scientific reports

OPEN

XGBoost models based on non imaging features for the prediction of mild cognitive impairment in older adults

Miguel A. Fernández-Blázquez[1]✉, José M. Ruiz-Sánchez de León[2], Rubén Sanz-Blasco[2], Emilio Verche[3], Marina Ávila-Villanueva[1], María José Gil-Moreno[4,5], Mercedes Montenegro-Peña[2,6], Carmen Terrón[7,8], Cristina Fernández-García[9] & Jaime Gómez-Ramírez[10]

The global increase in dementia cases highlights the importance of early detection and intervention, particularly for individuals at risk of mild cognitive impairment (MCI), a precursor to dementia. The aim of this study is to develop and validate machine learning (ML) models based on non-imaging features to predict the risk of MCI conversion in cognitively healthy older adults over a three-year period. Using data from 845 participants aged 65 to 87 years, we built five eXtreme Gradient Boosting (XGBoost) models of increasing complexity, incorporating demographic, self-reported, medical, and cognitive variables. The models were trained and evaluated using robust preprocessing techniques, including multiple imputation for missing data, Synthetic Minority Oversampling Technique (SMOTE) for class balancing, and SHapley Additive exPlanations (SHAP) for interpretability. Model performance improved with the inclusion of cognitive assessments, with the most comprehensive model (Model 5) achieving the highest accuracy (86%) and area under the curve (AUC = 0.8359). Feature importance analysis revealed that variables such as memory tests, depressive symptoms, and age were significant predictors of MCI conversion. In addition, an online risk calculator has been developed and made available free of charge to facilitate clinical use and provide a practical, cost-effective tool for early detection in diverse healthcare settings (https://aimar-project.shinyapps.io/MCI-risk-calculator/). This study highlights the potential of non-imaging ML models for early detection of MCI and emphasizes their accessibility and clinical utility. Future research should focus on validating these models in different populations and examining their integration with personalized intervention strategies to reduce dementia risk.

Keywords Aging, Dementia, Early diagnosis, Machine learning, Mil cognitive impairment, XGBoost

The progressive aging of the global population is a phenomenon driven by increased life expectancy and declining birth rates, leading to a demographic shift with significant implications for public health systems worldwide and the maintenance of well-being and quality of life standards. According to the United Nations (2020), the number of people aged 65 years or older is expected to double from 703 million in 2019 to over 1.5 billion in 2050[1], representing a growing portion of the population vulnerable to age-related diseases, including dementia, and in particular Alzheimer's disease (AD). Dementia affects millions of individuals worldwide, leading to

[1]Department of Biological and Health Psychology, School of Psychology, Universidad Autónoma de Madrid, C/ Iván Pavlov, 6, Madrid 28049, Spain. [2]Department of Experimental Psychology, Cognitive Processes and Speech Therapy, Universidad Complutense de Madrid, Madrid, Spain. [3]Department of Psychobiology and Methodology in Behavioural Sciences, Universidad Complutense de Madrid, Madrid, Spain. [4]Health Research Institute of the San Carlos Clinic Hospital (IdISSC), Madrid, Spain. [5]Neurology Department, Hospital Clínico San Carlos, Madrid, Spain. [6]Centre for the Prevention of Cognitive Impairment, Madrid Salud, Madrid City Council, Madrid, Spain. [7]Nuestra Señora del Rosario University Hospital, Madrid, Spain. [8]Sanitas La Zarzuela University Hospital, Madrid, Spain. [9]Cognitive Impairment Unit, Neurology Service, Sanitas La Moraleja University Hospital, Madrid, Spain. [10]Psychophysiology and Neuroimaging Group, Institute of Biomedical Research Cadiz (INiBICA), Cadiz, Spain. ✉email: miguelangel.fernandezb@uam.es

significant cognitive decline and impacting daily functioning and quality of life[2]. It not only causes a progressive and irreversible deterioration of the patient's mental faculties but is also the leading cause of dependency and, collaterally, represents a substantial economic, social, and personal cost. Its prevalence is expected to rise substantially, with estimates suggesting that the number of people with dementia will triple in the next 30 years, from 50 million to 152 million by 2050[3]. The situation is even more concerning considering that this type of dementia does not appear to be reversible once diagnosed, and there have been no truly significant scientific advances in this regard for over 20 years. However, the onset of dementia does not occur suddenly; on the contrary, the initial pathophysiological and cognitive changes associated with dementia, although subtle, can be observed up to two decades before clinical diagnosis in the absence of clearly perceptible symptoms[4,5]. Therefore, there is a large window of time in which it is possible to identify individuals at risk of future dementia and then subject them to personalized prevention programs. The problem is that there are currently no accurate tools for early detection of these individuals in preclinical stages. It is precisely these pre-dementia stages that are of utmost importance. Implementing effective therapeutic interventions at these stages could potentially halt or even reverse disease progression[6].

Apart from age, which is the main known risk factor for the development of dementia, potentially modifiable risk factors have been identified that show a consistent association with AD and other related dementias. Precisely, the Lancet Commission on Dementia Prevention, Intervention, and Care has been studying this issue in the last years. Since its inception in 2017, the Commission has assessed evidence related to potentially modifiable dementia risk factors, synthesizing scientific data from various meta-analyses and intervention studies. So, the 2017 report found the potential for dementia risk reduction by addressing twelve established risk factors across the lifespan: low education, hearing loss, hypertension, smoking, obesity, depression, physical inactivity, diabetes, excessive alcohol consumption, traumatic brain injury, air pollution, and social isolation[7]. The recent 2024 report reaffirms the role of the previously identified risk factors and adds two more: midlife elevated LDL cholesterol and untreated vision loss[8]. This finding means that through a combination of factors, it would be possible to identify individuals at high risk of developing dementia very early. This combination of factors could be non-linear and, therefore, not intuitive. For example, unlike in younger age groups, from the age of 85, elevated levels of hemoglobin A1c[9] cholesterol[10] or homocysteine[11] are associated with a lower risk of dementia. Thus, to unravel the role and actual weight of each factor in the onset of dementia, computational techniques based on artificial intelligence (AI) should be used, as they allow for handling massive data and developing dynamic algorithms that help trace personalized evolutionary trajectories[12].

Recent advancements in AI and machine learning (ML) offer promising alternatives for early dementia diagnosis using non-imaging-based features. ML has emerged as a powerful tool in medical diagnostics, offering the potential to analyze complex datasets and identify patterns that may not be apparent to human observers. In the context of dementia diagnosis, ML algorithms can be trained on large datasets containing demographic, cognitive, socioeconomic, clinical information, lifestyles, and neuroimaging to identify patterns indicative of dementia and develop predictive models. By leveraging ML, it is possible to develop diagnostic tools that are not only accurate but also cost-effective and easily integrated into clinical practice to enable early intervention in dementia[13]. In fact, in the last few years, several studies have been published that build predictive models using AI and ML to identify individuals at high risk of dementia. These risk prediction models have been developed primarily based on a combination of cognitive test scores, sociodemographic, health, lifestyle, and environmental variables, with varying levels of accuracy in distinguishing outcomes[14]. For example, Barnes et al. (2009) developed one of the first models to consider a late-life dementia risk index[15]. The model used logistic regression on 3,375 subjects and features such as demographics, cognition, physical activity, neuroimaging, social behavior, among others, to identify individuals at risk of dementia within six years. Exalto et al. (2013) reported a Cox proportional hazards model using a combination of 45 predictors from self-reported data, medication, laboratory data, and clinical histories in a cohort of 29,961 individuals with type 2 diabetes for a 10-year dementia risk[16]. Nori et al. (2019) used ML with a large dataset including over 10,000 features to predict 5-year dementia risk[17]. A more recent model presented a multi-diagnostic approach using structural magnetic resonance imaging (MRI) and ML to diagnose mild cognitive impairment (MCI) and AD[18]. The researchers used data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) and the Open Access Series of Imaging Studies (OASIS), training classifiers, such as support vector machines (SVMs) and neural networks. The classifiers demonstrated high accuracy, with the "healthy controls (HC) vs. AD" classifier achieving a balanced accuracy of 90.6%.

Although the results of these models are promising, the major limitation is that it may be very difficult for a clinician to apply the algorithms as their availability via online applications is quite limited. In addition, another weakness is that it may not be possible to collect a large set of variables for each patient, which usually include MRIs, cognitive tests, self-reported data, laboratory results, etc. A final limitation is that most ML algorithms rely on imaging techniques such as MRI and positron emission tomography (PET), which may not be readily available to physicians, especially in primary care. On the contrary, incorporating non-imaging features such as demographic information, cognitive test results, clinical history and lifestyles are often easier and less expensive to collect compared to imaging data, although the relationship between all of these features is not yet fully understood.

Objectives of our study

The main goal of the study was to develop ML models to predict MCI in people over 65 years of age over a three-year period. For this purpose, only non-imaging features such as demographic, clinical and cognitive variables as well as lifestyle data were analyzed. All of this data can be collected relatively easily through a short clinical interview and the use of specific cognitive tests. A total of five prediction models were developed, which included a smaller or larger number of variables depending on the ease of collecting information.

The specific objectives were:

1. To identify the most relevant features for MCI diagnosis.
2. To develop and train five ML models using these features.
3. To evaluate the performance of the models using multiple metrics and selecting the best hyperparameters.
4. To implement the most powerful models as an online diagnostic tool and make them available to primary and specialized healthcare professionals to assist them in data-driven clinical decision-making.

Developing an online ML diagnostic tool for MCI offers several potential benefits for physicians. First, it can improve the accessibility and affordability of early dementia diagnosis, particularly in low-resource settings. Second, it can facilitate early intervention and management, potentially slowing the progression of the disease and improving patient outcomes. Finally, it can reduce the burden on healthcare systems by enabling more efficient and cost-effective screening and diagnosis.

Methods

In this study, we adopted a robust machine learning framework that leverages the eXtreme Gradient Boosting (XGBoost) algorithm to predict conversion to MCI. As shown in Fig. 1, the methodology encompasses several critical stages, including data split, data preparation, feature engineering, model training with hyperparameter optimization, evaluation, and interpretability analysis. We built all of our machine-learning models in R v.4.2.1.

Data source

The participants of this study comprised 845 cognitively healthy individuals aged 65 to 87 years at baseline, living at home, without relevant psychiatric, neurological or systemic disorders. All of them were part of two equivalent datasets from two previous funded projects namely *Elderly Relatives of Alzheimer's Disease (ERAD): Influence of Subjective Cognitive Decline* and *Factors associated with healthy and pathologically aging in a sample of elderly*
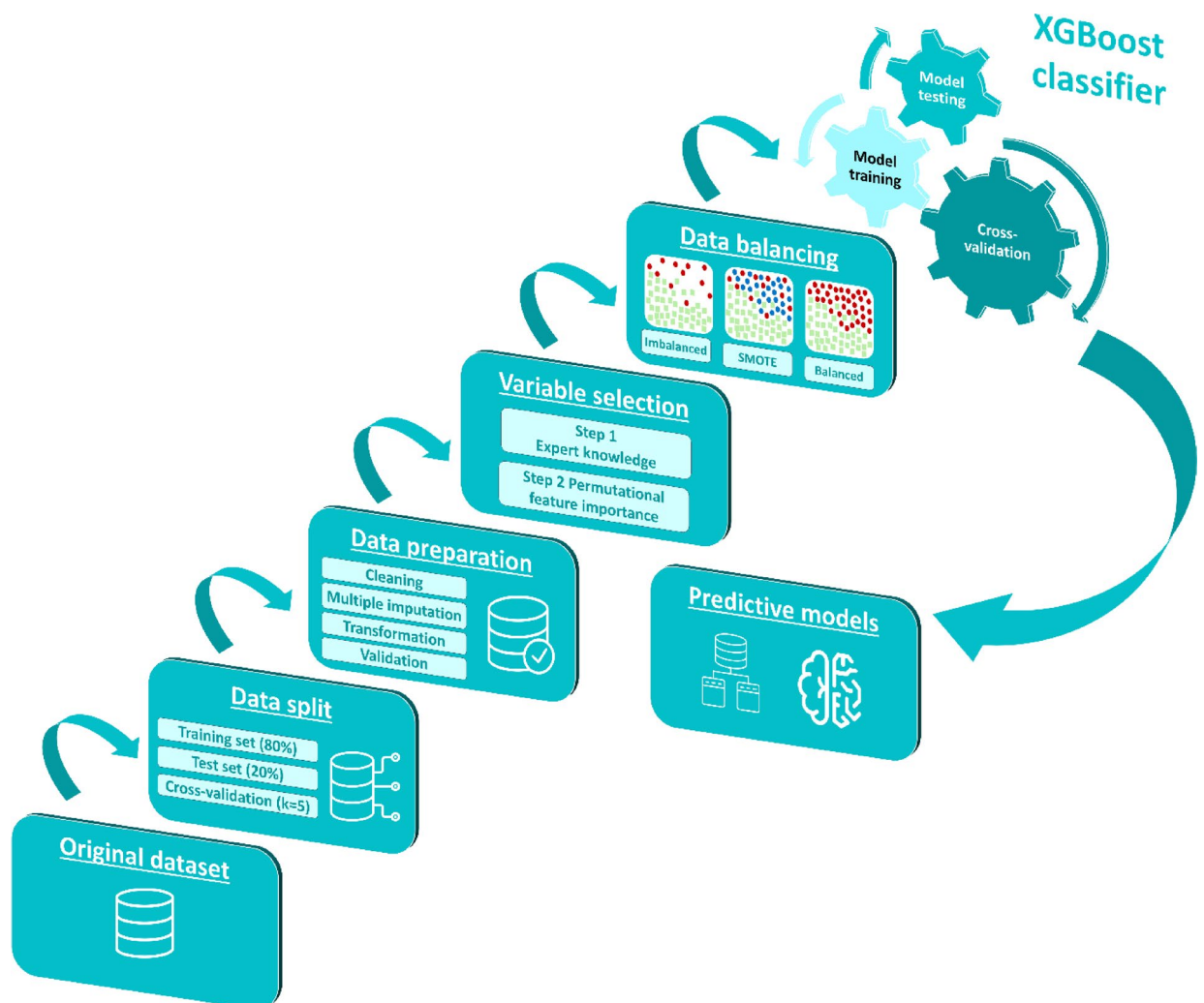
Fig. 1. Schematic diagram of the study workflow.

people over 90 in the city of Madrid (MADRID + 90). The main goal of the two projects was to both improve the early detection and prevention of MCI and dementia and to increase knowledge about the determinants of healthy aging[19].

After signing the informed consent, all participants underwent a detailed assessment protocol, which was usually completed within four hours with appropriate breaks. The full assessment included a clinical semi-structured interview and a neurological and neuropsychological examination. Therefore, our dataset contains information on a variety of variables grouped in five categories (Supplementary Table S1): Sociodemographic, Self-reported from a clinical interview, Medical examination, and Cognitive assessment.

The participants were followed for three years. Clinical diagnoses were always agreed upon in consensus discussions between neurologists and neuropsychologists. When available, MRIs were considered to exclude the presence of macroscopic lesions or significant vascular damage. Each participant was then independently diagnosed after each visit based on their age, gender, cognitive reserve, functional information, and cognitive scores. However, the diagnoses were not based on psychometrically stable thresholds values, but rather on clinical impressions. The NIA-AA's criteria were used to diagnose core MCI[20] which included: (i) self- or informant-reported cognitive complaints; (ii) objective evidence of impairment in one or more cognitive domains; (iii) preserved independence in functional abilities; and (iv) not demented. All participants were diagnosed as cognitively healthy at baseline.

Data preprocessing

First, the dataset was divided into training and testing sets by 80% and 20%, respectively. Then, essential R packages were loading, including *xgboost*[21] for model training, *ParBayesianOptimization*[22] for hyperparameter tuning, *caret*[23] for data partitioning, *pROC*[24] for performance metrics, *ggplot2*[25] for visualization, and additional packages such as *mice*[26] *smotefamily*, and *ROSE*[27] for data preprocessing and balancing the training dataset. A seed value of 42 was set to ensure that our results were reproducible across different runs.

The dataset included various features along with the target variable, mci_conversion, which was transformed into a factor with levels "No" and "Yes" to represent non-conversion and conversion to MCI, respectively. A methodology was used to clean and prepare the features, which included the identification of numerical and categorical variables. We then conducted a preliminary analysis of the baseline data to find out their distribution and examine the nature and distribution of missing values. Then, we performed a multiple imputation procedure under a fully conditional specification method to impute values that are as close as possible to the ideal predicted observations. These imputed values were generated based on existing variables. The imputation procedure replaced each missing observation with a set of plausible values representing the uncertainty about the appropriate value to be imputed. The process was repeated five times, yielding the corresponding five data sets, the coefficients of which varied from set to set. The imputed data sets were analyzed using the usual full data procedure. Distributions of imputed and observed values were compared using density plots, showing similar patterns across all variables. Convergence was monitored across − 20 iterations and confirmed using a convergence threshold of 1e-4. In addition, a sensitivity analysis comparing results across different imputation models showed comparable conclusions, supporting the robustness of our approach. Finally, the results of these analyses were combined to draw valid statistical inferences of the data. The *mice* package for R was used to perform the multiple imputation procedure[26]. No normalizations of variables were applied because the tree models in the various experiments did not show significant improvements when these transformations were added to the pipeline.

Selection of variables

A two-step process was used to select those variables that finally were included in the prediction models. First, a set of variables were selected if scientific evidence and expert knowledge give them a relevant role in the diagnosis of MCI (e.g. age, education, SCD, etc.). Second, a permutational feature importance method was then applied so that two random variables were generated and added to the original dataset. Next, the model was trained and the feature importance was calculated. Variables were selected if their feature importance exceeded that of at least one of the previously included random variables.

Class imbalance

Our dataset exhibited class imbalance, with a higher proportion of cognitively healthy stable individuals compared to MCI converters. To address this class imbalance, we decided to apply the Synthetic Minority Oversampling Technique (SMOTE) on the training set, a widely used method in machine learning for generating synthetic samples to balance the distribution of classes. This technique generates synthetic samples for the minority class to balance the class distribution and thereby improve the performance of classifiers. The SMOTE algorithm enables the generation of new, unique samples and reduces the risk of overfitting using the nearest neighbors technique.

Model training

After experimenting with several machine-learning algorithms (i.e., random forest, logistic regression, and gradient boosting), our choice was XGBoost, a high-performance supervised learning algorithm based on decision trees, designed to optimize both speed and performance. It is based on the principle of boosting, a technique that combines multiple weak models (in this case, decision trees) to create a stronger model[21]. The key idea is to sequentially add trees to the model, where each new tree aims to correct the errors made by the previous trees. This iterative process continues until a specified number of trees are added or the model's performance stops improving. As compared to other gradient-boosting algorithms, previous literature has found that XGBoost provides an optimal balance between accuracy and training speed[28].

We built five XGBoost models with the goal of predicting MCI conversion over a three-year period. The XGBoost classifier had two potential classes eligible for its output: stable cognitively healthy individuals and MCI converters. The reasons for developing the five prediction models, which incorporate increasing complexity and a larger number of variables, are described in detail below. Initially it was considered appropriate to design predictive models based on the potential availability of necessary information. Some features can be easily and directly obtained from a basic clinical interview conducted by any healthcare professional. However, other features require more effort and a certain level of clinical knowledge, as they require either a short medical examination or the application of specific cognitive tests. By integrating all these possibilities, the development of the following five models was deemed appropriate:

- Model 1: Demographics and self-reported features.
- Model 2: Demographics, self-reported and medical features.
- Model 3: Demographics, self-reported and cognitive screening features.
- Model 4: Demographics, self-reported, medical and cognitive screening features.
- Model 5: Demographics, self-reported, medical, cognitive screening and further cognitive assessment features.

Hyperparameter optimization and cross-validation

Hyperparameter optimization plays a crucial role in machine learning model development, particularly for complex algorithms like XGBoost, which contains over thirty configurable hyperparameters. These parameters significantly influence model performance and must be carefully tuned to account for the unique characteristics of each dataset. Our study implemented a comprehensive optimization and validation strategy that integrated Bayesian optimization with robust cross-validation methodology.

The optimization process was conducted using the *ParBayesianOptimization* package, with an objective function designed to maximize the area under the receiver operating characteristic curve (AUC). The process began with an initial data split of 80–20 for training and testing sets, respectively. Within the training phase, we implemented 5-fold cross-validation to ensure robust performance estimation during hyperparameter optimization. Our optimization strategy focused on six key hyperparameters selected for their significant impact on model complexity and generalization ability: maximum depth (ranging from 3 to 10), learning rate (between 0.01 and 0.3), gamma (from 0 to 0.2), column sampling ratio (between 0.6 and 1.0), minimum child weight (from 1 to 10), and subsample ratios (between 0.6 and 1.0).

The Bayesian optimization process consisted of 20 total iterations, beginning with an exploration phase of 10 random configurations to establish a baseline understanding of the hyperparameter space. The subsequent optimization phase was guided by the "expected improvement" acquisition function, systematically exploring promising regions of the parameter space. The objective function implementation used 'binary: logistic' with AUC as the evaluation metric, returning the maximum AUC score achieved during cross-validation for each parameter combination evaluated.

To ensure model robustness and prevent overfitting, several safeguards were implemented. First, early stopping mechanism that halted training if no improvement was observed after 10 rounds. Second, parameter validation checks that returned -Inf for invalid configurations, steering the optimization toward feasible solutions. Third, bootstrap validation ($n = 1000$) to calculate 95% confidence intervals for all performance metrics. Fourth, comprehensive cross-validation during the optimization process to ensure reliable performance estimation. Thus, this integrated approach to hyperparameter optimization and validation ensured both optimal model performance and reliable generalization capability. The combination of Bayesian optimization with rigorous cross-validation provided a robust framework for model development while maintaining computational efficiency.

Final model training and evaluation

Upon identifying the optimal hyperparameters, the final XGBoost model was trained. The model was configured using binary logistic loss as the objective function for training. This loss function transforms the model's outputs through a logistic function to produce probabilities between 0 and 1, making it appropriate for our binary classification task of predicting MCI conversion. The model's performance was evaluated using the AUC metric, which measures the model's ability to distinguish between the two classes. The model was configured with a maximum limit of 100 boosting rounds, with the actual number of iterations determined dynamically by the model's performance on the validation set. Predictions on the test set were generated, and a threshold of 0.5 was applied to classify predictions as "Yes" or "No". The model's performance was evaluated using a confusion matrix to compare predicted classifications against actual outcomes. Key performance metrics, including accuracy, precision, recall, F1 score, and AUC, were calculated to provide a comprehensive evaluation of the model's predictive power. The ROC curve, plotted using *ggplot2* package, visualized the trade-off between true positive rate (sensitivity) and false positive rate (1-specificity).

Feature importance and SHAP analysis

XGBoost classifiers also allow for individual input variable importance to be evaluated, providing insights into which features had the most significant impact on model predictions. To further enhance model interpretability, we conducted feature selection using SHapley Additive exPlanations (SHAP)[29] a unified, model-agnostic framework for evaluating variable importance at both the model and individual prediction levels. SHAP values, which offer a detailed view of how individual features contribute to each prediction, were calculated using the *SHAPforxgboost* package[30]. The SHAP summary plot depicted the distribution of SHAP values across all features,

highlighting those with the most substantial impact on predictions. This analysis not only aids in validating the model's decisions with domain knowledge but also enhances transparency and trust in the model's predictions.

Ethics approval

Ethical approval for the ERAD and MADRID + 90 studies was granted by the Research Ethics Committee of the Hospital Universitario Clínico San Carlos (Madrid, Spain) and the Ethics Committee of the Carlos III Institute of Health (Madrid, Spain), respectively. Written informed consent was obtained from all participants. The authors confirm that all procedures contributing to this work comply with the ethical standards of the relevant national and institutional committees on human experimentation and with the 1964 Helsinki Declaration and its later amendments. All methods were performed in accordance with the relevant guidelines and regulations.

Results

The study commenced with an examination of baseline characteristics between cognitively stable individuals and those who eventually converted to MCI over time (14.2% of the sample). As depicted in Table 1, the baseline characteristics were analyzed using t-tests for continuous variables and chi-square tests for categorical variables. These assessments provided foundational insights into the demographic and clinical attributes differentiating the two groups at the study's onset.

Preprocessing and selection of variables

The original dataset comprised 219 variables, necessitating substantial preprocessing and cleaning to enhance analytical clarity and model efficiency. A total of 148 variables were excluded due to a high incidence of missing values, redundancy, or lack of informativeness. This reduction process resulted in a refined dataset containing 71 variables. These underwent multiple imputation to handle missing data.

An important step in preprocessing was the application of the permutational feature importance method, shown in Fig. 2. This technique prioritized features based on their predictive power and revealed cognitive variables, particularly immediate free recall on the Free and Cued Selective Reminding Test (FCSRT), as very informative. Integration of demographic, self-reported, and medical parameters was also found to be crucial in predicting MCI risk. Expert knowledge and these analytical insights guided the selection of 29 features for building of the predictive models, as detailed in Table 2. In the table, these features are categorized into demographic, self-reported, medical, cognitive screening and further cognitive assessment groups. Model 1, the simplest, uses 13 features from the demographic and self-reported categories. In contrast, Model 5, the most complex, includes all 29 features and effectively integrates demographic, self-reported, medical, and complete cognitive data.

Models performance

The study developed five predictive models, each with different levels of complexity, to predict the conversion to MCI. To this end, we introduced a robust machine learning framework using XGBoost classifiers. The performance of each model was improved through careful hyperparameter tuning. The process included optimization of parameters such as maximum depth, learning rate (eta), gamma, column sampling ratio, minimum child weight, and subsample ratios. These parameters were fine-tuned using Bayesian optimization, which efficiently explored the hyperparameter space to find the optimal settings for each model configuration. Table 3 shows the best combination of hyperparameters for each model.

As shown in Table 4, Model 1 achieved 82.25% accuracy, 93.10% recall, 87.10% precision, 89.99% F1-score, and an AUC of 0.6687 across all models, indicating solid underlying performance. Model 2, which included additional medical parameters, showed slight improvements with an accuracy of 82.30%, a recall of 92.41%, a precision of 87.58%, an F1-score of 89.90%, and an AUC of 0.7293, reflecting improved sensitivity and discrimination ability. Model 3, which included cognitive screening, showed an accuracy of 82.35%, recall of 89.66%, precision of 89.66%, F1-score of 89.66%, and AUC of 0.7682, highlighting the crucial influence of cognitive features. Model 4, which included medical and cognitive screening measures, showed a better general performance with an accuracy of 82.84%, a recall of 86.84, precision of 93.62%, F1-score of 89.93%, and AUC of 0.7744. Finally, Model 5, the most comprehensive, achieved the highest metrics with an accuracy of 86.39%, recall of 90.13%, precision of 94.48%, F1-score of 92.22%, and AUC of 0.8359, demonstrating superior balance and robustness in predictive performance by integrating a full spectrum of demographic, medical, and cognitive data. Confusion matrices and ROC curves for all models are shown in the Supplementary Fig. S1.

### Model validation and overfitting assessment

To ensure robust generalization of our models, we conducted comprehensive analyses using training and validation loss curves. The pattern of misclassifications errors for learning curves revealed low training errors at the beginning followed by stabilization while the validation performance was improving as sample sizes increased, obtaining final validation errors ranging from 0.10 to 0.15. This approach demonstrated that our models maintain their predictive power when applied to new, unseen data. Detailed visualizations and metrics are provided in the Supplementary Fig. S2.

Feature importance and SHAP values

The SHAP value analysis revealed distinct patterns of feature importance across five increasingly complex models. As shown in Fig. 3, GDS emerged as the dominant predictor across Models 1–4, demonstrating consistent and robust influence in MCI prediction. Age, SCD, and family history of AD also maintained notable relevance across different model configurations. Clinical self-reported parameters demonstrated model-specific

	Total (N = 845)	Cognitively healthy (n = 725)	MCI converters (n = 120)	p-value§
Age [mean (SD)]	74.01 (3.77)	73.74 (3.64)	75.66 (4.10)	< 0.001
Gender (%)				
Female	520 (61.5)	444 (61.2)	76 (63.3)	0.687
Male	325 (38.5)	281 (38.8)	44 (36.7)	
Education (%)				
Illiterate	143 (16.9)	121 (16.7)	22 (18.3)	0.057
Primary education	232 (27.5)	188 (25.9)	44 (36.7)	
Secondary education	224 (26.5)	200 (27.6)	24 (20.0)	
Higher education	246 (29.1)	216 (29.8)	30 (25.0)	
Occupation (%)				
Unskilled manual labor	208 (24.6)	171 (23.6)	37 (30.8)	0.355
Skilled manual labor	203 (24.0)	178 (24.6)	25 (20.8)	
Office clerk	157 (18.6)	132 (18.2)	25 (20.8)	
Professional	189 (22.4))	167 (23.0)	22 (18.3)	
Management	88 (10.4)	77 (10.6)	11 (9.2)	
SCD scale[31] [mean (SD)]	4.22 (2.24)	4.07 (2.16)	5.13 (2.45)	< 0.001
Age at SCD onset [mean (SD)]	67.12 (7.13)	66.71 (6.94)	69.56 (7.78)	< 0.001
GDS-15[32] score [mean (SD)]	1.55 (2.20)	1.40 (2.11)	2.45 (2.46)	< 0.001
EQ5D[33] thermometer [mean (SD)]	76.01 (14.96)	76.55 (14.86)	72.72 (15.20)	0.009
Hypertension (%)				
No	415 (49.1))	356 (49.1)	59 (49.2)	0.990
Yes	430 (50.9)	369 (50.9)	61 (50.8)	
Cholesterol (%)				
No	384 (45.4)	319 (45.4)	55 (45.8)	0.926
Yes	461 (54.6)	396 (54.6)	65 (54.2)	
Diabetes (%)				
No	756 (89.5)	654 (90.2)	102 (85.09	0.085
Yes	89 (10.5)	71 (9.8)	18 (15.0)	
Number of drugs [mean (SD)]	4.07 (2.55)	4.20 (2.51)	4.36 (2.78)	0.182
Family history of dementia (%)				
No	580 (68.6)	500 (69.0)	80 (66.7)	0.615
Yes	265 (31.4)	225 (31.0)	40 (33.3)	
Body mass index [mean (SD)]	27.27 (3.56)	27.38 (3.54)	26.62 (3.60)	0.030
Tapping right hand [mean (SD)]	60.17 (10.26)	60.37 (10.29)	58.94 (10.06)	0.158
Tapping left hand [mean (SD)]	56.19 (9.20)	56.41 (9.35)	54.89 (8.11)	0.095
Systolic blood pressure [mean (SD)]	143.31 (20.97)	143.08 (20.67)	144.67 (22.79)	0.444
Diastolic blood pressure [mean (SD)]	78.60 (11.93)	78.62 (11.65)	78.43 (13.56)	0.871
MMSE[34] [mean (SD)]	28.69 (1.42)	28.84 (1.27)	27.82 (1.88)	< 0.001
Animals [mean (SD)]	18.82 (4.95)	19.30 (4.86)	15.88 (4.44)	< 0.001
FCSRT[35] free immediate recall [mean (SD)]	23.95 (6.20)	24.81 (5.85)	18.75 (5.70)	< 0.001
FCSRT total immediate recall [mean (SD)]	41.58 (5.16)	42.18 (4.67)	37.98 (6.37)	< 0.001
FCSRT free delayed recall [mean (SD)]	9.56 (2.59)	9.92 (2.40)	7.41 (2.66)	< 0.001
FCSRT total delayed recall [mean (SD)]	14.28 (1.72)	14.69 (1.54)	13.20 (2.17)	< 0.001
ROCF[36] time [mean (SD)]	246.54 (118.70)	239.18 (113.86)	290.97 (136.88)	< 0.001
ROCF copy [mean (SD)]	30.42 (6.53)	30.90 (6.07)	27.47 (8.25)	< 0.001
ROCF immediate recall [mean (SD)]	12.90 (6.14)	13.42 (6.09)	9.71 (5.50)	< 0.001
ROCF delayed recall [mean (SD)]	12.71 (6.21)	13.26 (6.16)	9.37 (5.47)	< 0.001
Digit-Symbol[37] [mean (SD)]	19.90 (7.44)	20.64 (7.40)	15.41 (5.99)	< 0.001

Table 1. Baseline characteristics of the sample for cognitively healthy individuals and MCI converters over time. §t-tests were used for continuous variables, while the chi-square test was used for categorical variables. SCD: Subjective Cognitive Decline; EQ-5D: EuroQol 5D; GDS: Geriatric Depression Scale; MMSE: Mini Mental State Examination; FCSRT: Free and Cued Selective Reminding Test; ROCF: Rey-Osterrieth Complex Figure.
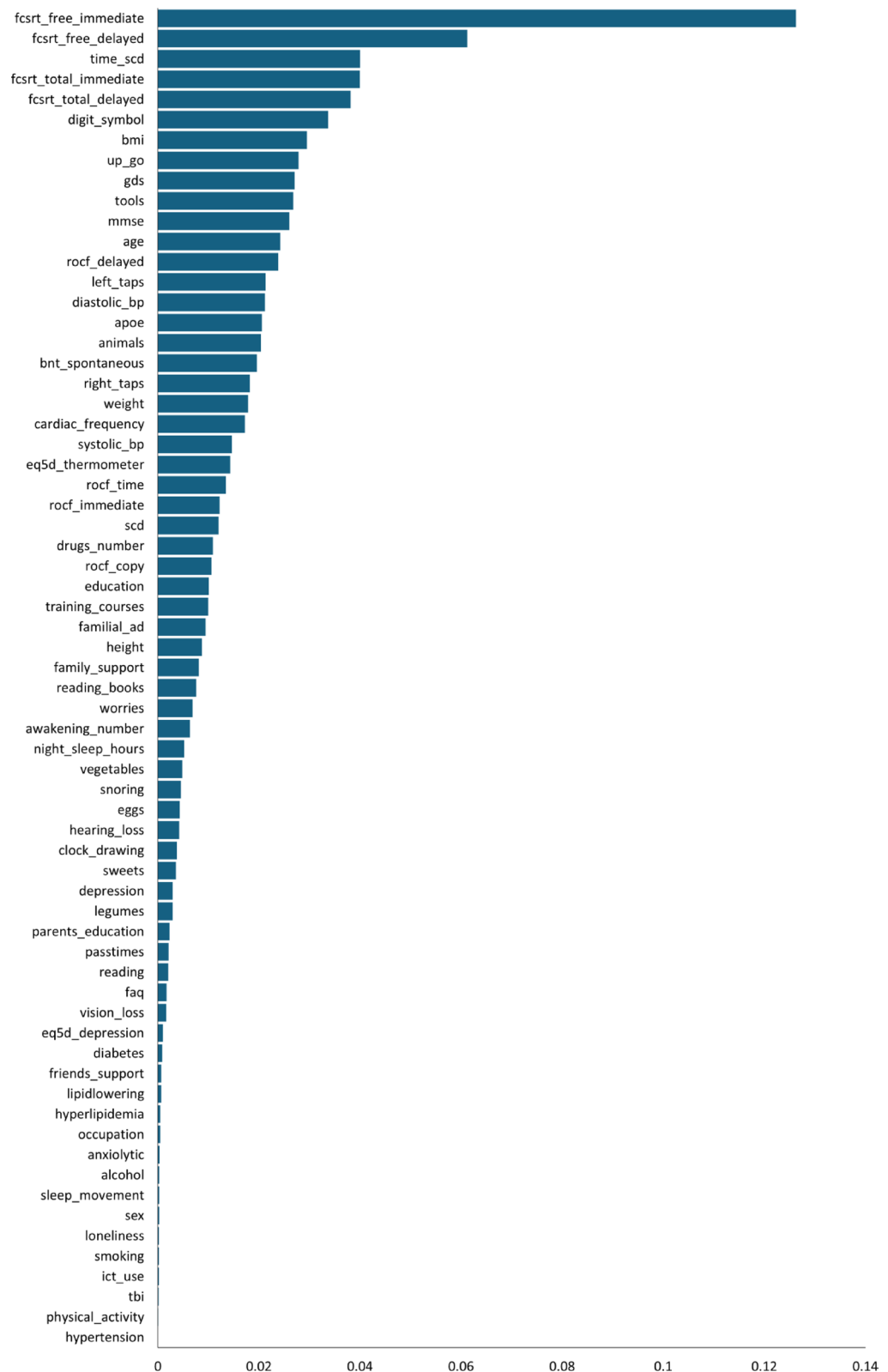
Fig. 2. Feature importance for variable selection. The figure shows the order of importance of the features. The cognitive variables appear to be the most informative, particularly the immediate free recall of the FCSRT. It also shows that a combination of demographic, self-reported and medical parameters are relevant for predicting MCI risk. SCD: Subjective Cognitive Decline; EQ-5D: EuroQol 5D; GDS: Geriatric Depression Scale; MMSE: Mini Mental State Examination; FCSRT: Free and Cued Selective Reminding Test; ROCF: Rey-Osterrieth Complex Figure.

importance (e.g. hyperlipidemia in Model 1, hypertension in Model 2, diabetes in Model 4), but their influence lacked consistency across different model configurations.

In Model 1, beyond GDS's prominence, hyperlipidemia showed significant contribution among self-reported factors. Model 2, incorporating medical variables, showed right-hand tapping emerged as the most influential medical variable. Hypertension also exhibited notable contribution in this model configuration. However, the introduction of cognitive screening variables in Model 3 marked a significant inflection point. MMSE and semantic fluency test (animals) positioned themselves as crucial predictors alongside GDS, while previously important variables assumed secondary roles. This pattern persisted in Model 4, where diabetes emerged as a relevant clinical factor. Model 5, the most comprehensive, revealed a significant shift in predictor hierarchy. For the first time, immediate free episodic memory performance surpassed GDS as the most important predictor. Information processing speed also emerged as a crucial predictor. Notably, even in this more complex model, both GDS and family history of AD maintained their relevance as important predictors.

Discussion

This study implements an advanced machine learning framework utilizing the XGBoost algorithm to predict the conversion to MCI in cognitively healthy older adults. Five predictive models were developed, each progressively adding complexity by incorporating demographic, medical, and cognitive information. The methodology addresses challenges such as class imbalance and missing values by employing modern techniques like SMOTE and multiple imputation. The results highlight that the multimodal approach and the use of tools like SHAP for interpretability significantly improve predictive accuracy. An original aspect of this work is the provision of a free online calculator (https://aimar-project.shinyapps.io/MCI-risk-calculator/), enabling interested readers to apply the model in practice (instructions and more details about the calculator can be found in the Supplementary Note). This study represents progress toward the early detection of MCI and the personalization of preventive interventions.

Main findings

In terms of performance, Models 1 and 2 demonstrated solid baseline accuracy using easily acquired demographic, self-reported, and basic medical data for practitioners. The incorporation of cognitive screening measures in Model 3 and 4 and more complex cognitive tests in Model 5 led to significant improvements in prediction accuracy. Thus, our results highlight the importance of integrating a wide range of data types. Importantly, supplementary analyses confirmed no overfitting across all five models.

Distinct patterns of feature importance across models were revealed. GDS emerged as the dominant predictor across Models 1–4, while immediate free episodic memory performance became the most influential factor in Model 5. This progression highlights how the integration of increasingly sophisticated cognitive assessments enhances predictive accuracy while maintaining the relevance of affective symptoms. Although demographic and self-reported information provide a useful baseline, our comprehensive models that integrate cognitive assessments demonstrated superior prognostic utility in identifying individuals at risk for MCI. This finding aligns with the broader literature emphasizing the importance of cognitive performance in predicting MCI[38,39]. In our study, cognitive variables, particularly memory tests and digit-symbol substitution tasks, were identified as the most critical predictors in the more complex model.

The consistent importance of GDS across models, coupled with the emergence of memory performance as the leading predictor in Model 5, suggests that both affective and cognitive domains play crucial roles in MCI prediction. Notably, family history of AD maintained its relevance as an important predictor even in the most comprehensive model, highlighting the value of considering genetic risk factors. Moreover, while clinical parameters showed model-specific importance (e.g., hyperlipidemia, hypertension, diabetes), their influence varied across different model configurations.

### Variable importance and clinical implications

Our SHAP analysis revealed a hierarchical pattern of predictive factors that provide valuable insights for clinical practice. The basic model identified three critical predictors that should trigger initial clinical attention: GDS (with scores ≥ 5 warranting special attention), age (particularly significant for patients over 70 years), and the presence of cognitive complaints, especially when these factors appear in combination. Family history of AD also maintained consistent importance across models, suggesting its value as an early screening indicator.

Building upon these foundational indicators, our analysis revealed that certain self-reported clinical parameters showed model-specific importance (hyperlipidemia in Model 1, hypertension in Model 2, diabetes in Model 4). Right-hand tapping emerged as the most relevant medical variable in Model 2. These factors should be evaluated as secondary indicators, especially in patients already showing basic risk factors.

Our more sophisticated Models 3–5 highlighted specific cognitive patterns that deserve particular attention in clinical settings. The introduction of cognitive screening measures (MMSE and animals) in Model 3 marked a significant shift in predictive patterns. In Model 5, the most comprehensive, immediate free episodic memory performance emerged as the strongest predictor, surpassing GDS, while information processing speed also showed crucial predictive value.

Based on these findings, we propose a stepwise evaluation approach for practitioners: (i) Initial screening focusing on depressive symptoms, advanced age, and positive family history of AD; (ii) Secondary assessment incorporating specific clinical parameters based on individual risk profiles; and (iii) Comprehensive cognitive evaluation, with particular emphasis on episodic memory and processing speed (i.e. FCSRT and digit-symbol scores below normative ranges). Moreover, it would be advisable to continuously monitor patients presenting multiple risk factors, especially those showing high-risk combinations identified in Models 3–5. This structured

a) Model 1: demographic + self-reported
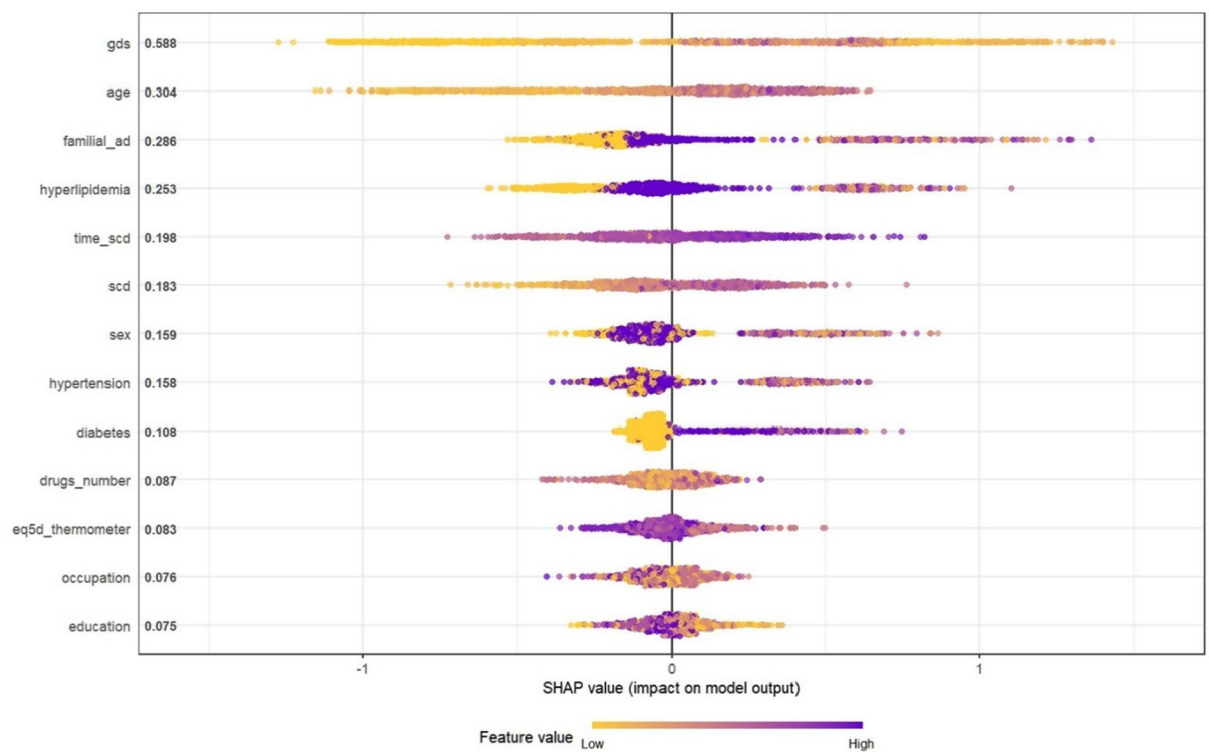
b) Model 2: demographic + self-reported + medical
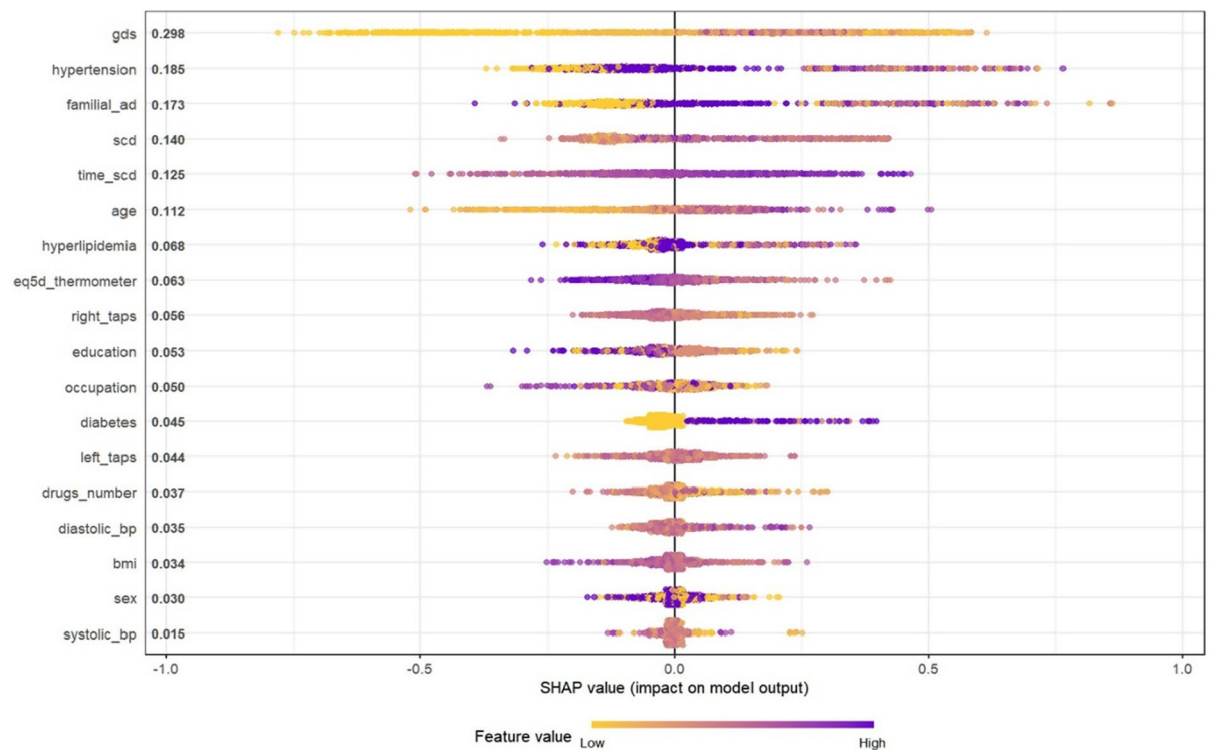
approach emphasizes the importance of both affective symptoms and cognitive performance in MCI prediction, while acknowledging the role of demographic and clinical factors in creating a comprehensive risk profile.

◄**Fig. 3.** SHAP distributions for all models. SHAP beeswarm plots for the five models developed to predict MCI conversion. The x-axis in all subplots represents the SHAP values, which measure the contribution of each feature to the model's predictions: positive values indicate that the feature increases the prediction of the positive class (i.e. MCI conversion), while negative values are assigned to the negative class (i.e. remaining cognitively healthy). The y-axis lists the features used in each model, ranked by their importance based on the mean absolute SHAP values. Finally, the color gradient (from purple to yellow) represents the feature values: purple corresponds to lower feature values, while yellow corresponds to higher feature values. Each subplot corresponds to a different model, with progressively more complex feature sets: (**a**) SHAP values are relatively concentrated, suggesting moderate variability in feature importance, with GDS, age and familial AD appearing to make significant contributions to the model's predictions; (**b**) Inclusion of medical data increases the variability in SHAP values, suggesting that these features provide additional predictive power; (**c**) and (**d**) SHAP values for cognitive screening features are more dispersed, suggesting that these features have a differential influence on predictions; e) In the most comprehensive model, SHAP values show significant variability, indicating that the combination of all feature sets produces the most sophisticated predictions. SCD: Subjective Cognitive Decline; EQ-5D: EuroQol 5D; GDS: Geriatric Depression Scale; MMSE: Mini Mental State Examination; FCSRT: Free and Cued Selective Reminding Test; ROCF: Rey-Osterrieth Complex Figure.

Clinical utility and accessibility

The findings of this study highlight the critical importance of early detection and targeted prevention strategies in addressing the growing burden of MCI and dementia. The successful implementation of an online risk calculator provides a practical solution for clinicians. By relying exclusively on non-imaging features, the calculator overcomes the cost, accessibility, and infrastructure barriers typically associated with neuroimaging-based approaches. In particular, the availability of simpler models (e.g., Model 1) broadens the accessibility of these tools to primary care settings, where cognitive assessments or specialized medical equipment may not be readily available. However, the incorporation of cognitive variables in Models 3-5significantly enhances predictive accuracy, making them more suitable for specialized settings or when cognitive testing can be performed. These findings suggest that the tool's flexibility allows it to be adapted to different healthcare contexts, from resource-limited environments to advanced clinical facilities.

Ethical and data protection implications

The implementation of machine learning models for predicting cognitive decline raises critical ethical considerations that extend beyond technical performance. While our models demonstrate robust predictive accuracy and stable validation performance, this high accuracy carries significant responsibility in clinical practice. The psychological impact of early MCI diagnosis demands careful handling of both positive and negative predictions, necessitating clear protocols for result communication and appropriate support systems. Furthermore, ensuring fairness across different demographic groups is paramount, requiring regular validation to prevent potential healthcare disparities and maintain trust in clinical settings.

From a data protection perspective, the implementation of these predictive models demands a comprehensive framework that addresses both privacy concerns and practical clinical integration. This includes robust security protocols for data transmission and storage, strict adherence to GDPR and HIPAA requirements, and clear procedures for data retention and deletion. The successful integration of these models in clinical practice requires maintaining a balance between algorithmic predictions and human oversight, ensuring that AI augments rather than replaces clinical judgment. Additionally, healthcare providers must be equipped with proper training and support systems to effectively communicate AI-derived predictions while maintaining transparency about the model's capabilities and limitations. This approach ensures that technology serves to enhance patient care while protecting individual privacy and maintaining ethical standards in medical practice.

Limitations

While the results of this study are promising, several limitations must be acknowledged. First, the study cohort was derived from two datasets in Madrid, Spain, which may limit the generalizability of the findings to other populations. Validation in diverse cohorts and geographic regions is essential to ensure broader applicability. To address the inherent geographical limitations of our current dataset, we are planning a comprehensive validation strategy to address this concern that will involve two parallel approaches: first, external retrospective validation using established international longitudinal databases (such as the National Alzheimer's Coordinating Center (NACC) and the Survey of Health, Ageing and Retirement in Europe (SHARE)) that contain equivalent variables to those in our study; and second, clinical prospective validation using patient data from multiple Spanish medical centers. This dual validation approach will enable both model refinement through new data incorporation and verification of generalizability across diverse populations. This systematic validation process will strengthen the robustness and broader applicability of our predictive models across different geographical and demographic contexts.

Second, while our models demonstrate strong predictive performance within a three-year window, we acknowledge that this follow-up period may not fully capture the progression of cognitive decline, particularly in individuals with slower disease trajectories. Longer follow-up periods could provide valuable insights into different progression patterns and potentially improve the calibration and predictive performance of our models. This limitation highlights the importance of future longitudinal studies with extended follow-up periods to validate and refine these predictive tools.

c) Model 3: demographic + self-reported + cognitive screening
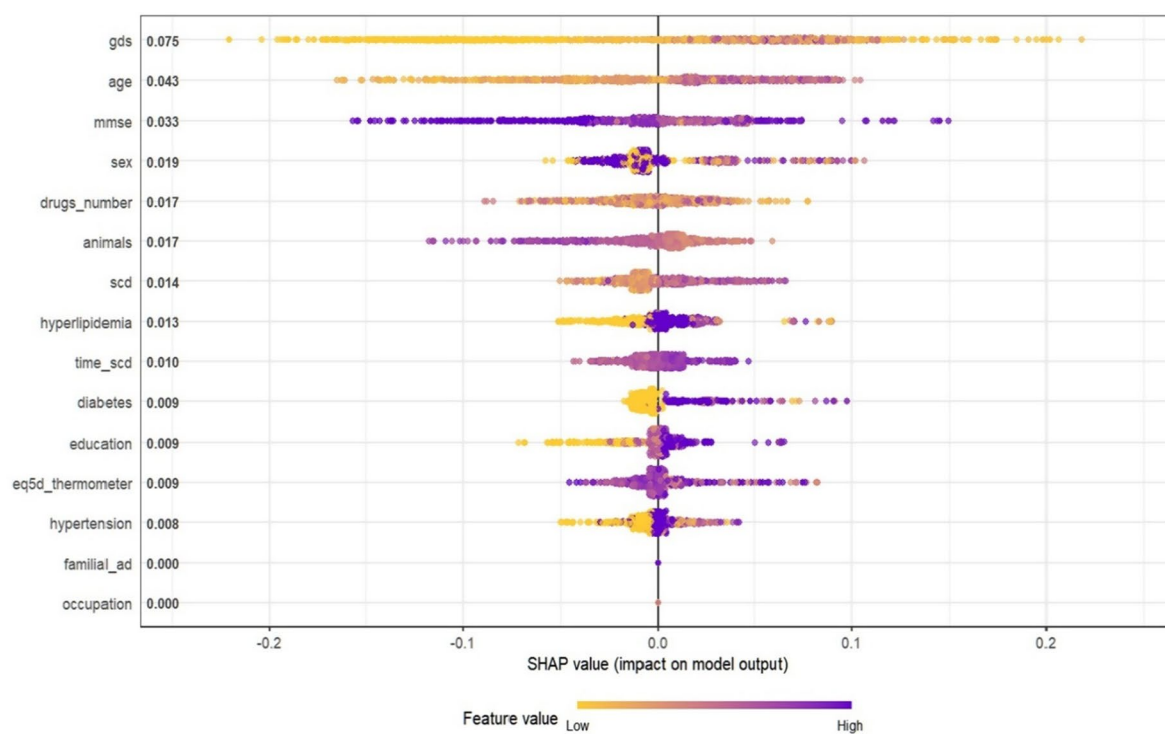
d) Model 4: demographic + self-reported + medical + cognitive screening
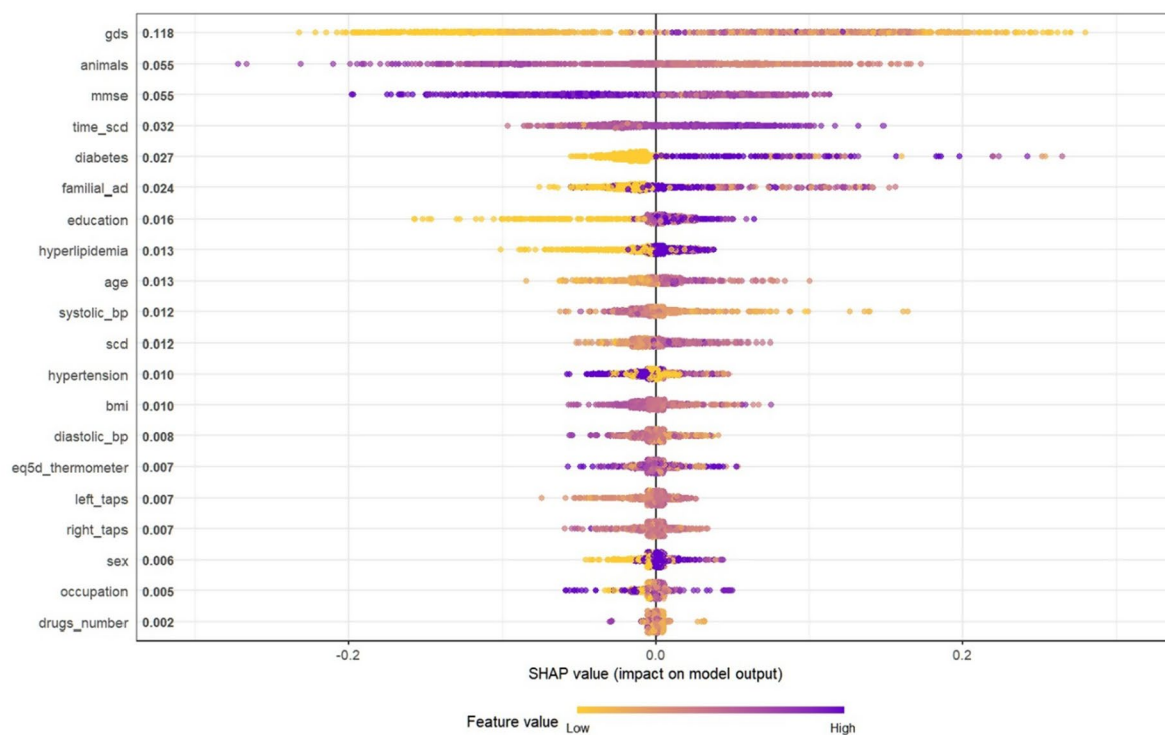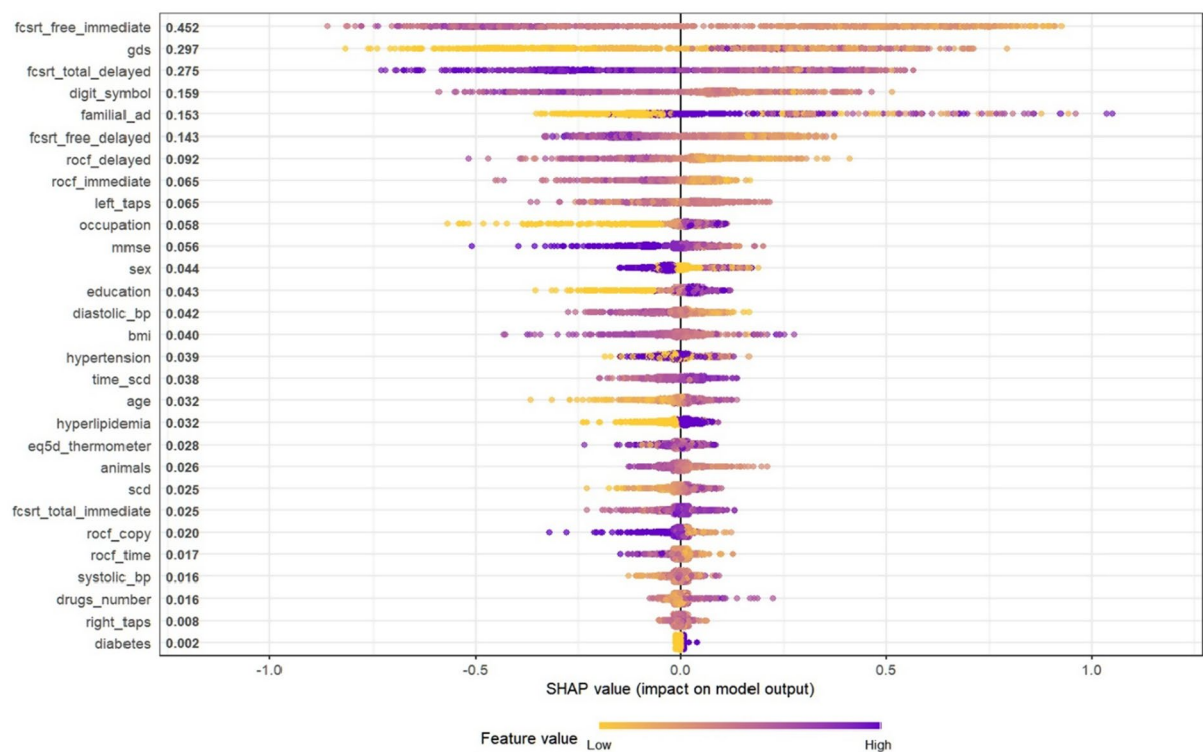

Fig. 3. (continued)

Fig. 3. (continued)

Finally, while our models incorporating cognitive assessments demonstrate superior predictive performance, we acknowledge the potential implementation challenges these assessments may present in various clinical settings. To address this limitation, we have developed a flexible, multi-tiered approach that accommodates different levels of assessment capability. For settings where comprehensive cognitive testing may not be feasible such as primary care, our simpler models (Models 1 and 2) provide reliable predictive performance using readily available clinical and demographic data. In any event, several emerging alternatives could complement or enhance traditional cognitive assessments in future iterations of these models. Digital physiological markers from wearable devices have shown promising correlations with cognitive function and offer the advantage of continuous, passive monitoring without requiring specialized testing environments. The incorporation of cognitive reserve proxies, such as educational attainment and occupational history, could also provide valuable predictive information without direct cognitive testing.

Broader implications and future research

The comparative analysis across our five models reveals two key insights. First, cognitive features were found to be the dominant predictors, particularly those assessing memory, which emphasizes their critical role in predicting MCI conversion. Second, there is also an interplay of demographic, self-reported, and medical factors, particularly in simpler models. This interplay suggests that while cognitive decline is central to MCI, other factors such as age, depression, cognitive complaints, family history of AD, and cardiovascular health significantly contribute to risk stratification.

According to Livingston et al. (2024), 45% of future dementia cases could potentially be prevented if the 14 risk factors identified by the Lancet Commission on Dementia Prevention, Intervention, and Care were eliminated[8]. This emphasizes the need to integrate machine learning frameworks, such as the one developed in this study, into public health initiatives aimed at reducing modifiable risk factors. Future research should explore how predictive models can be combined with personalized interventions to mitigate these risks at an individual level. Furthermore, the development of user-friendly tools, such as the online calculator proposed in this work, will facilitate the practical application of advanced algorithms, empowering healthcare professionals and policymakers to make data-driven decisions. Expanding the dataset to include diverse populations and longitudinal outcomes beyond three years could enhance the generalizability and robustness of predictive models. Extended follow-up studies could enable the development of more sophisticated time-dependent models capable of predicting cognitive decline across various timeframes. Such models could potentially be calibrated to provide predictions for different temporal horizons (e.g., 3-year, 5-year, and 10-year risk), offering clinicians more comprehensive prognostic information. Additionally, longer follow-up periods would allow for

Categories	Features	Included in model#
Demographics	Age	1, 2, 3, 4 & 5
	Gender	
	Education	
	Occupation	
Self-reported	SCD	1, 2, 3, 4 & 5
	Age at SCD onset	
	GDS-15	
	Equation 5D thermometer	
	Hypertension	
	Cholesterol	
	Diabetes	
	Number of drugs	
	Family history of dementia	
Medical	Body mass index	2, 4 & 5
	Tapping right hand	
	Tapping left hand	
	Systolic blood pressure	
	Diastolic blood pressure	
Cognitive screening	MMSE	3, 4 & 5
	Animals	
Further cognitive assessment	FCSRT free imm.	5
	FCSRT total imm.	
	FCSRT free del.	
	FCSRT total del.	
	ROCF time	
	ROCF copy	
	ROCF imm.	
	ROCF del.	
	Digit-Symbol	

Table 2. List of features used to build the five predictive models grouped by their typology. SCD: Subjective Cognitive Decline; EQ-5D: EuroQol 5D; GDS: Geriatric Depression Scale; MMSE: Mini Mental State Examination; FCSRT: Free and Cued Selective Reminding Test; ROCF: Rey-Osterrieth Complex Figure.

	Model 1	Model 2	Model 3	Model 4	Model 5
max_depth	10	10	9	9	8
Eta	0.103178	0.182763	0.048609	0.039159	0.09991
Gamma	0.167760	0.009576	0.159068	0.2	0.087543
colsample_bytree	0.702655	0.6	0.668692	0.893329	0.757451
min_child_weight	1	1	2	1	2
subsample	0.727134	0.677331	0.849401	0.6	0.978085

Table 3. Best combination of hyperparameters by each model.

the development of dynamic prediction models that could be updated as new patient data becomes available, potentially improving the accuracy and clinical utility of the predictions over time. Finally, interdisciplinary collaboration between clinicians, researchers, and technology developers is essential to ensure that these innovations translate effectively into real-world applications, fostering a global effort to reduce the incidence of dementia.

Conclusion

The early diagnosis of dementia, particularly AD, is critical for effective intervention and management. Traditional diagnostic methods, while effective, have several limitations that make them less feasible for widespread use. ML offers a promising alternative, with the potential to develop accurate, interpretable, and online accessible diagnostic tools using non-imaging features. This study aims to address the gaps in current research by developing and validating a non-imaging-based diagnostic tool for dementia that represents a

	Model 1	Model 2	Model 3	Model 4	Model 5
Accuracy (95% CI)	0.8225 (0.7564–0.8769)	0.8230 (0.7572–0.8783)	0.8235 (0.7601–0.8801)	0.8284 (0.7629–0.8820)	0.8639 (0.8028–0.9117)
Recall (95% CI)	0.9310 (0.8881–0.9718)	0.9241 (0.8784–0.9645)	0.8966 (0.8461–0.9424)	0.8684 (0.8117–0.9211)	0.9013 (0.8543–0.9481)
Precision (95% CI)	0.8710 (0.8170–0.9195)	0.8758 (0.8228–0.9241)	0.8966 (0.8446–0.9433)	0.9362 (0.8905–0.9718)	0.9448 (0.9048–0.9796)
F1-score (95% CI)	0.8999 (0.8639–0.9355)	0.8990 (0.8592–0.9320)	0.8966 (0.8550–0.9307)	0.8993 (0.8622–0.9342)	0.9222 (0.8866–0.9537)
AUC (95% CI)	0.6687 (0.5584–0.7789)	0.7293 (0.6151–0.8435)	0.7682 (0.6826–0.8539)	0.7744 (0.6717–0.8771)	0.8359 (0.7464–0.9254)

Table 4. Predictive performance of the five models for MCI prediction.

significant step toward early detection and personalized prevention of cognitive decline. With further validation and refinement, this tool could play a pivotal role in mitigating the growing global burden of dementia.

Data availability
The datasets generated during the current study are not publicly available due to privacy restrictions but are available from the corresponding author on reasonable request.

References
1. Affairs, U. N. D. of E. and S. *World Population Ageing 2019*. United Nations, Department of Economic and Social Affairs, Population Division (2020). https://doi.org/10.18356/6A8968EF-EN
2. Better, M. A. Alzheimer's disease facts and figures. *Alzheimers Dement* **19**, 1598–1695 (2023).
3. World Health Organization. *Global Action Plan on the Public Health Response To Dementia 2017–2025. Geneva: World Health Organization* (World Health Organization, 2017).
4. Hampel, H., Lista, S. & Khachaturian, Z. S. Development of biomarkers to chart all Alzheimer's disease stages: The Royal road to cutting the therapeutic gordian knot. *Alzheimers Dement.* **8**, 312–336 (2012).
5. Porsteinsson, A. P., Isaacson, R. S., Knox, S., Sabbagh, M. N. & Rubino, I. Diagnosis of early Alzheimer's disease: Clinical practice in 2021. *J. Prev. Alzheimers Dis.* **8**, 371–386 (2021).
6. Sanz-Blasco, R. et al. Transition from mild cognitive impairment to normal cognition: Determining the predictors of reversion with multi-state Markov models. *Alzheimers Dement.* **18**, 1177–1185 (2022).
7. Livingston, G. et al. Dementia prevention, intervention, and care. *Lancet* **390**, 2673–2734 (2017).
8. Livingston, G. et al. Dementia prevention, intervention, and care: 2024 report of the lancet standing commission. *Lancet* **404**, 572–628 (2024).
9. Huang, J. et al. Haemoglobin A(1c) and cognitive function in very old, cognitively intact men. *Age Ageing.* **41**, 125–128 (2012).
10. West, R. et al. Better memory functioning associated with higher total and low-density lipoprotein cholesterol levels in very elderly subjects without the Apolipoprotein e4 allele. *Am. J. Geriatr. Psychiatry.* **16**, 781–785 (2008).
11. West, R. K. et al. Homocysteine and cognitive function in very elderly nondemented subjects. *Am. J. Geriatr. Psychiatry.* **19**, 673–677 (2011).
12. Gómez-Ramírez, J., Ávila-Villanueva, M. & Fernández-Blázquez, M. Á. Selecting the most important self-assessed features for predicting conversion to mild cognitive impairment with random forest and permutation-based methods. *Sci. Rep.* 10, (2020).
13. Wang, H. et al. Develop a diagnostic tool for dementia using machine learning and non-imaging features. *Front. Aging Neurosci.* **14** (2022).
14. Stephan, B. C. M., Kurth, T., Matthews, F. E., Brayne, C. & Dufouil, C. Dementia risk prediction in the population: Are screening models accurate? *Nat. Rev. Neurol.* **6**, 318–326 (2010).
15. Barnes, D. E. et al. Predicting risk of dementia in older adults: The late-life dementia risk index. *Neurology* **73**, 173–179 (2009).
16. Exalto, L. G. et al. Risk score for prediction of 10 year dementia risk in individuals with type 2 diabetes: A cohort study. *Lancet Diabetes Endocrinol.* **1**, 183–190 (2013).
17. Nori, V. S., Hane, C. A., Martin, D. C., Kravetz, A. D. & Sanghavi, D. M. Identifying incident dementia by applying machine learning to a very large administrative claims dataset. *PLoS One* **14** (2019).
18. Diogo, V. S., Ferreira, H. A. & Prata, D. Early diagnosis of Alzheimer's disease using machine learning: A multi-diagnostic, generalizable approach. *Alzheimers Res. Ther* **14** (2022).
19. Fernández-Blázquez, M. A. et al. MADRID + 90 study on factors associated with longevity: Study design and preliminary data. *PLoS One* **16** (2021).
20. 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.* **7**, 270–279 (2011).
21. Chen, T. & Guestrin, C. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Dis* 13–17-August-2016, 785–794 (2016).
22. Snoek, J., Larochelle, H. & Adams, R. P. Practical bayesian optimization of machine learning algorithms. *Adv. Neural Inf. Process. Syst.* **4**, 2951–2959 (2012).
23. Kuhn, M. Building predictive models in R using the caret package. *J. Stat. Softw.* **28**, 1–26 (2008).
24. Robin, X. et al. pROC: An open-source package for R and S + to analyze and compare ROC curves. *BMC Bioinform.* **12** (2011).
25. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis* (Springer, 2016).
26. van Buuren, S. & Groothuis-Oudshoorn, K. Mice: Multivariate imputation by chained equations in R. *J. Stat. Softw.* **45**, 1–67 (2011).
27. Lunardon, N., Menardi, G. & Torelli, N. ROSE: A package for binary imbalanced learning. *R J.* **6**, 79–89 (2014).
28. Bentéjac, C., Csörgő, A. & Martínez-Muñoz, G. A comparative analysis of gradient boosting algorithms. *Artif. Intell. Rev.* **54**, 1937–1967 (2021).

29. Lundberg, S. M. & Lee, S. I. A unified approach to interpreting model predictions. *Adv. Neural Inf. Process. Syst.* 4766–4775 (2017).
30. Yang, L. & Just, A. *SHAPforxgboost: SHAP Plots for 'XGBoost'* R package version 0.1.3 Preprint at (2023). https://CRAN.R-project.org/package=SHAPforxgboost
31. Fernández-Blázquez, M. A., Ávila-Villanueva, M. & Medina, M. The dimensional structure of subjective cognitive decline. *Neuromethods* **137**, 45–62 (2018).
32. Yesavage, J. et al. Development and validation of a geriatric depression screening scale: A preliminary report. *J. Psychiatr Res.* **17**, 37–49 (1983).
33. Rabin, R. & De Charro, F. EQ-5D: A measure of health status from the EuroQol group. *Ann. Med.* **33**, 337–343 (2001).
34. Folstein, M. F., Folstein, S. E. & McHugh, P. R. Mini-mental state: A practical method for grading the cognitive state of patients for the clinician. *J. Psychiatr Res.* **12**, 189–198 (1975).
35. Buschke, H. Cued recall in amnesia. *J. Clin. Neuropsychol.* **6**, 433–440 (1984).
36. Rey, A. L'examen psychologique Dans les Cas d'encéphalopathie traumatique. *Arch. Psychol. (Geneve).* **28**, 215–285 (1941).
37. Wechsler, D. *Wechsler Adult Intelligence Scale-III* (The Psychological Corporation, 1997).
38. Belleville, S., Fouquet, C., Hudon, C., Zomahoun, H. T. V. & Croteau, J. Neuropsychological measures that predict progression from mild cognitive impairment to Alzheimer's type dementia in older adults: A systematic review and meta-analysis. *Neuropsychol. Rev.* **27**, 328–353 (2017).
39. Gomar, J. J., Bobes-Bascaran, M. T., Conejero-Goldberg, C., Davies, P. & Goldberg, T. E. Utility of combinations of biomarkers, cognitive markers, and risk factors to predict conversion from mild cognitive impairment to Alzheimer disease in patients in the Alzheimer's disease neuroimaging initiative. *Arch. Gen. Psychiatry.* **68**, 961–969 (2011).

Author contributions

M.A.F-B. and J.M.R-SdL. conceptualized and designed the study. M.J.G-M., M.M-P., C.T., and C.F-G. were responsible for data collection. M.A-V., E.V., and R.S-B. managed the database. M.A.F-B. and J.G-R. performed the data analysis, including the application of machine learning techniques and the development of the risk calculator. J.M.R-SdL., E.V., and R.S-B. contributed to the interpretation of the results and provided critical revisions of the manuscript. M.A-V. and J.G-R. participated in drafting and final editing of the manuscript. M.A.F-B. supervised the overall project. All authors read and approved the final version of the manuscript.

Funding

Declarations

Competing interests
The authors declare no competing interests.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-14832-0.

Correspondence and requests for materials should be addressed to M.A.F.-B.

Reprints and permissions information is available at www.nature.com/reprints.

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