2012). Leakage in data mining: Formulation, detection, and avoidance. ACM Transactions on Knowledge Discovery from Data, 6, 1–21.

[8]. Marcus, D. S., Wang, T. H., Parker, J., Csernansky, J. G., Morris, J. C., & Buckner, R. L. (2007). Open Access Series of Imaging Studies (OASIS): Cross-sectional MRI data in young, middle aged, nondemented, and demented older adults. Journal of Cognitive Neuroscience, 19, 1498–1507.

[9]. Marcus, D. S., Fotenos, A. F., Csernansky, J. G., Morris, J. C., & Buckner, R. L. (2010). Open Access Series of Imaging Studies: Longitudinal MRI data in nondemented and demented older adults. Journal of Cognitive Neuroscience, 22, 2677–2684.

[10]. Breiman, L. (2001). Random forests. Machine Learning, 45, 5–32.

[11]. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). San Francisco, CA, USA.

[12]. Lee, G., Nho, K., Kang, B., Sohn, K.-A., & Kim, D. (2019). Predicting Alzheimer's disease progression using multi-modal deep learning approach. Scientific Reports, 9, 1952.

[13]. Jiang, S., Li, Z., & Qian, X. (2017). Three-dimensional convolutional neural networks for Alzheimer's disease diagnosis. IEEE Transactions on Medical Imaging, 36, 2388–2400.

[14]. Zhang, F., Li, Z., Zhang, B., Du, H., Wang, B., & Zhang, X. (2019). Multi-modal deep learning model for auxiliary diagnosis of Alzheimer's disease. Neurocomputing, 361, 185–195.

[15]. Collins, G. S., Moons, K. G. M., Dhiman, P., Riley, R. D., Beam, A. L., Van Calster, B., Ghassemi, M., Liu, X., Reitsma, J. B., van Smeden, M., et al. (2024). TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ, 385, e078378.

[16]. Collins, G. S., Reitsma, J. B., Altman, D. G., & Moons, K. G. M. (2015). Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD). Annals of Internal Medicine, 162, 55–63.

[17]. Van Calster, B., McLernon, D. J., van Smeden, M., Wynants, L., & Steyerberg, E. W. (2019). Calibration: The Achilles heel of predictive analytics. BMC Medicine, 17, 230.

[18]. Moons, K. G. M., Altman, D. G., Reitsma, J. B., Ioannidis, J. P. A., Macaskill, P., Steyerberg, E. W., Vickers, A. J., Ransohoff, D. F., & Collins, G. S. (2015). Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): Explanation and elaboration. Annals of Internal Medicine, 162, W1–W73.

[19]. Collins, G. S., Dhiman, P., Ma, J., Schlussel, M. M., Archer, L., Van Calster,

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2026.0412
e ISSN: 2584-2854
Volume: 04 Issue: 10
October 2026
Page No: 2923 - 2939

B., Harrell, F. E., Martin, G. P., Moons, K. G. M., van Smeden, M., et al. (2024). Evaluation of clinical prediction models: How to undertake and interpret external validation. BMJ, 384, e074819.

[20]. Whiting, P., Rutjes, A. W. S., Reitsma, J. B., Glas, A. S., Bossuyt, P. M. M., & Kleijnen, J. (2004). Sources of variation and bias in studies of diagnostic accuracy: A systematic review. Annals of Internal Medicine, 140, 189–202.

[21]. Sirago, G., Vinci, A., Visci, P., Solarino, B., Dell'Erba, A., & Ferorelli, D. (2026). From black-box clinical AI to algorithmovigilance: A minimum-viable governance protocol for safety, accountability, and prevention of algorithmic harm. Journal of Medical Systems, 50, 52.

[22]. Morris, J. C. (1993). The Clinical Dementia Rating (CDR): Current version and scoring rules. Neurology, 43, 2412–2414.

[23]. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.

[24]. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems (Vol. 30, pp. 4765–4774). Curran Associates, Red Hook, NY, USA.

[25]. Vickers, A. J., & Elkin, E. B. (2006). Decision curve analysis: A novel method for evaluating prediction models. Medical Decision Making, 26, 565–574.

[26]. Steyerberg, E. W. (2019). Clinical prediction models: A practical approach to development, validation, and updating (2nd ed.). Springer, Cham, Switzerland.