ACCSCIENCE PUBLISHING

Design+

ARTICLE

Machine learning strategies for predicting Alzheimer's disease progression

Adhinrag Kalarikkal Induchudan and Kevin Curran*

Faculty of Computing, Engineering and The Built Environment, School of Computing, Engineering, and Intelligent Systems, Ulster University, Londonderry, Ireland

Abstract

Alzheimer's disease (AD) represents a significant global health challenge, affecting millions of individuals worldwide through progressive cognitive decline and behavioral changes. The burden extends beyond patients to caregivers and healthcare systems. While traditional diagnostic methods pose financial obstacles, emerging non-imaging techniques show promise. Machine learning has emerged as a transformative approach for enhancing both diagnosis and management. This study aims to develop a robust multi-class classification model using random forest (RF) and extreme gradient boosting algorithms on non-imaging data from the Australian AD Neuroimaging Initiative, with emphasis on the Australian Imaging, Biomarkers, and Lifestyle Study of Aging. Extensive data analysis was conducted, including feature importance and selection, to improve interpretability and classification accuracy. Synthetic oversampling was applied to address class imbalance. The findings indicate the superiority of the tuned RF model, achieving 90% in accuracy, precision, recall, and F1 scores. In addition, cost-effective diagnostic variables were explored, with neuropsychology assessment variables demonstrating exceptional accuracy (90%). This research contributes to early AD detection, personalized treatment, and optimized resource allocation.

Keywords: Alzheimer's disease; Machine learning; Python classification model; Non-imaging data; Random Forest; Extreme gradient boosting; Australian imaging biomarkers and lifestyle study of aging; Diagnosis

***Corresponding author:**
Kevin Curran
(kj.curran@ulster.ac.uk)

Citation: Induchudan AK, Curran K. Machine learning strategies for predicting Alzheimer's disease progression. *Design+*. 2025;2(3):025270031. doi: 10.36922/DP025270031

Received: July 3, 2025

Revised: July 23, 2025

Accepted: August 1, 2025

Published online: August 21, 2025

1. Introduction

Alzheimer's disease (AD) represents a significant global health challenge, affecting millions of individuals worldwide. This condition progressively impairs memory, cognitive function, and behavior, ultimately leading to severe disability and death. AD not only affects those diagnosed but also places considerable strain on caregivers and healthcare systems, escalating the burden of care and resource allocation. Initially, AD may manifest as mild forgetfulness, but it gradually progresses to encompass a wide range of symptoms that deteriorate over time, subjecting both patients and their families to a distressing trajectory of decline and loss. The emotional toll of AD extends beyond cognitive impairment, significantly affecting the well-being of families and caregivers.[1] The continuous demands of caregiving challenge emotional resilience and endurance. However, amidst these challenges, there is a shared commitment to confronting AD with resolve and innovation.

One significant obstacle in addressing AD is the high cost associated with traditional imaging techniques and diagnostic procedures. While these methods are beneficial, they are often highly expensive for patients and healthcare systems. Nevertheless, emerging alternatives, such as genetic markers, neuropsychological assessments, and biomarker analysis, show promise as more accessible and cost-effective diagnostic tools.[2] By prioritizing these non-imaging methods, the financial burden of diagnosis may be alleviated, thereby broadening access to care for individuals with AD.

In this landscape of challenges and opportunities, machine learning (ML) has emerged as a transformative tool. With its ability to process complex datasets and extract valuable insights, ML holds the potential to improve AD diagnosis and management. Through the utilization of novel data and rigorous training, ML algorithms excel at predicting outcomes and providing invaluable guidance for decision-making processes. Moreover, ML enables earlier disease detection and intervention, thereby contributing to improved patient outcomes and enhanced quality of life.[3] The adaptability of ML models further allows for continual refinement and optimization, ensuring ongoing improvements in prediction accuracy and diagnostic efficacy.

The primary objective of this study is to develop a robust multi-class classification model for predicting AD among three distinct groups: Healthy control (HC), individuals with mild cognitive impairment (MCI), and those diagnosed with AD. Leveraging non-imaging data from the Australian AD Neuroimaging Initiative,[4] with a particular emphasis on the Australian Imaging Biomarkers and Lifestyle Study of Aging (AIBL),[5] this study utilizes random forest (RF) and Extreme Gradient Boosting (XGBoost) algorithms, along with their optimized models. Through comparative analysis, the most effective classification model is identified. In addition, this study aims to enhance interpretability through feature importance analysis and the evaluation of various classifiers. These efforts are expected to streamline the predictive process for AD, facilitate early detection, enable personalized treatment strategies, and optimize resource allocation. The ultimate goal is to provide valuable insights to inform the development of improved, cost-effective diagnostic and therapeutic approaches for addressing this debilitating condition.

2. Existing work

Many researchers have conducted studies on classifying AD using various datasets. In alignment with the present study's objectives, Rahman and Prasad[6] addressed the challenge of accurately diagnosing AD—a disease that severely impacts cognitive and behavioral abilities—as a binary classification problem. Utilizing non-imaging data from the AIBL, they built RF models employing different combinations of data and preprocessing steps. An RF is an ML algorithm that uses an ensemble of decision trees to make predictions. It is a supervised learning method, trained on labeled data to classify or predict outcomes. RFs are known for their accuracy and ability to handle complex datasets.

Their approach included using scaled and unscaled data for simple RF classifiers, tuned RF classifiers, and RF classifiers with selected features using DALEX and Boruta packages in R software. Their results showed that the tuned RF classifier, which utilized the original data, achieved an impressive 96% accuracy in classifying AD into HC and non-HC categories, with precision and recall scores exceeding 97%. Model evaluation was primarily focused on accuracy, in line with their research objective of effectively classifying instances of AD. Furthermore, they developed multiple diagnostic classifiers and evaluated them to streamline the prediction process, aiming to create a cost-effective diagnosis method.

Notably, their classifier based on neuropsychological assessment variables demonstrated exceptional performance, achieving an accuracy of 93.68%. This model required only 4 out of 30 test variables, highlighting its potential to increase efficiency in diagnostic processes.

3. Dataset description

The AIBL study commenced in 2006 with the aim of investigating the origins of AD and developing tools for identifying cognitive decline at its early stages.[4,5] The study includes a diverse population comprising healthy individuals, those with MCI, and those diagnosed with AD. With over 1,000 participants, the AIBL dataset represents a comprehensive resource for AD research. It supports investigations into the associations between lifestyle factors and cognitive impairment and facilitates the development and evaluation of algorithms for early AD detection. A summary of the dataset is presented in Table 1.

4. Methodology

The Cross-Industry Process for Data Mining (CRISP-DM), a widely adopted methodology recognized for its effectiveness across industries, was employed in this study. It offers flexibility while maintaining a comprehensive and structured approach compared to other methods.[7] The method comprises distinct phases: business understanding, data understanding, data preparation,

Table 1. Dataset summary

Variable	Description
Demographics	
Age	55–96 years old
Gender	Categorized as "Female" or "Male"
Medical history	
Psychiatric (MH_PSYCH)	Binary features
Neurologic (MH_NEURL)	
Cardiovascular (MH_CARD)	
Hepatic (MH_HEPAT)	
Musculoskeletal (MH_MUSCL)	
Endocrine–metabolic (MH_ENDO)	
Gastrointestinal (MH_GAST)	
Renal–genitourinary (MH_RENA)	
Smoking (MH_SMOK)	
Malignancy (MH_MALI)	
ApoE genotype	
Two-allele genotype	Each individual carries two ApoE alleles, and each allele can be E2, E3, or E4
Neuropsychological assessments	
Clinical dementia rating (CDGLOBAL)	Total number of story units recalled immediately; scores ranged from 0 to 25
Mini-mental state exam (MMSCORE)	Total number of story units recalled after a delay; scores ranged from 0 to 25
Logical memory immediate recall (LIMMTOTAL)	-
Logical memory delayed recall (LDELTOTAL)	
Blood analysis	
Thyroid stimulating hormone (AXT117)	
Vitamin B12 (BAT126)	
Red blood cell count (HMT3)	
White blood cell count (HMT7)	
Platelet count (HMT13)	
Hemoglobin (HMT40)	
Mean corpuscular hemoglobin (HMT100)	
Mean corpuscular hemoglobin concentration (HMT102)	
Urea nitrogen (RCT6)	
Serum glucose (RCT11)	
Cholesterol (high performance; RCT120)	
Creatinine (rate blanked; RCT329)	
Diagnosis	
Diagnostic results	Categorized into healthy control, mild cognitive impairment, and Alzheimer's disease

Abbreviation: ApoE: Apolipoprotein E.

modeling, evaluation, and deployment. Figure 1 illustrates a graphic representation of these CRISP-DM phases.

4.1. Business understanding

The business understanding phase involves defining business objectives, assessing the current context, establishing data mining goals, and formulating a project plan. As outlined in the introduction, a background study was conducted, and the research objectives were clearly defined. The success criteria for this study involved benchmarking classifier performance against the AD classification model presented by Rahman and Prasad[6] and comparing the best diagnosis classifier with the one identified in their study.

This comparison focused on four key metrics critical for evaluating classifier performance: (i) Accuracy, indicating the proportion of correctly predicted instances relative to the total number of instances in the dataset; (ii) precision, a measure of prediction reliability, reflecting the ratio of true positive predictions to all positive predictions; (iii) recall, also referred to as sensitivity, measuring the classifier's ability to identify actual positive cases; and (iv) F1-score, the harmonic mean of recall and precision, which balances the trade-off between these two metrics.[8]

A comprehensive project plan was formulated based on available resources, requirements, and risk assessments. The plan encompassed tasks across each CRISP-DM phase, including the selection of appropriate tools, methodologies, and risk mitigation strategies. The primary tools utilized were Google Colab and Python, with tasks involving data preparation, cleaning, and analysis. Python libraries, particularly functionalities

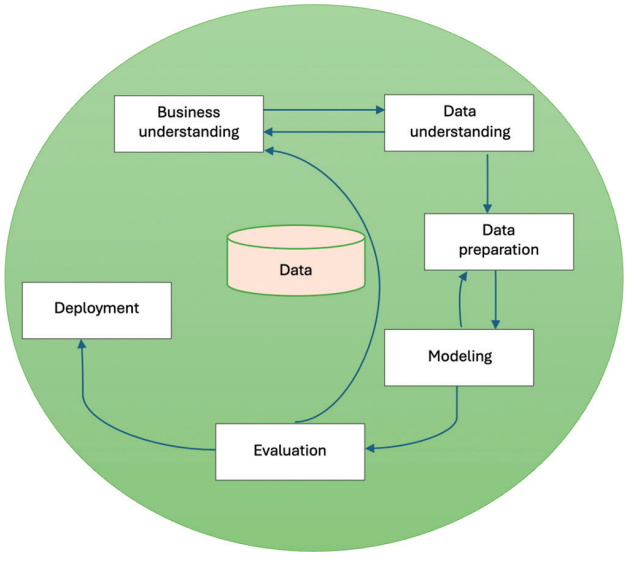

Figure 1. Phases of the cross-industry process for data mining

from Scikit-learn,[9] were employed for modeling and evaluation purposes.

4.2. Data understanding

In the second phase of the methodology, we began by familiarizing ourselves with the collected data using a comprehensive data dictionary, which outlined feature descriptions and properties. The dataset comprised eight distinct CSV files, imported into Google Colab via file synchronization from Google Drive using the PyDrive library in Python.[10] Subsequently, these CSV files were merged to construct a master dataframe, facilitated by shared key columns such as RID, SITEID, and VISCODE, resulting in a unified dataframe containing 1,688 rows and 36 columns.

Given our focus on baseline data, we filtered the dataset for baseline entries using the VISCODE column, yielding 862 observations. To prepare for pre-analysis, we systematically transformed several features into categorical formats based on predefined values. Medical history variables—including MHPSYCH, MH2NEURL, MH4CARD, MH6HEPAT, MH8MUSCL, MH9ENDO, MH10GAST, MH12RENA, MH16SMOK, and MH17MALI—were categorized as "No" or "Yes" based on their respective binary values. Apolipoprotein E (ApoE) genotypes (e.g., APGEN1, APGEN2) were labeled as "E2," "E3," or "E4," corresponding to their genetic variants. MMSCORE was segmented into severity levels (e.g., "Severe," "Moderate," "Mild," "Normal") based on predefined score ranges. PTGENDER was categorized as "Male" or "Female" according to gender data. CDGLOBAL was classified into health status categories (e.g., "Healthy," "Very Mild," "Mild," "Moderate," "Severe") based on clinical assessment scores. DXCURREN was mapped to clinical stages (e.g., "HC," "MCI," "AD") using a predefined mapping dictionary. These transformations enhance the interpretability of the dataset by aligning feature values with clinically relevant categories for subsequent analysis. During this preparatory phase, it was observed that 2.28% of the data were missing; however, no duplicates were detected.

Before the exploratory data analysis, a few data cleaning procedures were performed to enhance the interpretability of the findings. This step was essential to ensure the accuracy and reliability of the analyses by removing any inconsistencies and inaccuracies within the dataset. Initially, the age of patients was calculated by comparing examination dates with their respective birthdates. This process involved cleansing the date of birth column to remove unnecessary characters, followed by the creation of the EXAMYEAR column to compute the age as a distinct feature. Furthermore, noisy values of "–4"—recurrent across multiple columns—were identified and replaced with NaN. Concurrently, redundant columns such as RID, SITEID, VISCODE, EXAMDATE, EXAMYEAR, APTESTDT, and PTDOB, among others, were eliminated to streamline the dataset for analysis.

In the exploratory data analysis, the distribution of output classes was visualized, revealing a significant class imbalance among HC, MCI, and AD. Specifically, HC emerged as the predominant class with 609 instances, followed by MCI with 144 instances and AD with 105 instances. A subsequent review of summary statistics for numerical features revealed slight discrepancies in feature counts, suggesting the presence of missing values. Moreover, notable differences in scales and variances were observed across many features.

Upon delving further into the distributions of numerical features, distinctive patterns were observed. Variables such as AXT117, BAT126, and HMT7, alongside RCT6 and RCT11, displayed a notable tendency toward higher values, suggesting a right-skewed distribution. Similarly, RCT392 exhibited a comparable pattern, indicating a concentration of data at the lower end with potential outliers extending toward higher values.

In contrast, the distributions of HMT13, HMT40, HMT100, HMT102, RCT20, RCT392, and AGE showed a unimodal pattern, indicative of relatively normal distributions with a pronounced peak at the center. This characteristic suggests the presence of a central value around which the data clusters. Furthermore, LIMMTOTAL displayed a unimodal distribution with an additional smaller peak, while LDELTOTAL exhibited a similar pattern with a slightly less distinct secondary peak.

The analysis was extended using box plots to assess the spread of numerical variables. Except for LIMMTOTAL, LDELTOTAL, and AGE, potential outliers were observed in the remaining variables at both ends of the distribution.

To assess multicollinearity,[11] a correlation matrix was constructed and visualized using a heatmap (Figure 2). LIMMTOTAL and LDELTOTAL exhibited a strong positive correlation, indicating a close relationship between these variables. Additionally, HMT3 and HMT40, HMT100 and HMT102, as well as RCT6 and RCT392, demonstrated strong positive correlations, further highlighting interdependencies within the dataset. Conversely, strong negative correlations were observed between HMT100 and HMT3, HMT40 and HMT3, as well as HMT13 and HMT3, suggesting inverse relationships between these variables.

Finally, the association between categorical variables and the target variables was evaluated. As shown in Figure 3,

CDGLOBAL and MMSCORE displayed significant Chi-square statistics[12] with extremely low *p*-values, indicating a robust association with the target variable. This suggests that these variables hold substantial predictive power with respect to the target outcome. In addition, MH2NEURL, APGEN1, and APGEN2 exhibited moderate chi-square statistics accompanied by small *p*-values, indicating a noticeable association with the target variable, although not as strong as that of CDGLOBAL and MMSCORE. However, MH8MUSCL and PTGENDER demonstrated relatively smaller Chi-square statistics along with higher *p*-values, suggesting a weaker association with the target variable.

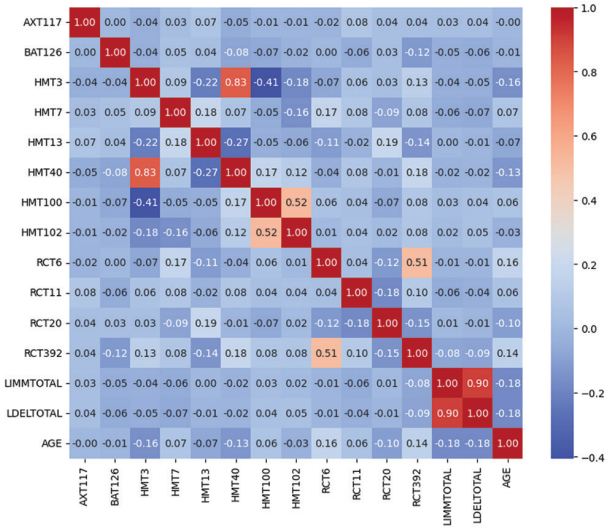

Figure 2. Correlation matrix heatmap of numerical features

Overall, the exploratory data analysis identified several areas for improvement, including class imbalance, missing values, outliers, multicollinearity, and skewed distributions.

4.3. Data preparation

Data preparation, the third phase of the CRISP-DM methodology, began by following the basic data cleaning steps conducted during the data understanding phase. The initial step involved converting categorical data into a numerical format to facilitate model development. However, it was noted that the "pd.factorize"[13] function assigned "−1" in place of missing values, necessitating further replacement with NaN values to enable imputation at a later stage.

To prevent data leakage and assess the model's efficacy in generalizing to previously unseen data, a critical first step was to split the data before implementing any preprocessing techniques.[14] The data was split in an 80:20 ratio, allocating 80% for training and the remaining 20% for testing. Subsequently, we focused on handling missing values within the training dataset, acknowledging the potential impact on predictive accuracy due to data loss if inadequately addressed. To address this, the MissForest imputation technique,[15] an algorithmic approach that initially uses mean and mode values to replace missing data, was applied. This was followed by the implementation of an RF methodology to iteratively predict missing values, prioritizing data accuracy over processing speed.

Given the high dimensionality of the dataset, an analysis of feature importance was conducted to determine the most

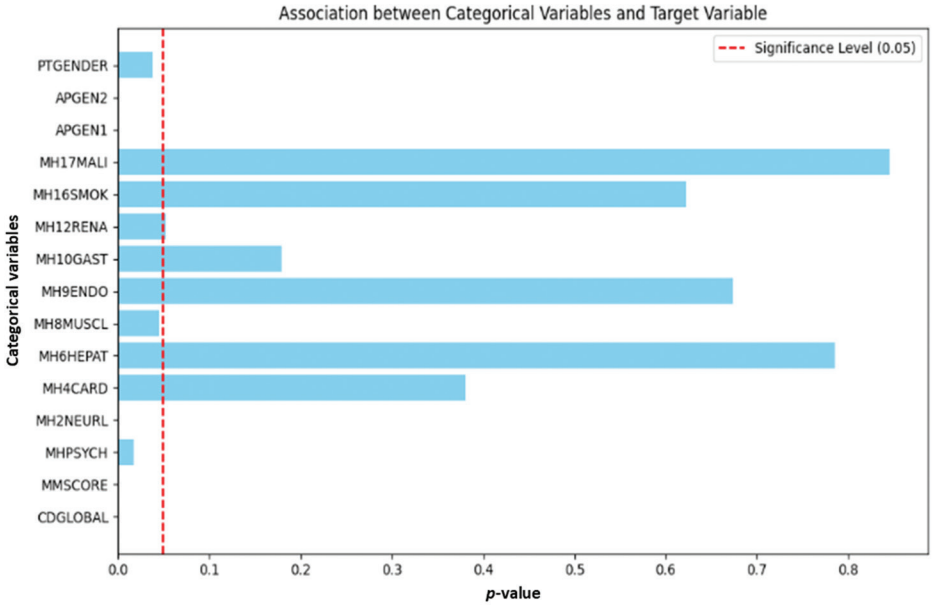

Figure 3. Visualization of the results from the Chi-square test

influential features for accurate predictions. By combining the permutation feature importance technique[16] with an RF classifier over 150 iterations, this analysis revealed the significance of specific features in influencing predictive accuracy, thereby guiding further modeling decisions. The importance of each feature was systematically assessed to ensure a comprehensive understanding of its contribution to the overall predictive capability of the model.

Subsequently, feature selection was performed to streamline computational resources and optimize model performance. Using an RF feature selection technique[17] with 100 estimators and a maximum depth of 5, the algorithm evaluated each feature's contribution to impurity reduction (Gini impurity) before decision tree construction, thereby identifying the most significant features for predictive modeling. By selecting the most informative, non-redundant features, data utilization was optimized, resulting in improved computational efficiency and enhanced model performance.

Addressing class imbalance, a common challenge in ML, was essential to ensuring model robustness across all classes. The Synthetic Minority Over-Sampling Technique (SMOTE)[18] was applied to oversample minority classes (e.g., MCI and AD) by generating synthetic samples, yielding a balanced representation across all classes. Following resampling, further adjustments were made to facilitate model training and evaluation, resulting in the creation of two distinct dataframes for analysis.

4.4. Modeling

In selecting ML models during the data preparation phase, non-parametric models were prioritized due to their flexibility in handling complex datasets.[19] Outliers, multicollinearity, and skewness were identified as key challenges that were unaddressed in the previous phase. Therefore, tree-based models were considered suitable due to their adaptability to such issues. For multiclass classification, the RF and XGBoost algorithms were selected.[20,21]

RF[21] is an ensemble learning algorithm that combines multiple decision trees to yield more accurate and reliable predictions. By training each decision tree on randomly selected subsets of the training data, RF reduces overfitting and enhances model generalizability.

XGBoost, commonly known as XGBoost,[22] is another powerful algorithm in the gradient boosting family. XGBoost is a widely used open-source software library that implements a gradient boosting algorithm. It is commonly applied to ML tasks such as classification, regression, and ranking, particularly when dealing with tabular or structured data. XGBoost is known for its speed, efficiency, and ability to handle large datasets. It sequentially builds a strong predictive model by aggregating the predictions of multiple weak decision trees. Through advanced feature selection and regularization techniques, XGBoost minimizes overfitting and improves model performance.

Two models were developed for the prepared data: baseline models and their fine-tuned equivalents. For fine-tuning, the "RandomizedSearchCV" function was used.[23] This method selects random combinations of hyperparameter values from a grid, trains the model on a subset of the training data, and evaluates its performance on a different subset using cross-validation. The combination that yields the best performance metric represents the optimized set of hyperparameters for the model.

In addition, three distinct diagnosis classifiers were developed to identify the most reliable method for reducing the number of tests required for disease detection, thereby lowering diagnostic costs. These classifiers utilize medical history variables, blood analysis, ApoE genotype variables, and neuropsychological assessment variables individually. To model these classifiers, we employed the fine-tuned version of the best-performing algorithm, ensuring optimal predictive performance. This approach aims to streamline the diagnostic process while maintaining diagnostic accuracy.

4.5. Model evaluation

In this phase, a comprehensive evaluation of the models was conducted to guide future actions. Predictions from all models were compared against actual values using the "classification_report" function.[24] The evaluation included accuracy, as well as weighted-average precision, recall, and F1-score, offering a detailed overview of overall model performance. This approach accounts for class imbalances, ensuring robustness across all classes.[25] In addition, macro-average and class-wise performance metrics were emphasized when further insights were required. Given the research focus of this study, the deployment phase was omitted. A detailed analysis of the models is presented in the Results and Discussion section.

5. Results

5.1. Feature importance

The Chi-square test results revealed a significant association between the target variable and two key features, CDGLOBAL and MMSCORE, as indicated by their strong chi-square statistics and extremely low p-values. This finding was further confirmed by the permutation feature importance test. Figure 4 shows the feature importance ranking, highlighting CDGLOBAL as the most influential feature, followed by LDELTOTAL, MMSCORE, and

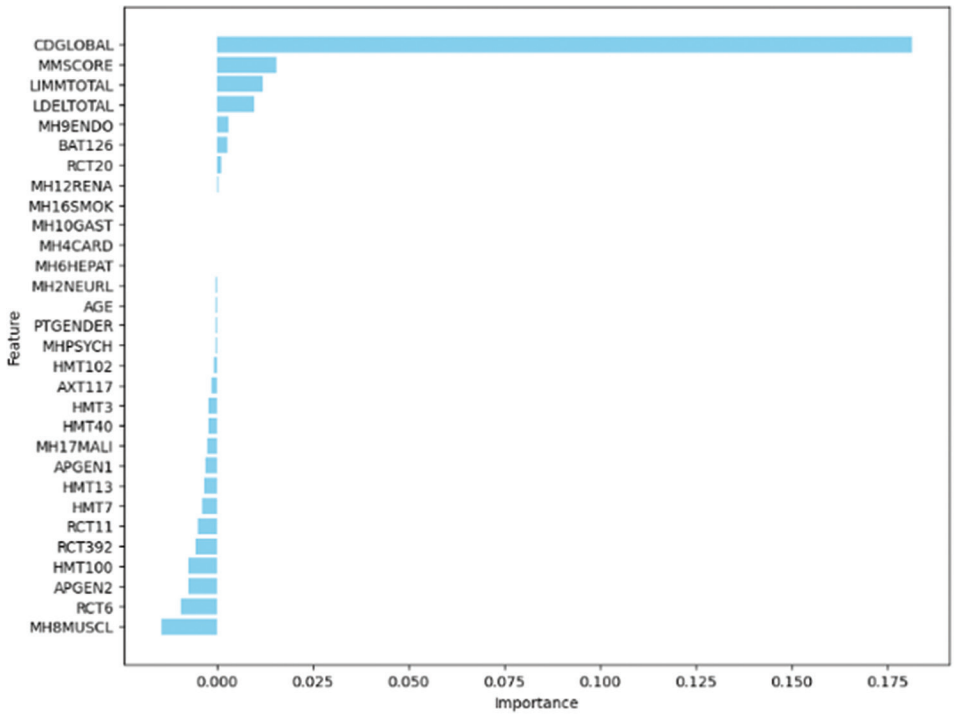

Figure 4. Feature importance using permutation method

LIMMTOTAL. While MH9ENDO ranked next in importance, its contribution during the imputation process was relatively less significant compared to the other features. This comprehensive analysis underscores the pivotal role of these features in predicting the target outcome, thereby guiding subsequent steps in the analysis.

5.2. Feature selection

The RF-based feature selection process identified four key features as crucial for predicting the output class, DXCURREN: CDGLOBAL, MMSCORE, LIMMTOTAL, and LDELTOTAL. These features demonstrated significant importance in accurately predicting the target outcome. Additionally, the feature importance analysis revealed MH9ENDO as an additional feature, though its contribution was relatively minor compared to the others. Implementing this feature selection approach supported decision-making by excluding MH9ENDO from the final feature set.

5.3. Class balancing

In the dataset exhibiting class imbalance, the sample distribution was skewed, with 609 samples for HC, 144 for MCI, and 105 for AD. By generating additional synthetic samples using SMOTE, each class was balanced to contain 490 samples.

5.4. Performance evaluation

As shown in Table 2, the performance evaluation metrics indicate that the tuned RF model with selected features outperformed the other models. The best hyperparameters included "n_estimators" = 100, "min_samples_split" = 15, "min_samples_leaf" = 1, and "max_depth" = 50—identified through randomized search with five-fold cross-validation over 100 iterations. Furthermore, the tuned RF model with selected features demonstrated exceptional performance across multiple evaluation metrics.

In Class 0 (HC), the model exhibited high precision (97%) and recall (93%), ensuring accurate identification of HCs. For Class 1 (MCI), while precision was moderate (69%), the model displayed commendable recall (89%), which is crucial for identifying MCI instances. Class 2 (AD) demonstrated balanced precision (95%) and recall (78%), essential for accurately identifying AD cases.

The overall accuracy of 90% underscores the model's efficiency, with both macro-average (precision: 0.87; recall: 0.87; F1-score: 0.86) and weighted-average metrics (precision: 0.91; recall: 0.90; F1-score: 0.90) confirming its consistency and superior performance across all classes. This comprehensive evaluation highlights the effectiveness of the tuned RF model in accurately distinguishing between different diagnostic categories. The Appendix

Table 2. Performance metrics of the machine learning models

Machine learning model	Complete features					Selected features				
	Accuracy	Weighted average				Accuracy	Weighted average			
		Precision	Recall	F1-score	Support		Precision	Recall	F1-score	Support
Simple RF	0.88	0.90	0.88	0.88	74	0.86	0.86	0.86	0.86	134
Tuned RF	0.88	0.90	0.88	0.88	74	0.90[a]	0.91[a]	0.90[a]	0.90[a]	134[a]
Simple XGBoost	0.86	0.87	0.86	0.87	74	0.85	0.85	0.85	0.85	134
Tuned XGBoost	0.85	0.86	0.85	0.85	74	0.89	0.90	0.89	0.89	134

Notes: This table presents the performance of machine learning models evaluated on two datasets—one with complete features and one with selected features. "Tuned" models refer to those that were optimized via hyperparameter tuning using "RandomizedSearchCV" function. Metrics include accuracy, precision, recall, and F1-score. The "weighted average" accounts for class imbalance, while "support" indicates the number of test samples.
[a]Indicates the tuned RF model with selected features outperformed the other models.
Abbreviations: RF: Random forest; XGBoost: Extreme gradient boosting.

outlines the macro-average metrics and provides a detailed classification report.

The evaluation of the diagnostic classifiers highlighted the superior performance of the "neuropsychological assessment" classifier compared to the other two. Leveraging the variables CDGLOBAL (clinical dementia rating [CDR]), MMSCORE (mini-mental state examination [MMSE]), LIMMTOTAL (logical memory immediate recall), and LDELTOTAL (logical memory delayed recall), this classifier achieved a remarkable 90% accuracy in classifying AD cases. These variables were modeled using optimal hyperparameters—"n_estimators" = 100, "min_samples_split" = 15, "min_samples_leaf" = 1, and "max_depth" = 50—identified through randomized search with five-fold cross-validation and 100 iterations. Performance metrics of the diagnosis classifiers are presented in Table 3.

In terms of macro-average metrics, precision, recall, and F1 scores were all approximately 0.86, indicating consistent and balanced performance across all classes. Furthermore, the weighted-average precision, recall, and F1 scores exceeded 0.90, demonstrating excellent overall performance, with precision slightly surpassing recall. This detailed evaluation supports the effectiveness of the "neuropsychological assessment" classifier in accurately classifying AD cases. The Tables A1 and A2 outline the macro-average metrics and provide a detailed classification report.

6. Discussion

This study focused on developing robust multi-class classification models to predict AD across three distinct groups—HC, individuals with MCI, and diagnosed AD patients—and selecting the best-performing model based on its evaluation metrics. The results obtained from the optimal model could contribute to the early diagnosis of disease progression and provide valuable insights for advancing diagnostic methods and treatment strategies.

Table 3. Performance metrics of the diagnosis classifiers

Diagnostic classifier	Accuracy	Weighted average			
		Precision	Recall	F1-score	Support
Medical history variables	0.52	0.43	0.52	0.46	111
Neuropsychological assessment variables	0.90[a]	0.91	0.90	0.90	134
Blood analysis and ApoE genotype variables	0.65	0.85	0.68	0.66	148

Notes: This table presents the performance of three classifiers, each constructed using a single feature group—medical history variables, blood analysis and ApoE genotype data, and neuropsychological/clinical test results. The "neuropsychological assessment" classifier is further broken down into four individual cognitive tests: CDGLOBAL (clinical dementia rating), MMSCORE (mini-mental state examination), LIMMTOTAL (logical memory immediate recall), and LDELTOTAL (logical memory delayed recall). All classifiers were developed using the tuned Random Forest algorithm.
Abbreviation: ApoE: Apolipoprotein E.

Several data mining techniques used in this research, particularly feature importance and feature selection, yielded information that may inform further studies on this debilitating condition.

The comparative analysis between RF and XGBoost models, using the complete dataset, revealed detailed differences in their performance metrics, offering valuable insights into their predictive capabilities. Initially, both the simple RF and tuned RF models demonstrated a commendable overall accuracy of 88%, reflecting their ability to generate accurate predictions. This finding underscores the robustness of the RF algorithm in identifying complex patterns within the dataset. Furthermore, their high precision scores (90%) highlight the model's effectiveness in minimizing false positives—a critical factor in healthcare applications and resource optimization decision-making.

Additionally, the notable recall scores (88%) confirm the model's ability to correctly identify relevant cases in the dataset. The consistent F1-scores of 88% across both RF models further validate their balance between precision and recall, indicating their resilience to class imbalances and capacity to maintain predictive integrity. In contrast, the simple and tuned XGBoost models, while showing competitive performance, exhibited slightly lower accuracy scores of 86% and 85%, respectively. This indicates a slight reduction in overall predictive capability compared to the RF models. Nevertheless, the XGBoost models maintained comparable precision and recall scores—approximately 87% and 86%, respectively—demonstrating a consistent ability to minimize false positives and accurately detect relevant cases. Despite this slight decrement in accuracy, the F1-scores of 87% (simple XGBoost) and 85% (tuned XGBoost) indicate a well-maintained balance between precision and recall, affirming their reliability to sustain predictive accuracy across multiple evaluation metrics.

Both the simple RF and simple XGBoost models, utilizing selected variables, exhibited comparable accuracies of 86% and 85%, respectively, suggesting similar predictive performance. However, upon tuning, the RF model demonstrated a notable improvement, achieving an impressive accuracy of 90% and outperforming the tuned XGBoost model, which attained a respectable score of 89%. This enhancement underscores the effectiveness of fine-tuning in optimizing the RF algorithm's predictive capabilities, potentially making it a preferred choice in scenarios where maximizing prediction accuracy is crucial.

Additionally, evaluating precision, recall, and F1-score metrics provided a more comprehensive understanding of model performance beyond overall accuracy. Both the simple and tuned RF models consistently achieved higher precision, recall, and F1 scores compared to their XGBoost counterparts. Specifically, the tuned RF model yielded the highest scores across all three metrics, indicating superior ability to minimize false positives while effectively capturing relevant instances from the dataset. While the XGBoost models also demonstrated good precision, recall, and F1 scores, they slightly underperformed relative to the RF models, suggesting a moderate reduction in their effectiveness at minimizing misclassifications and accurately detecting relevant cases.

The predictive simple RF model from a previous study achieved an impressive accuracy of 96.05% for a binary classification task using all features of the AIBL non-imaging dataset.[3] In comparison, our best model—the tuned RF model using selected features—achieved a slightly lower accuracy of 90%. When comparing equivalent models from both studies, our simple RF model using all features yielded an accuracy of 88%. However, it is important to note that the prior study addressed a binary classification problem, whereas our study considered all three AD-related classes. This difference in classification scope likely accounts for the observed decrease in accuracy. The added complexity of distinguishing among three classes inherently increases the challenge and may reduce model performance relative to a binary setting.

Therefore, while our model's accuracy may appear slightly lower, its ability to classify across multiple classes provides valuable insight into the severity of AD. Furthermore, in our study, the train–test split was performed prior to preprocessing, supporting the model's ability to generalize to unseen data. In contrast, the previous study preprocessed the entire dataset, except for SMOTE, which may have contributed to their enhanced performance. Nonetheless, both studies consistently identified CDGLOBAL (CDR), MMSCORE (MMSE score), LIMMTOTAL (logical memory immediate recall), and LDELTOTAL (logical memory delayed recall) as the most informative predictors.

The comparison across classifiers based on medical history, neuropsychological assessment, and blood analysis with ApoE genotype variables offered valuable insights for medical diagnostics and predictive modeling. Initially, the classifier utilizing neuropsychological assessment variables emerged as the top performer, displaying impressive accuracy, precision, recall, and F1-score metrics, all exceeding 90%. This underscores the robust predictive capability of neuropsychological assessment data, highlighting its potential as a crucial diagnostic tool for AD. However, the classifier relying on medical history variables exhibited substantially lower performance metrics, with accuracy, precision, recall, and F1 scores hovering around 52%. This indicates its limited predictive accuracy when used in isolation. Despite its relatively lower accuracy, the classifier based on blood analysis and ApoE genotype variables demonstrated notable improvement. With precision at 85% and recall at 68%, resulting in an F1-score of 66%, the classifier shows promise in enhancing predictive accuracy and diagnostic capabilities by incorporating blood analysis and genetic data.

Both the existing and the present study identified that the classifier based on neuropsychological assessment variables as the most effective, consistently demonstrating exceptional performance metrics. Palmqvist[26] underscored the significance of the MMSE score in predicting the transition from MCI to AD. Similarly, Bloch and Friedrich[27] concluded that cognitive test results, including MMSE and CDR values, were the most informative features for effectively classifying AD. These findings highlight the

critical role of cognitive assessments in the early detection and diagnosis of AD.

7. Conclusion and future work

In this study, non-imaging data from the AIBL were analyzed to classify AD into three classes: HC, individuals with MCI, and those diagnosed with AD. Extensive data cleaning and exploration were conducted to reveal underlying patterns and extract information using various data mining techniques and statistical methods, including correlation analysis, feature association analysis, feature importance analysis, and feature selection.

To address the challenge of class imbalance, synthetic oversampling methods were employed to generate artificial samples to balance the target classes. Subsequently, the data were modeled and evaluated using advanced non-parametric ML algorithms, such as RF and XGBoost, first with complete feature set and then with selected features obtained through the feature selection process. Fine-tuning techniques were applied to enhance predictive accuracy. The results from these models underwent thorough evaluation to determine the most effective algorithm. The tuned RF model emerged as the top performer, achieving an accuracy of 90%, with precision, recall, and F1 scores also reaching 90%.

Furthermore, to reduce the diagnosis cost of AD and provide valuable insights toward developing a more reliable and affordable diagnostic tool, the data were segmented into three main variable groups: medical history, neuropsychological assessment, and blood analysis with ApoE genotype variables. Corresponding ML models were then developed using the fine-tuned model of the best-performing algorithm. The "neuropsychological assessment" classifier emerged as the most effective, exhibiting an exceptional accuracy of 90%.

Beyond its strong classification performance, this study presents a replicable and cost-effective methodology that may benefit other research groups, clinical practitioners, and public health systems aiming to improve early detection of AD. By leveraging non-imaging data—including widely available neuropsychological assessments—our approach avoids the high costs and limited accessibility associated with imaging-based diagnostics. The use of interpretable ML models, combined with robust feature selection and data preprocessing techniques, facilitates deployment in diverse clinical or research settings without the need for extensive computational infrastructure. Moreover, the proposed methodology can be adapted to other populations or datasets, supporting generalizability studies and cross-cohort validation efforts. This makes it particularly relevant for low-resource settings or large-scale screening efforts where rapid, accurate, and affordable tools are essential. Ultimately, this framework serves as a foundation for developing intelligent, personalized diagnostic support systems that prioritize early intervention and optimized resource allocation.

Given the time constraints of this research, explicit handling of outliers, multicollinearity, or distribution abnormalities was not performed. However, the selected models possess built-in capabilities to address these issues. Involving domain experts to directly address these factors could further enhance model performance. Moreover, the models were fine-tuned using the "RandomizedSearchCV" function; however, a more exhaustive approach, such as "GridSearchCV,"[28] could potentially yield better parameters by exploring a wider range of combinations. Although this study focused solely on direct multi-class classification, alternative approaches such as one-versus-one and one-versus-the-rest[29] may offer additional insights. Acknowledging these limitations provides a pathway for future research and further exploration of the problem.

Acknowledgments

None.

Funding

None.

Conflict of interest

The authors declare that they have no competing interests.

Author contributions

Conceptualization: All authors
Formal analysis: Adhinrag Kalarikkal Induchudan
Investigation: All authors
Methodology: All authors
Writing–original draft: Adhinrag Kalarikkal Induchudan
Writing–review & editing: All authors

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data

Data is available from the corresponding author upon reasonable request.

References

1. Alzheimer's Association. 2023 Alzheimer's disease facts and

figures. *Alzheimers Dement*. 2023;19(4):1598-1695.

doi: 10.1002/alz.13016.

2. Yang Q, Li X, Ding X, Xu F, Ling Z. Deep learning-based speech analysis for Alzheimer's disease detection: A literature review. *Alzheimers Res Ther.* 2022;14(1):186.

doi: 10.1186/s13195-022-01131-3

3. Shahbaz M, Ali S, Guergachi A, Niazi A, Umer A. Classification of Alzheimer's Disease Using Machine Learning Techniques. In: *Proceedings of the 12th International Joint Conference on Biomedical Engineering Systems and Technologies*. Prague, Czech Republic; 2019. p. 296-303.

doi: 10.5220/0007949902960303

4. *AIBL Study ADNI Non-imaging Data*. aibl.csiro.au. Available from: https://aibl.csiro.au/adni/nonimaging.php [Last accessed on 2024 Apr 30].

5. *ADNI. About*. Available from: https://adni.loni.usc.edu/about [Last accessed on 2024 Apr 30].

6. Rahman M, Prasad G. Comprehensive study on machine learning methods to increase the prediction accuracy of classifiers and reduce the number of medical tests required to diagnose Alzheimer's disease. *arXiv (Machine Learning).* 2022;1-10.

doi: 10.48550/arXiv.2212.00414

7. Wirth R, Hipp J. *CRISP-DM: Towards a Standard Process Model for Data Mining*. Available from: https://cs.unibo.it/~danilo.montesi/CBD/Beatriz/10.1.1.198.5133.pdf [Last accessed on 2024 Apr 30].

8. Harrikrishna NB. *Confusion Matrix, Accuracy, Precision, Recall, F1 Score*. Medium; 2020. Available from: https://medium.com/analytics-vidhya/confusion-matrix-accuracy-precision-recall-f1-score-ade299cf63cd [Last accessed on 2024 Apr 30].

9. Scikit-Learn. *Scikit-learn: Machine Learning in Python*. Scikit-Learn; 2019. Available from: https://scikit-learn.org/stable [Last accessed on 2024 Apr 30].

10. Nabel R. *PyDrive: Google Drive API Made Easy*. PyPI. Available from: https://pypi.org/project/pydrive [Last accessed on 2024 Apr 30].

11. Bhandari A. *Multicollinearity. Causes, Effects and Detection Using VIF*. Analytics Vidhya; 2023. Available from: https://www.analyticsvidhya.com/blog/2020/03/what-is-multicollinearity/#:~:text=multicollinearity%20is%20a%20statistical%20phenomenon [Last Accessed on 2024 Apr 30].

12. Turney S. *Chi-Square (χ2) Tests. Types, Formula and Examples*. Scribbr; 2022. Available from: https://www.scribbr.com/statistics/chi-square-tests [Last accessed on 2024 Apr 30].

13. *Pandas.factorize -- Pandas 1.5.3 Documentation*. Available from: https://pandas.pydata.org/docs/reference/api/pandas.factorize.html [Last accessed on 2024 Apr 30].

14. Goyal C. *Data Leakage and Its Effect on the Performance of an ML Model*. Analytics Vidhya; 2021. Available from: https://www.analyticsvidhya.com/blog/2021/07/data-leakage-and-its-effect-on-the-performance-of-an-ml-model [Last accessed on 2024 Apr 30].

15. Stekhoven D, Bühlmann P. MissForest--non-parametric missing value imputation for mixed-type data. *Bioinformatics*. 2012;28(1):112-118.

doi: 10.1093/bioinformatics/btr597

16. Saeys Y, Inza I, Larranaga P. A review of feature selection techniques in bioinformatics. *Bioinformatics*. 2007;23(19):2507-2517.

doi: 10.1093/bioinformatics/btm344

17. Malato G. *Feature Selection with Random Forest*. Your Data Teacher; 2021. Available from: https://www.yourdatateacher.com/2021/10/11/feature-selection-with-random-forest [Last accessed on 2024 Apr 30].

18. Tanuja D, Goutam S. Classification of imbalanced big data using SMOTE with rough random forest. *Int J Eng Adv Technol.* 2019;9:5174.

doi: 10.35940/ijeat.B4096.129219

19. Brownlee J. *Parametric and Nonparametric Machine Learning Algorithms*. Machine Learning Mastery; 2016. Available from: https://machinelearningmastery.com/parametric-and-nonparametric-machine-learning-algorithms [Last accessed on 2024 Apr 30].

20. Breiman L. Random forests. *Mach Learn*. 2001;45(1):5-32.

doi: 10.1023/a:1010933404324

21. N. Room. *How to Use XGBoost for Time-Series Forecasting?* Datadance; 2024. Available from: https://datadance.ai/machine-learning/how-to-use-xgboost-for-time-series-forecasting/#step-3-handling-missing-values-and-outliers [Last accessed on 2024 Apr 30].

22. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16*. 2016. 785-794.

doi: 10.1145/2939672.2939785

23. Scikit-Learn. *Sklearn.Model_Selection. RandomizedSearchCV - Scikit-Learn 0.21.3 Documentation*. Scikit-Learn; 2019. Available from: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.RandomizedSearchCV.html [Last accessed on 2024 Apr 30].

24. Scikit. *3.3. Metrics and Scoring: Quantifying the Quality of Predictions - Scikit-Learn 0.23.2 Documentation*. Scikit-Learn. Available from: https://scikit-learn.org/stable/modules/model_evaluation.html#classification-report [Last accessed on 2024 Apr 30].

25. Lanier ST. *Choosing Performance Metrics*. Medium; 2020. Available from: https://towardsdatascience.com/choosing-

performance-metrics-61b40819eae1 [Last accessed on 2024 Apr 30].

26. Palmqvist S. Comparison of brief cognitive tests and CSF biomarkers in predicting Alzheimer's disease in mild cognitive impairment: Six-year follow-up study. *PLoS One.* 2012;7(6):e38639.

    doi: 10.1371/journal.pone.0038639

27. Bloch L, Friedrich CM. Data analysis with shapley values for automatic subject selection in Alzheimer's disease data sets using interpretable machine learning. *Alzheimers Res Ther.* 2021;13(1):155.

    doi: 10.1186/s13195-021-00879-4

28. Shah R. *GridSearchCV. Tune Hyperparameters with GridSearchCV*. Analytics Vidhya; 2021. Available from: https://www.analyticsvidhya.com/blog/2021/06/tune-hyperparameters-with-gridsearchcv [Last accessed on 2024 Apr 30].

29. Brownlee J. *How to Use One-vs-Rest and One-vs-One for Multi-Class Classification*. Machine Learning Mastery; 2020. Available from: https://machinelearningmastery.com/one-vs-rest-and-one-vs-one-for-multi-class-classification [Last accessed on 2024 Apr 30].

Appendices

Table A1. Classification performance of machine learning models using complete and selected features

Dataset type		Classification report								
Complete features		Simple RF					Tuned RF			
		Precision	Recall	F1-score	Support		Precision	Recall	F1-score	Support
	HC	0.95	0.93	0.94	40	HC	0.95	0.93	0.94	40
	MCI	0.70	0.89	0.78	18	MCI	0.70	0.89	0.78	18
	AD	1.00	0.75	0.86	16	AD	1.00	0.75	0.86	16
	Macro-average	0.88	0.85	0.86	74	Macro-average	0.88	0.85	0.86	74
		Confusion matrix					Confusion matrix			
		37	3	0			37	3	0	
		2	16	0			2	16	0	
		0	4	12			0	4	12	
		Simple XGBoost					Tuned XGBoost			
		Precision	Recall	F1-score	Support		Precision	Recall	F1-score	Support
	HC	0.93	0.93	0.93	40	HC	0.90	0.93	0.91	40
	MCI	0.71	0.83	0.77	18	MCI	0.70	0.78	0.74	18
	AD	0.92	0.75	0.83	16	AD	0.92	0.75	0.83	16
	Macro-average	0.85	0.84	0.84	74	Macro-average	0.84	0.82	0.83	74
		Confusion matrix					Confusion matrix			
		37	3	0			37	3	0	
		2	15	1			3	14	1	
		1	3	12			1	3	12	
Selected features		Simple RF					Tuned RF			
		Precision	Recall	F1-score	Support		Precision	Recall	F1-score	Support
	HC	0.93	0.94	0.93	84	HC	0.97	0.93	0.95	84
	MCI	0.63	0.70	0.67	27	MCI	0.69	0.89	0.77	27
	AD	0.89	0.74	0.81	23	AD	0.95	0.78	0.86	23
	Macro-average	0.82	0.79	0.80	134	Macro-average	0.87	0.87	0.86	134
		Confusion matrix					Confusion matrix			
		79	5	0			78	6	0	
		6	19	2			2	24	1	
		0	6	17			0	5	18	
		Simple XGBoost					Tuned XGBoost			
		Precision	Recall	F1-score	Support		Precision	Recall	F1-score	Support
	HC	0.91	0.95	0.93	84	HC	0.97	0.93	0.95	84
	MCI	0.63	0.63	0.63	27	MCI	0.68	0.85	0.75	27
	AD	0.89	0.74	0.81	23	AD	0.90	0.78	0.84	23
	Macro-average	0.81	0.77	0.79	134	Macro-average	0.85	0.85	0.85	134
		Confusion matrix					Confusion matrix			
		80	4	0			78	6	0	
		8	17	2			2	23	2	
		0	6	17			0	5	18	

Abbreviations: AD: Alzheimer's disease; HC: Healthy control; MCI: Mild cognitive impairment; RF: Random forest.

Table A2. Classification report of diagnosis classifiers

Diagnosis classifier		Classification report			
Medical history variables		Precision	Recall	F1-score	Support
	HC	0.60	0.80	0.68	69
	MCI	0.09	0.04	0.06	24
	AD	0.25	0.11	0.15	18
	Macro-average	0.31	0.32	0.30	111
		Confusion matrix			
		55	10	4	
		21	1	2	
		16	0	2	
Neuropsychological assessment variables		Precision	Recall	F1-score	Support
	HC	0.97	0.93	0.95	84
	MCI	0.69	0.89	0.77	27
	AD	0.95	0.78	0.86	23
	Macro-average	0.87	0.87	0.86	134
		Confusion matrix			
		78	6	0	
		2	24	1	
		0	5	18	
Blood analysis and ApoE genotype variables		Precision	Recall	F1-score	Support
	HC	0.78	0.85	0.81	107
	MCI	0.19	0.14	0.16	22
	AD	0.47	0.37	0.41	19
	Macro-average	0.48	0.45	0.46	148
		Confusion matrix			
		91	11	5	
		16	3	3	
		10	2	7	

Abbreviations: AD: Alzheimer's disease; ApoE: Apolipoprotein E; HC: Healthy control; MCI: Mild cognitive impairment.