

Assessment-Driven Machine Learning For Dementia Prediction: A Study On ADNI Data

Chanidapa Winalai^{1*} and Dominique J. Bicut^{1,2}

¹ EPSP, TIMC Laboratory, UMR CNRS5525, Grenoble Alpes University, VetAgro Sup, Grenoble, France

² Institute Laue-Langevin, 71 Avenue des Martyrs, 38042 Grenoble, France

1 <https://orcid.org/0009-0008-1435-7445>; 2 <https://orcid.org/0000-0003-0750-997X>

*Corresponding Author: Chanidapa Winalai

Abstract

Early dementia screening remains limited by the cost, invasiveness, and accessibility of neuroimaging and molecular biomarkers. Cognitive assessments offer a scalable alternative, but their interpretation is challenged by diagnostic uncertainty and variability across clinical visits. In this study, we present an assessment-driven machine learning framework for dementia screening based exclusively on cognitive assessments.

Using longitudinal cognitive and demographic data from the ADNI cohort, we train lightweight classification models to estimate probabilistic diagnostic states across the cognitive spectrum (cognitively normal, mild cognitive impairment, Alzheimer's disease). Rather than enforcing rigid class boundaries, the proposed approach outputs diagnosis probabilities that evolve across visits, supporting dynamic clinical interpretation.

To ensure transparency and usability, we apply SHAP (SHapley Additive exPlanations) to quantify feature contributions and establish a ranking of cognitive assessments driving model predictions. Our results show that a small subset of cognitive scores consistently accounts for most of the predictive power, achieving competitive performance while reducing assessment burden.

Overall, this work demonstrates how explainable, assessment-based machine learning can be deployed as a practical decision-support tool for dementia screening and follow-up monitoring, particularly in pre-clinical settings and resource-limited environments.

Keywords: Alzheimer's disease screening; ADNI; cognitive assessment; machine learning; clinical decision support.

1. INTRODUCTION

Dementia affects approximately 55 million people worldwide, with nearly 10 million new cases annually [1]. Early detection of cognitive decline, particularly at the mild cognitive impairment (MCI) stage, remains crucial for timely intervention and patient management [2]. However, differentiating normal aging from pathological cognitive decline presents significant clinical challenges, with substantial inter-rater variability even among experienced clinicians [3].

Clinical practice relies on standardized cognitive assessments, each with established diagnostic thresholds. The Mini-Mental State Examination (MMSE) screens for global cognitive impairment, with scores below 24 indicating possible dysfunction [4]. The Functional Activities Questionnaire (FAQ) evaluates independence in daily activities, where scores above 5 suggest functional decline [5]. Neuropsychological tests such as Logical Memory from the Wechsler Memory Scale assess episodic memory through delayed recall [6], while ADAS-Cog evaluates word recall and other cognitive domains [7]. While these

instruments have well-defined cutoffs, integrating multiple assessments while considering patient history represents a complex task often subject to clinical variability.

Machine learning approaches offer potential to support clinical decision-making by identifying patterns across assessments and patient histories [8, 9]. Previous studies applying various algorithms to dementia prediction have reported accuracies between 70-85% [10-12]. However, most approaches either disregard established clinical knowledge or treat patients as independent observations, neglecting the longitudinal nature of cognitive decline.

This study presents a systematic approach to dementia prediction that explicitly incorporates clinical diagnostic criteria while maintaining model interpretability. We first conduct feature selection to identify the most informative cognitive assessments from a comprehensive battery. Using these selected assessments, we then develop an enhanced model that integrates clinical thresholds with longitudinal patient trajectories to predict cognitive status at each clinical visit. This two-stage methodology aims to balance predictive accuracy with clinical usability, providing a practical decision-support tool that respects established diagnostic frameworks while learning from longitudinal data patterns.

2. MATERIALS AND METHODS

2.1 Data Source and Study Population

Data were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu), a multi-center longitudinal study launched in 2004 to investigate biomarkers for Alzheimer's disease diagnosis and progression [13]. ADNI recruited participants aged 55-90 years across three diagnostic categories: cognitively normal (CN), mild cognitive impairment (MCI), and Alzheimer's disease (AD). All participants provided written informed consent, with institutional review board approval at each participating site. We extracted data from the ADNIMERGE dataset, which consolidates cognitive assessments, demographics, and clinical diagnoses across all ADNI phases. The dataset included 11,135 visits from 2,086 unique participants who completed multiple assessments over time. Diagnoses were harmonized into three categories: Normal (including CN and subjective memory complaints), MCI (including early and late MCI), and AD (including dementia). Only participants with at least two visits were retained to enable longitudinal analysis. The final dataset comprised visits with diagnosis distributions of: MCI (4,875 visits, 43.8%), Normal (3,861 visits, 34.7%), and AD (2,399 visits, 21.5%).

2.2 Cognitive Assessments

We initially considered multiple validated cognitive instruments available in ADNI covering different cognitive domains. Global cognition was assessed using the Mini-Mental State Examination (MMSE), a 30-point scale evaluating orientation, memory, attention, language, and visuospatial function [4], and the Montreal Cognitive Assessment (MoCA) as an alternative screening tool [14]. Memory assessments included: Logical Memory from the Wechsler Memory Scale-Revised with LDELTOTAL representing delayed recall score (0-25) [6]; Rey Auditory Verbal Learning Test (RAVLT) for immediate and delayed recall; and ADAS-Cog Word Recall (ADAS-Q4) scored as number of words not recalled (0-10, lower is better) [7]. Executive function was evaluated using Trail Making Tests A and B [15] and Category Fluency tests [16]. Functional status was assessed via the Functional Activities Questionnaire (FAQ), a 10-item informant-based evaluation of instrumental activities of daily living (0-30,

higher indicates greater impairment) [5]. All assessments followed standardized administration protocols established by ADNI.

2.3 Feature Selection

To identify the most informative cognitive assessments while minimizing redundancy, we implemented a systematic three-step feature selection process as shown in Figure 1. First, Pearson correlation coefficients were computed between each cognitive assessment and diagnosis labels to identify features with strong diagnostic relationships. Features showing weak correlations ($|r| < 0.30$) were excluded from further consideration. Second, we evaluated each remaining feature's individual predictive strength by training simple Random Forest classifiers using single features, retaining those achieving over 60% classification accuracy when used alone. This step ensured that retained features carried sufficient diagnostic information independently. Third, to address multicollinearity among the selected features, we calculated Variance Inflation Factors (VIF). Features with VIF scores above 5 were systematically removed to ensure statistical independence among the final feature set, as multicollinearity can distort model interpretations and reduce generalization performance.

Using the features identified through this selection process, we evaluated four machine learning algorithms: Random Forest (RF), Extreme Gradient Boosting (XGBoost), Support Vector Machine (SVM), and Logistic Regression (LR) for individual predictive strength. To assess the impact of feature set size on performance, we systematically tested all possible combinations ranging from two to five features. For each combination and algorithm, hyperparameters were optimized using grid search combined with 10-fold stratified cross-validation. The dataset was divided into 80% training and 20% testing sets, with stratification ensuring balanced diagnostic category representation in both sets. This evaluation framework enabled identification of the optimal balance between predictive performance and feature parsimony, which is critical for clinical deployment where assessment time and patient burden must be minimized.

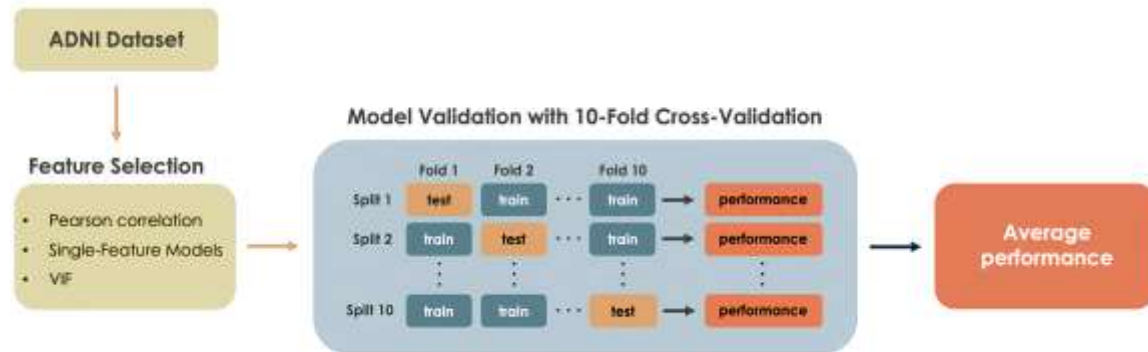


Figure 1: Workflow of feature selection and model evaluation using stratified cross-validation.

2.4 SHAP Analysis

To quantify each feature's contribution to predictions and ensure model interpretability, we applied Shapley Additive Explanations (SHAP) to the best-performing model configuration. SHAP values provide a unified measure of feature importance derived from game theory, offering both global feature importance rankings (average absolute impact across all predictions) and local explanations (contribution to individual predictions). This analysis

allowed us to identify which cognitive assessments most strongly influenced diagnostic predictions and to reveal threshold effects corresponding to established clinical diagnostic criteria, thereby validating that the model learned clinically meaningful decision boundaries.

2.5 Clinical Threshold Integration and Enhanced Feature Engineering

Following feature selection, we developed an enhanced model that explicitly incorporated established clinical diagnostic thresholds. For each selected assessment, we created multiple derived features based on published diagnostic criteria and clinical practice guidelines. Severity categories were defined as follows:

- MMSE: Severe (<10), Moderate (10-19), Mild (20-23), Normal (≥ 24) [4, 17]
- FAQ: Normal (≤ 5), Mild (6-9), Moderate (10-15), Severe (>15) [5]
- LDELTOTAL: Severe (<2), Moderate (2-3), Mild (4-7), Normal (≥ 8) [6]
- ADAS-Q4: Normal (≤ 3), Mild (4-6), Moderate (7-8), Severe (>8) [7].

These categories reflect clinically recognized levels of cognitive impairment documented in the diagnostic literature.

In addition to severity categories, we created binary threshold indicators for key clinical cutoffs that trigger diagnostic concern in practice (e.g., MMSE < 24 indicating cognitive impairment, FAQ > 5 suggesting functional decline, LDELTOTAL < 8 indicating memory impairment, ADAS-Q4 > 3 suggesting word recall deficit). We also computed a cognitive composite score as a weighted combination of normalized assessment scores, with weights determined by SHAP importance rankings from the initial feature selection model. Finally, we created an aggregate clinical risk category (low risk, moderate risk, high risk) based on the number and severity of impaired domains, providing a holistic measure of cognitive status. This multi-level feature representation allowed the model to learn from both continuous scores and categorical clinical interpretations, mirroring how expert clinicians integrate quantitative test results with qualitative severity judgments during diagnosis.

2.6 Longitudinal Sliding Window Framework

To capture cognitive trajectories over time, we developed a sliding window approach that incorporates each patient's visit history as features for prediction at the current visit. For each visit (excluding the baseline visit where no history exists), we extracted several categories of historical features. Visit context features included: visit number (position in patient timeline), total number of previous visits, and last diagnosis prior to the current visit. Historical diagnosis distribution was captured through counts of previous visits diagnosed as Normal, MCI, or AD, along with proportions of previous visits showing impaired scores on each assessment (e.g., proportion with MMSE < 24 , FAQ > 5 , LDELTOTAL < 8). We also included raw scores from up to five previous visits for each assessment to allow the model to detect non-linear patterns in recent progression.

This sliding window framework generated a comprehensive feature vector for each visit that combined: (1) current visit assessment scores, (2) clinical threshold-based derived features, and (3) longitudinal historical features capturing trajectory and context. The final model used Random Forest with 600 trees, maximum depth of 25, minimum samples per split of 10, minimum samples per leaf of 4, and balanced class weights to account for diagnostic category imbalances in the dataset. The model was trained on 80% of visits (8,908 visits) and evaluated on a held-out test set (20%, 2,227 visits) using stratified sampling to maintain diagnostic proportions and ensure fair evaluation across all categories.

2.7 Model Evaluation and Statistical Analysis

Model performance was evaluated using accuracy, precision, recall, and F1-score for each diagnostic category. Confusion matrices quantified misclassification patterns. Wilson score

intervals were computed for test set accuracy to provide conservative confidence bounds accounting for finite sample size [18]. Statistical significance of performance differences between models was assessed using McNemar's test for paired predictions [19]. All analyses were conducted using Python 3.8 with scikit-learn, pandas, and numpy libraries.

3. RESULTS AND DISCUSSION

3.1 Feature Selection Identifies Four Key Cognitive Assessments

Our systematic feature selection process successfully identified a parsimonious set of cognitive assessments with strong diagnostic value while minimizing redundancy. Initial correlation analysis revealed several features demonstrating strong relationships with diagnosis ($|r| > 0.65$), including measures from ADAS, FAQ, MMSE, RAVLT domains, and composite cognitive scores (Figure 2). Single-feature Random Forest models confirmed that these assessments individually achieved classification accuracies exceeding 60% when used alone, validating their diagnostic utility. However, substantial intercorrelations among certain features suggested potential multicollinearity, particularly among subdomains of informant-based measures and composite scores that aggregate multiple assessments.

Application of Variance Inflation Factor analysis to address multicollinearity yielded eight candidate features: ADAS-Q4, MMSE, RAVLT_immediate, LDELTOTAL, FAQ, and three ECog subscales (language, visuospatial, divided attention). We then systematically evaluated all possible feature combinations ranging from two to five variables across four algorithms (Random Forest, XGBoost, SVM, Logistic Regression) using 10-fold stratified cross-validation. Table 1 presents the top-performing configurations for each algorithm across different feature set sizes.

Random Forest consistently achieved the highest classification accuracy across feature set sizes. The optimal configuration used four features: LDELTOTAL, FAQ, MMSE, and ADAS-Q4, yielding 90.10% training accuracy and 85.64% testing accuracy (Table 1). This four-feature model outperformed all other combinations tested. Notably, adding a fifth feature did not improve and sometimes reduced testing accuracy, likely due to introducing redundancy and overfitting. This diminishing return with additional features has been documented in previous cognitive modeling studies [10] and emphasizes the importance of feature parsimony for maintaining model generalization to new patients.

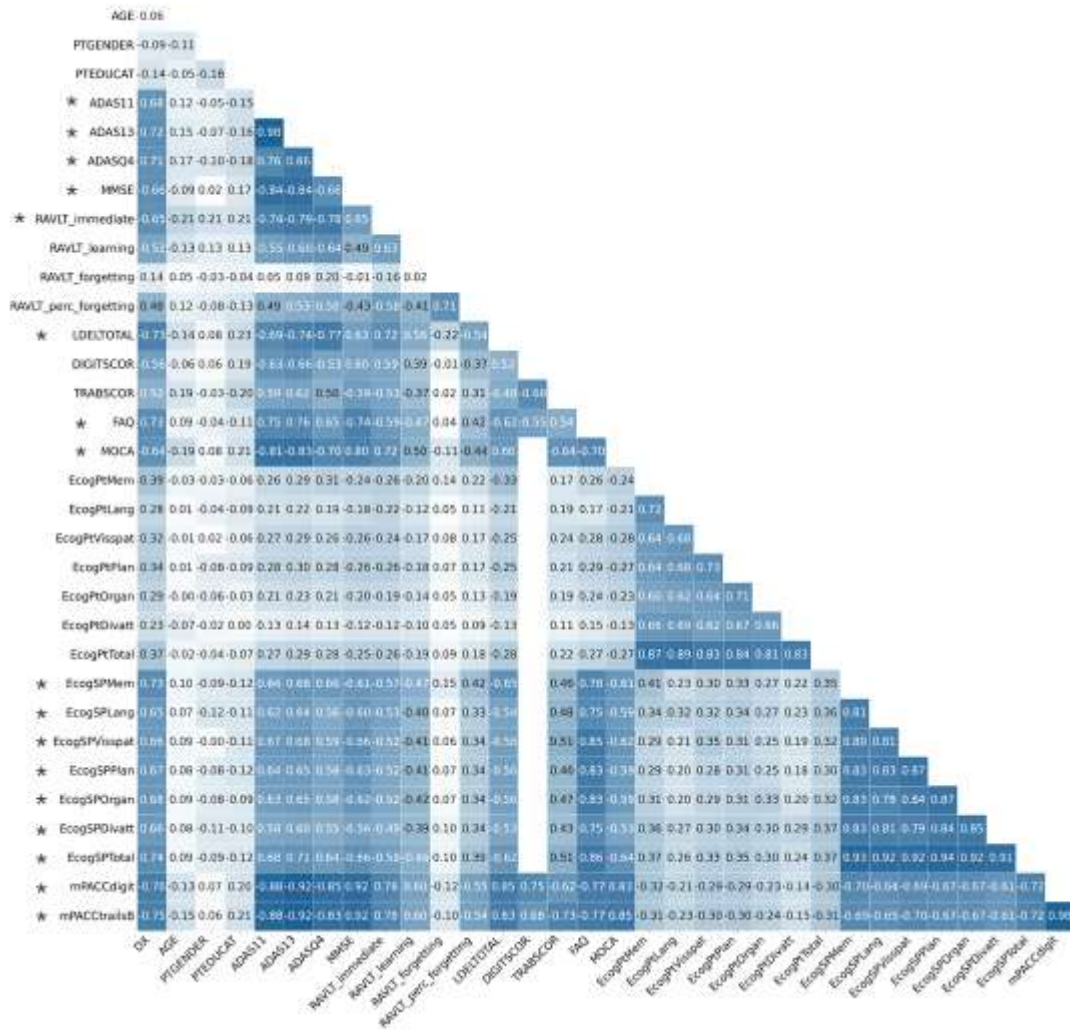


Figure 2: Pearson correlation matrix between cognitive assessments and diagnostic labels

Table 1: Performance Comparison of Models with Varying Feature Combinations

Model	Number of feature combination	Best Feature Set	Training Accuracy	Testing Accuracy
RF	2	LDELTOTAL, FAQ	79.06%	76.38%
	3	LDELTOTAL, FAQ, MMSE	84.36%	81.52%
	4	LDELTOTAL, FAQ, MMSE, ADASQ4	90.10%	85.64%
	5	LDELTOTAL, FAQ, MMSE, ADASQ4, RAVLT_immediate	88.18%	82.11%
XGBoost	2	LDELTOTAL, FAQ	77.83%	75.40%
	3	LDELTOTAL, FAQ, MMSE	80.51%	76.75%

SVM	4	LDELTOTAL, FAQ, MMSE, ADASQ4	83.16%	77.88%
	5	LDELTOTAL, FAQ, MMSE, ADASQ4, RAVLT_immediate	86.57%	76.86%
	2	LDELTOTAL, FAQ	76.59%	74.88%
	3	LDELTOTAL, FAQ, MMSE	77.74%	75.02%
	4	LDELTOTAL, FAQ, MMSE, ADASQ4	79.60%	78.31%
	5	LDELTOTAL, FAQ, MMSE, ADASQ4, RAVLT_immediate	78.10%	77.77%
LR	2	LDELTOTAL, FAQ	76.34%	76.30%
	3	LDELTOTAL, FAQ, MMSE	77.85%	77.76%
	4	LDELTOTAL, FAQ, MMSE, ADASQ4	78.40%	78.31%
	5	LDELTOTAL, FAQ, MMSE, ADASQ4, RAVLT_immediate	78.15%	78.03%

The four selected assessments capture distinct and complementary cognitive domains that align with established diagnostic frameworks [20, 21]. LDELTOTAL assesses delayed verbal memory, a sensitive marker of hippocampal dysfunction and early Alzheimer's pathology. The FAQ evaluates functional capacity in daily activities, capturing real-world cognitive impact that distinguishes MCI from dementia. MMSE provides a global cognitive screen assessing orientation, attention, language, and visuospatial function. ADAS-Q4 measures immediate word recall, complementing the story-based delayed recall of LDELTOTAL. Together, these assessments provide comprehensive coverage of memory, function, and global cognition while avoiding redundancy. The superiority of Random Forest over other algorithms likely stems from its ability to capture non-linear relationships and interactions between assessments without requiring explicit specification of these relationships, while maintaining interpretability through feature importance rankings showed in Figure 3.

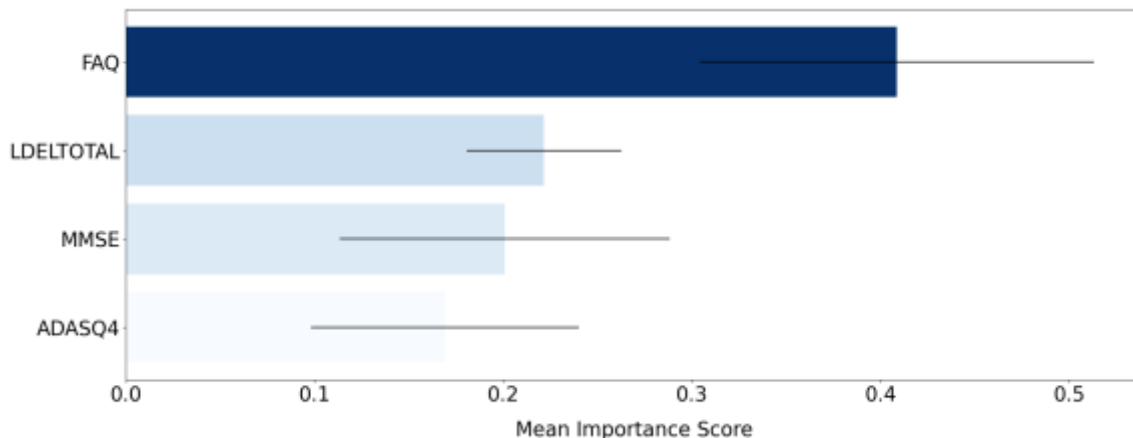


Figure 3: Feature importance scores from the Random Forest model with error bars indicating uncertainty (standard deviation) across trees.

3.2 SHAP Analysis and Clinical Threshold Effects

SHAP analysis quantified each assessment's contribution to diagnostic predictions (see Figure 4). LDELTOTAL demonstrated the highest importance (SHAP value: 0.42), aligning with literature identifying delayed memory recall as a sensitive marker of early cognitive decline and transition from normal cognition to MCI [22]. FAQ ranked second (0.38), capturing real-world functional impact that becomes pronounced as individuals progress from MCI to dementia [23]. MMSE showed moderate importance (0.31) despite widespread clinical use, likely reflecting known limitations in detecting early changes, particularly in educated individuals [24]. ADAS-Q4 contributed least (0.24) but provided complementary information about immediate word recall.

SHAP dependence plots revealed threshold effects corresponding closely to established clinical cutoffs. For LDELTOTAL, a sharp transition occurred around score 8, matching the clinical threshold for impaired delayed memory. FAQ showed marked changes at score 5, the cutoff for functional independence. MMSE showed transitions at scores 24 and 20, corresponding to mild and moderate impairment thresholds. These patterns indicate the model learned clinically meaningful decision boundaries rather than arbitrary splits, validating alignment with established diagnostic knowledge and motivating our subsequent integration of explicit threshold-based features.

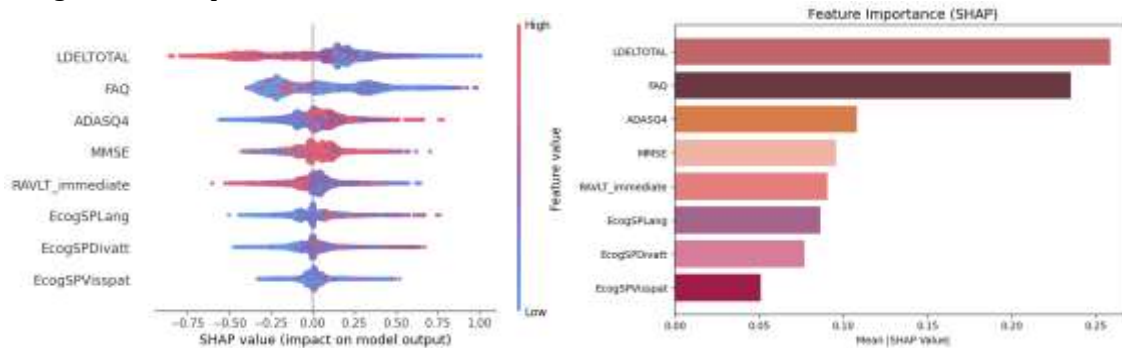


Figure 4: SHAP feature importance and dependence plots showing threshold effects

3.3 Clinical Threshold Integration and Longitudinal Data Improve Accuracy

Building upon the four selected assessments, we developed an enhanced model that explicitly incorporated clinical diagnostic thresholds and longitudinal patient history through our sliding window framework. Ten-fold cross-validation yielded a mean accuracy of 90.48% (SD: 1.04%, range: 88.33-91.69%). The held-out test set achieved 90.03% accuracy (95% CI: 88.72-91.21%), representing a statistically significant 4.4 percentage point improvement over the baseline model, which used only raw scores (85.64%, $p < 0.001$).

Performance was well-balanced across diagnostic categories. AD classification achieved a precision of 88.5% and a recall of 87.9% (F1-score: 0.882). The MCI classification achieved a precision of 89.8% and a recall of 87.3% (F1-score: 0.885), notably exceeding the typical 0.70-0.80 range reported in previous studies [25, 26]. Normal classification showed the strongest performance with a precision of 91.3% and a recall of 94.8% (F1-score: 0.930). Critically, the model rarely made gross errors: only 1 of 772 Normal visits was misclassified as AD, demonstrating clinically appropriate decision boundaries.

The confusion matrix revealed that misclassifications occurred predominantly between adjacent diagnostic categories (Normal-MCI or MCI-AD) rather than across the full spectrum (Table 2). Of 480 AD visits, 422 (87.9%) were correctly classified, with 58 (12.1%)

misclassified as MCI and none as Normal, a clinically reasonable error pattern suggesting the model appropriately recognizes borderline dementia cases. For 975 MCI visits, 851 (87.3%) were correctly classified, with 54 (5.5%) classified as AD and 70 (7.2%) as Normal. This pattern reflects the genuine heterogeneity of MCI, where some individuals show progressive decline toward dementia while others remain stable or even revert to normal cognition. The model's ability to capture this heterogeneity, rather than forcing rigid category boundaries, represents an important strength for clinical application, where recognizing diagnostic uncertainty is valuable.

3.4 Comparison with Rule-Based Classification

To quantify the advantage of machine learning over conventional diagnostic approaches, we compared our enhanced model with a simple rule-based classifier that applied standard threshold cutoffs to current visit assessments only. The rule-based approach achieved 58.46% accuracy (95% CI: 56.40-60.49%), substantially lower than our model's 90.03% ($p < 0.001$, McNemar's test). This represents an absolute improvement of 31.57 percentage points, or 54.0% relative improvement (Table 2).

Analysis of error patterns revealed systematic biases in rule-based classification. It showed high AD recall (91.7%) but poor precision (46.2%), over-diagnosing 411 of 975 MCI cases (42.2%) as AD. MCI detection was particularly weak (F1-score: 0.374), as borderline cases were misclassified as either AD or Normal. This poor MCI performance is concerning, given that MCI represents the primary target for early intervention.

These findings confirm that while clinical criteria provide valuable frameworks, their mechanical application yields poor performance. The 54% improvement validates the clinical value of integrating multiple assessments, threshold interactions, and longitudinal context, processes that experienced clinicians naturally employ but are difficult to capture in simple rule-based systems.

Table 2: ML Model vs. Rule-Based Classification

Method	Accuracy (%)	AD F1	MCI F1	Normal F1
Rule-Based	58.46	0.614	0.374	0.722
Stage 1 (4 features)	85.64	0.880	0.820	0.850
Stage 2 (4 features + History)	90.03	0.882	0.885	0.930
Improvement vs. Rules	+31.57	+0.268	+0.511	+0.208

3.5 Clinical Decision Support Application

To demonstrate practical clinical applicability, we developed an interactive web-based interface implementing the validated model. The interface accepts the four required cognitive assessments (LDELTOTAL, FAQ, MMSE, ADAS-Q4) along with an optional visit history. For first-time evaluations, the tool achieves approximately 82% accuracy using current scores only, improving to 90% when longitudinal data are available for returning patients.

The tool outputs probability distributions across diagnostic categories (Normal, MCI, AD) rather than hard classifications, supporting nuanced clinical interpretation. It displays current assessment scores with severity categories, cognitive composite score, clinical risk level, and model confidence. When longitudinal data are available, the interface visualizes assessment trends and flags significant changes, helping clinicians quickly grasp trajectory patterns.

A guided assessment mode provides standardized protocols with scoring instructions for each test [4-7], enabling consistent administration even by users unfamiliar with these instruments. This promotes reliability across different assessors and settings.

This decision-support tool addresses practical clinical needs for efficient assessment. Traditional neuropsychological batteries require 2-4 hours, creating barriers to routine screening. Our four-assessment battery requires approximately 30 minutes total, making it feasible for primary care settings. The tool's strong first-visit performance enables initial screening, with accuracy progressively improving as history accumulates, creating incentive for systematic longitudinal tracking to monitor cognitive trajectories.

ADNI Cognitive Assessment Predictor

Validated clinical assessments for cognitive diagnosis prediction

Assessment References

- MMSE: Folstein et al., 1975 - Mini-Mental State Examination
- FAQ: Pfeffer et al., 1982 - Functional Activities Questionnaire
- LDELTOTAL: Wechsler, 1987 - Logical Memory Delayed Recall (WMS-R)
- ADAS-Q4: Rosen et al., 1984 - ADAS-Cog Word Recall

Configuration

Is this your first visit?

☒ Yes ☐ No

Input method

☒ I have the scores ☐ Help me measure them

[Reset All](#)

Figure 5: Screenshot of ADNI Cognitive Assessment Predictor web interface

4. CONCLUSION

This study demonstrates that systematic integration of clinical diagnostic criteria with machine learning and longitudinal patient data substantially improves dementia prediction. Through rigorous feature selection, we identified four cognitive assessments: LDELTOTAL, FAQ, MMSE, and ADAS-Q4, that provide a comprehensive evaluation.

Our enhanced model achieved 90% accuracy, representing a 4.4 percentage point improvement over models using only raw scores and a 54% relative improvement over rule-based threshold application. This performance compares favorably with previous machine learning studies (typically 70-85% accuracy) [10-12, 27] and approaches the inter-rater reliability of expert clinicians (kappa: 0.78-0.88) [3]. Three methodological factors contribute to this advantage: explicit incorporation of clinical thresholds as engineered features, longitudinal trajectory capture through sliding windows, and systematic feature selection to minimize redundancy.

The model's strength lies in formalizing the integration of multiple assessments, diagnostic criteria, and patient history that characterizes expert clinical reasoning. By incorporating established clinical thresholds as features rather than external validation criteria, the model learns how clinicians weight different diagnostic information sources. The sliding window framework provides systematic quantification of cognitive trajectories, enabling evidence-based assessment of stability versus progression.

Our comparison with rule-based classification provides important context often missing from healthcare machine learning studies. That simple threshold-based rules achieve only 58% accuracy confirms that while diagnostic criteria define useful frameworks, their mechanical application yields poor performance, validating the need for decision-support tools that formalize clinicians' nuanced integration processes.

While external validation in diverse clinical populations remains necessary, this work establishes proof-of-concept for threshold-enhanced longitudinal machine learning in dementia diagnosis. The methodology is generalizable to other cognitive assessment batteries and clinical contexts. Implementation could facilitate earlier and more accurate detection, particularly at the MCI stage where interventions may have greatest impact, addressing practical barriers to systematic cognitive screening in settings where specialist evaluation is unavailable or delayed.

5. LIMITATION

Several limitations of this study should be acknowledged. First, the ADNI cohort consists primarily of highly educated, predominantly White research volunteers, which may limit generalizability to more diverse clinical populations. External validation in heterogeneous, community-based cohorts will be essential to assess robustness and determine whether model recalibration is needed for populations with different socioeconomic backgrounds or levels of cognitive reserve.

Second, ADNI diagnoses are based on expert clinical consensus rather than neuropathological confirmation. Although clinical diagnoses show good agreement with post-mortem findings (approximately 85% accuracy) [28], some misclassification is unavoidable and may constrain achievable model performance. Future work could evaluate whether incorporating biomarker data (e.g., APOE genotype, amyloid PET, CSF measures) improves diagnostic accuracy, particularly for early MCI detection. We intentionally focused on cognitive assessments alone, as these remain the most accessible and cost-effective screening tools in routine clinical practice.

Third, while the model performs well for first-visit patients, optimal performance is achieved when longitudinal data are available, highlighting the value of cognitive trajectory tracking. Importantly, first-visit accuracy (~82%) still substantially exceeds rule-based approaches (58%), supporting clinical utility even in initial screening.

Future work should prioritize external validation in diverse clinical settings, extension to prognostic modeling of disease progression (e.g., MCI-to-AD conversion), and refinement to distinguish Alzheimer's disease from other dementia subtypes. Prospective evaluation within real-world clinical workflows will be critical to assess feasibility and clinical impact.

DATA AVAILABILITY

Data used in this study are available from the ADNI database (adni.loni.usc.edu). Access requires submission of a research proposal and institutional review board approval.

ACKNOWLEDGEMENT

Chanidapa Winalai is a PhD student enrolled at the Ecole Doctorale Ingénierie pour la Santé, la Cognition et l'Environnement (ED-ISCE) of Grenoble Alpes University and conducted her research at the EPSP – TIMC Laboratory. She is supported by co-financing from the French Embassy in Thailand, managed by Campus France, and Naresuan University, Phitsanulok, Thailand. Their financial support and contributions have been invaluable to the completion of this research project.

FUNDING SOURCES

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

1. World Health, O. Dementia Fact Sheet. 2023; Available from: <https://www.who.int/news-room/fact-sheets/detail/dementia>.
2. Petersen, R.C., et al., Practice guideline update summary: Mild cognitive impairment [RETIRED]: Report of the Guideline Development, Dissemination, and Implementation Subcommittee of the American Academy of Neurology. *Neurology*, 2018. 90(3): p. 126-135.
3. Dubois, B., et al., Research criteria for the diagnosis of Alzheimer's disease: revising the NINCDS-ADRDA criteria. *Lancet Neurol*, 2007. 6(8): p. 734-46.
4. Folstein, M.F., S.E. Folstein, and P.R. McHugh, "Mini-mental state". A practical method for grading the cognitive state of patients for the clinician. *J Psychiatr Res*, 1975. 12(3): p. 189-98.
5. Pfeffer, R.I., et al., Measurement of functional activities in older adults in the community. *J Gerontol*, 1982. 37(3): p. 323-9.
6. Elwood, R.W., The Wechsler Memory Scale-Revised: psychometric characteristics and clinical application. *Neuropsychol Rev*, 1991. 2(2): p. 179-201.
7. Rosen, W.G., R.C. Mohs, and K.L. Davis, A new rating scale for Alzheimer's disease. *Am J Psychiatry*, 1984. 141(11): p. 1356-64.
8. Myszczyńska, M.A., et al., Applications of machine learning to diagnosis and treatment of neurodegenerative diseases. *Nat Rev Neurol*, 2020. 16(8): p. 440-456.
9. Rathore, S., et al., A review on neuroimaging-based classification studies and associated feature extraction methods for Alzheimer's disease and its prodromal stages. *Neuroimage*, 2017. 155: p. 530-548.
10. Moradi, E., et al., Machine learning framework for early MRI-based Alzheimer's conversion prediction in MCI subjects. *Neuroimage*, 2015. 104: p. 398-412.
11. Zhang, D., et al., Multimodal classification of Alzheimer's disease and mild cognitive impairment. *Neuroimage*, 2011. 55(3): p. 856-67.
12. Sarica, A., A. Cerasa, and A. Quattrone, Random Forest Algorithm for the Classification of Neuroimaging Data in Alzheimer's Disease: A Systematic Review. *Front Aging Neurosci*, 2017. 9: p. 329.
13. Petersen, R., et al., Alzheimer's Disease Neuroimaging Initiative ({ADNI}): clinical characterization. *Neurology*, 2010. 74(3): p. 201-209.
14. Nasreddine, Z.S., et al., The Montreal Cognitive Assessment, MoCA: a brief screening tool for mild cognitive impairment. *J Am Geriatr Soc*, 2005. 53(4): p. 695-9.

15. Reitan, R.M., Validity of the Trail Making Test as an Indicator of Organic Brain Damage. Perceptual and Motor Skills, 1958. 8(3): p. 271-276.
16. Morris, J.C., et al., The Consortium to Establish a Registry for Alzheimer's Disease (CERAD). Part I. Clinical and neuropsychological assessment of Alzheimer's disease. Neurology, 1989. 39(9): p. 1159-65.
17. Tombaugh, T.N. and N.J. McIntyre, The mini-mental state examination: a comprehensive review. J Am Geriatr Soc, 1992. 40(9): p. 922-35.
18. Brown, L.D., T.T. Cai, and A. DasGupta, Interval Estimation for a Binomial Proportion. Statistical Science, 2001. 16(2): p. 101-117.
19. Mc, N.Q., Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika, 1947. 12(2): p. 153-7.
20. McKhann, G.M., et al., The diagnosis of dementia due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimers Dement, 2011. 7(3): p. 263-9.
21. Albert, M.S., et al., The diagnosis of mild cognitive impairment due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimers Dement, 2011. 7(3): p. 270-9.
22. Petersen, R.C., et al., Apolipoprotein E status as a predictor of the development of Alzheimer's disease in memory-impaired individuals. Jama, 1995. 273(16): p. 1274-8.
23. Pernecky, R., et al., Impairment of activities of daily living requiring memory or complex reasoning as part of the MCI syndrome. Int J Geriatr Psychiatry, 2006. 21(2): p. 158-62.
24. Mitchell, A.J., A meta-analysis of the accuracy of the mini-mental state examination in the detection of dementia and mild cognitive impairment. J Psychiatr Res, 2009. 43(4): p. 411-31.
25. Young, J., et al., Accurate multimodal probabilistic prediction of conversion to Alzheimer's disease in patients with mild cognitive impairment. Neuroimage Clin, 2013. 2: p. 735-45.
26. Suk, H.I., S.W. Lee, and D. Shen, Deep ensemble learning of sparse regression models for brain disease diagnosis. Med Image Anal, 2017. 37: p. 101-113.
27. Kruthika, K.R., Rajeswari, and H.D. Maheshappa, Multistage classifier-based approach for Alzheimer's disease prediction and retrieval. Informatics in Medicine Unlocked, 2019. 14: p. 34-42.
28. Beach, T.G., et al., Accuracy of the clinical diagnosis of Alzheimer disease at National Institute on Aging Alzheimer Disease Centers, 2005-2010. J Neuropathol Exp Neurol, 2012. 71(4): p. 266-73.