NEUROSCIENCE
RESEARCH ARTICLE
R. Franciotti et al. / Neuroscience 514 (2023) 143–152

Comparison of Machine Learning-based Approaches to Predict the Conversion to Alzheimer's Disease from Mild Cognitive Impairment

Raffaella Franciotti,[a]* Davide Nardini,[b] Mirella Russo,[a,c] Marco Onofrj[a,c] and Stefano L. Sensi[a,c,d]*, for the Alzheimer's Disease Neuroimaging Initiative[1] the Alzheimer's Disease Metabolomics Consortium ADMC[2]

[a] *Department of Neuroscience, Imaging, and Clinical Sciences, G. d'Annunzio University of Chieti-Pescara, Italy*

[b] *Biomedical Unit, ASC27 s.r.l., Rome, Italy*

[c] *Center for Advanced Studies and Technology – CAST, G. d'Annunzio University of Chieti-Pescara, Italy*

[d] *Institute for Advanced Biomedical Technologies, G. d'Annunzio University of Chieti-Pescara, Italy*

Abstract—In Mild Cognitive Impairment (MCI), identifying a high risk of conversion to Alzheimer's Disease Dementia (AD) is a primary goal for patient management. Machine Learning (ML) algorithms are widely employed to pursue data-driven diagnostic and prognostic goals. An agreement on the stability of these algorithms –when applied to different biomarkers and other conditions– is far from being reached. In this study, we compared the different prognostic performances of three supervised ML algorithms fed with multimodal biomarkers of MCI subjects obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database. Random Forest, Gradient Boosting, and eXtreme Gradient Boosting algorithms predict MCI conversion to AD. They can also be simultaneously employed –with the voting procedure– to improve predictivity. AD prediction accuracy is influenced by the nature of the data (i.e., neuropsychological test scores, cerebrospinal fluid AD-related proteins and APOE ε4, cerebral structural MRI (sMRI) data). In our study, independent of the applied ML algorithms, sMRI data showed the lowest accuracy (0.79) compared to other classes. Multimodal data were helpful in the algorithms' performances by combining clinical and biological measures. Accordingly, using the three ML algorithms, the highest accuracy (0.90) was reached by employing neuropsychological and AD-related biomarkers. Finally, the feature selection procedure indicated that the most critical variables in the respective classes were the ADAS-Cog-13 scale, the medial temporal lobe and hippocampus atrophy, and the ratio between phosphorylated Tau and Aβ42 proteins. In conclusion, our data support the notion that using multiple ML algorithms and multimodal biomarkers helps make more accurate and solid predictions.

Key words: Alzheimer's disease dementia (AD) prediction, artificial intelligence, gradient boosting, machine learning (ML) algorithms, mild cognitive impairment (MCI), random forest (RF).

INTRODUCTION

Mild cognitive impairment (MCI) is a condition characterized by worsening cognition in one or more domains (e.g., memory, attention, language). At the same time, independence in daily living activities (Petersen et al., 2014) is maintained. This condition is

*Corresponding authors. Address: Department of Neuroscience, Imaging, and Clinical Sciences, University G. d'Annunzio of Chieti-Pescara Via Luigi Polacchi 11, Chieti 66100, Italy.
E-mail addresses: raffaella.franciotti@unich.it (R. Franciotti), ssensi@unich.it (S. L. Sensi).

[1] Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (http://adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in the analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how to apply/ADNI Acknowledgement List.pdf.

[2] Data used in the preparation of this article were generated by the Alzheimer's Disease Metabolomics Consortium (ADMC). As such, the investigators within the ADMC provided data but did not participate in the analysis or writing of this report. A complete listing of ADMC investigators can be found at: https://sites.duke.edu/adnimetab/team/.

Abbreviations: AdaBoost, Adaptive Boosting; ABETA, Ab42 peptide; AD, Alzheimer's Disease Dementia; ADAS-Cog-11 and 13, Alzheimer's Disease Assessment Scale-Cognitive Subscales-11 items and 13 items; ADNI, Alzheimer's Disease Neuroimaging Initiative; c-MCI, converting MCI; FAQ, Functional Activities Questionnaire; GB, Gradient Boosting; LM-DEL, Logical Memory-Delayed recall; MCI, Mild Cognitive Impairment; ML, Machine Learning; NPV, negative predictive value; PPV, positive predictive value; PTAU, phosphorylated tau protein; RAVLT-DEL and -IMM, Rey Auditory Verbal Learning Test-Delayed and Immediate recall; p-Tau, phosphorylated-Tau; RF, Random Forest; SHAP, SHapley Additive exPlanations; s-MCI, stable MCI; sMRI, structural MRI; t-Tau, total Tau; XGB, eXtreme Gradient Boosting.

https://doi.org/10.1016/j.neuroscience.2023.01.029
0306-4522/

synergistically driven by factors like primary or secondary neurodegenerative processes (Brem and Sensi, 2018), systemic alterations, and unhealthy lifestyle habits (Knopman, and Petersen, 2014). It can exhibit a stable course (s-MCI) or be prodromal (Albert et al., 2011) to Alzheimer's Disease Dementia (AD) (converting MCI, c-MCI). In recent years a large body of evidence indicates that the identification of MCI subjects at high risk of conversion can be highly improved with the use of Artificial Intelligence and Machine Learning (ML) algorithms (Varoquaux and Cheplygina, 2022). Thus, large datasets containing a wide array of biomarkers, combined with ML algorithms, are now widely employed to pursue data-driven diagnostic and prognostic goals (Rossini et al., 2022). However, although many studies have applied ML for predicting dementia, an agreement on the stability of these algorithms –when applied to different biomarkers and under other conditions– is far from being reached (Faouri et al., 2022). Selecting which ML algorithm to use is crucial, as each exhibits strengths and weaknesses. Specifically, for large and high-dimensional data, deep learning (DL) often outperforms shallow ML algorithms (Janiesch et al., 2021). However, the problematic availability of large amounts of clinical multimodal data, and the poor interpretability of the learning process and results make DL less exploitable in AD prediction. When the data size is small (lower than 1,000), Naive Bayes (Shree and Sheshadri, 2018), K-nearest neighbors (Dinu and Ganesan, 2019), and support vector machine (Syaifullah et al., 2021), are the most commonly used ML classifiers together with logistic regression (Rohini and Surendran, 2021). The most widely used tree-based ML algorithms are Random Forest (RF), Adaptive Boosting (AdaBoost), and Gradient Boosting (GB), which utilize decision trees within an ensemble through bagging (in RF) or boosting (in AdaBoost and GB) methods (Natras et al., 2022).

In this study, we applied RF and GB Machine which consistently outperform other models, as confirmed by many studies (Fernández-Delgado et al., 2014). In addition, the results of tree-based algorithms like RF and GB, are interpretable, unlike DL models like Deep Artificial Neural Network. Interpretability of results is a fundamental target for predicting dementia because it can point out new and unknown patterns in disease progression.

At the same time, these models are robust, and with a good performance so that they can achieve a good and excellent accuracy and prediction rate even with a modest number of sample data.

RF is an ensemble learning algorithm structured from a set of decision trees using the Bagging algorithm (Breiman, 2001). The trees grow by splitting a node with the best features chosen from a random subset of features and the best possible thresholds. The model's randomness can also be increased using random thresholds for each feature (Qi, 2012). GB and eXtreme Gradient Boosting (XGB) are ensemble learning algorithms that, unlike RF, are built on weak learners, i.e., shallow trees that sometimes can be made with only one level of decision. In addition, GB (Friedman, 2001) and XGB (Chen and Guestrin, 2016) are based on the boosting technique (Schapire and Freund, 2013). This method transforms many weak classifiers into one robust classifier, reducing bias and variance. The boosting method learns sequentially, building on the error of previous classifiers. XGB (Chen and Guestrin, 2016) is an optimization of GB with advanced regularization techniques to prevent overfitting.

In a previous study (Massetti et al., 2022), we obtained an 0.86 accuracy in predicting MCI to AD conversion when employing a RF algorithm to neuropsychological data, cerebrospinal fluid (CSF) levels of AD-related proteins, structural Magnetic Resonance Imaging (sMRI), omics data, and apolipoprotein E genotype, all data obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (https://adni.loni.usc.edu).

In this study, we compared different performances in MCI to AD prediction models obtained by using RF, GB and XGB. These approaches made use of combinations of multimodal biomarkers. The goal was to test their stability and reliability in different clinical environments. The results of the three ML algorithms were combined with a voting technique to reach the best possible performance.

As secondary outcomes, we also assessed variable importance in the prediction and cut-off prediction values of the essential features by applying the SHapley Additive exPlanations (SHAP) algorithm (Lundberg and Lee, 2017) to the model with the best accuracy.

EXPERIMENTAL PROCEDURES

Data

Data used in the preparation of this article were obtained from the ADNI database (https://adni.loni.usc.edu) and used in a previous paper (Massetti et al., 2022). The ADNI was launched in 2003 as a public–private partnership led by Principal Investigator Michael W. Weiner, MD. The clinical coordination center of ADNI established a network of clinical sites. It developed the clinical protocol and informed consent, which is distributed to the sites for local institutional review board ethical approval. One such institution is the Office for the Protection of Research Subjects at the University of Southern California. Participants provided written informed consent for the study. More details can be found at adni.loni.usc.edu. The primary goal of ADNI has been to test whether serial MRI, positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of MCI and early AD.

This study grouped MCI baseline data into three classes: neuropsychological features, sMRI-related data, and AD-related biomarkers. Neuropsychological variables are detailed in Table 1.

MRI data (Image Collections, https://adni.loni.usc.edu) were acquired with a Magnetization Prepared-RApid Gradient Echo (MP-RAGE) protocol by employing a Philips 3 T MRI scanner (https://adni.loni.usc.edu/wp-content/uploads/2010/05/ADNI2_MRI_Training_Manual_FINAL.pdf). T1-weighted images were obtained using 3D Turbo Field-Echo sequences (slice thickness of 1.2 mm,

Table 1. Neuropsychological measures included in the study and relative cognitive domains

Neuropsychological test Abbreviation	*Cognitive/functional domains*
Alzheimer's Disease Assessment Scale-Cognitive Subscales-11 items ***ADAS-Cog-11***	Verbal memory, gnosis (naming objects and fingers), praxis (constructional, ideational, ideomotor), orientation, and language
Alzheimer's Disease Assessment Scale-Cognitive Subscales-13 items ***ADAS-Cog-13***	Verbal memory, gnosis (naming objects and fingers), praxis (constructional, ideational, ideomotor), orientation, language, and selective attention/ psychomotor speed (number cancellation)
Boston Naming Test ***BNT***	Semantic memory and language
Category Fluency (Animals)	Semantic fluency
Clinical Dementia Rating Scale-Sum of boxes ***CDR-SB***	Staging of cognitive impairment through the assessment of memory, orientation, judgment/ problem solving, community life, domestic life/hobbies, and personal care
Clock Drawing Test ***CDT***	Constructional praxis
Clock Drawing Test ***CDT-DEL***	Constructional praxis, memory
Functional Activities Questionnaire ***FAQ***	Personal care
Logical Memory-Immediate recall ***LM-IMM***	Narrative episodic memory (immediate recall)
Logical Memory-Delayed recall ***LM-DEL***	Narrative episodic memory (delayed recall)
Mini-Mental State Examination ***MMSE***	Orientation, verbal memory, attention, language, constructional praxis, writing
Rey Auditory Verbal Learning Test-Immediate recall ***RAVLT-IMM***	Episodic memory (word list immediate recall)
Rey Auditory Verbal Learning Test-Delayed recall ***RAVLT-DEL***	Episodic memory (word list delayed recall)
Rey Auditory Verbal Learning Test-Total score ***RAVLT-TOT***	Episodic memory (word list recall, immediate + delayed recall)
Trail Making Test-Part A ***TMT-A***	Attention, processing speed, perceptual-scanning skills, cognitive flexibility
Trail Making Test-Part B ***TMT-B***	Attention, executive functions, processing speed, perceptual-scanning skills, cognitive flexibility

repetition time/echo time of 6.8/3.1 ms). Like the previous paper (Massetti et al., 2022), the sMRI analysis was performed with Freesurfer (version 6.0). Automatic reconstruction and labeling of cortical and subcortical regions was achieved with the ''recon-all-all'' command line, according to Desikan-Killiany Atlas (Desikan et al., 2006). The volumes of the brain regions, computed with *asegstats2table*, were normalized by dividing by the total intracranial volume of each patient, while the cortical thicknesses were calculated automatically by *aparcstats2table*. All the calculated MRI variables are reported in Table 2.

Table 2. List of brain areas extracted as sMRI variables parcellated by FreeSurfer.

	*Brain regions**
Normalized brain volume	Lateral ventricle
	Thalamus
	Hippocampus
	Amygdala
Cortical thickness *(Desikan-Killiany atlas)*	Entorhinal
	Fusiform
	Inferior temporal
	Isthmus cingulate
	Lateral orbitofrontal
	Medial orbitofrontal
	Middle Temporal
	Parahippocampal
	Posterior cingulate
	Precuneus
	Rostral anterior cingulate
	Superior temporal
	Supramarginal
	Temporal pole

*Right and left brain areas were separately parcellated.

AD-related biomarkers encompassed CSF proteins like Aβ42 (ABETA), total-Tau (t-Tau), phosphorylated-Tau (p-Tau), p-Tau/ABETA ratio (Delli Pizzi et al., 2019), and APOE ε4 genotype from blood samples. CSF protein levels were measured using the Roche fully-automated immunoassay platform (Cobas e601) and immunoassay reagents.

ML algorithms

Single classes or class combinations with a sample size lower than 300 were excluded from the analyses. ML algorithms implemented in Python were applied to identify the best strategy to divide MCI subjects into two groups: individuals who converted to AD within the 36-month follow-up (c-MCI) or not (s-MCI). From the entire dataset, information of 85% of subjects was randomly extracted and used for the training phase performed by RF, GB, and XGB separately. After the training stage, the testing phase was applied by each ML algorithm to the remaining 15% of the dataset.

We used a random search for RF, GB, and XGB as a hyperparameter optimization technique (Bergstra and Bengio, 2012). This step is critical because ML models need to set hyperparameters to best tune the algorithm for data testing. Random search is one of the most common hyperparameter optimization methods, based on randomly extracting values in a bounded domain of hyperparameter values. From several extractions, the algorithm fits the model for each combination of values and then selects the hyperparameter combination that provides the best performance.

Table 3. Performance measures among RF, GB, and XGB applied with multimodal biomarkers. For each performance index, the best values among RF, GB, and XGB are highlighted in bold.

	Neuropsychological data (n = 587)			*sMRI data (n = 318)*			*AD-related biomarkers (n = 422)*			*Neuropsychological + AD-related biomarkers (n = 422)*			*Neuropsychological + sMRI data (n = 318)*		
All features	**RF**	**GB**	**XGB**	**RF**	**GB**	**XGB**	**RF**	**GB**	**XGB**	**RF**	**GB**	**XGB**	**RF**	**GB**	**XGB**
Accuracy	**0.83**	0.82	0.81	**0.79**	**0.79**	0.77	**0.85**	0.74	0.74	0.87	0.85	**0.89**	0.83	**0.88**	0.83
PPV	0.78	0.78	**0.81**	**0.89**	**0.89**	0.84	0.83	**0.91**	**0.91**	**0.95**	0.90	**0.95**	0.82	**0.90**	0.89
NPV	**0.87**	0.85	0.81	**0.73**	**0.73**	0.72	**0.87**	0.70	0.70	0.83	0.83	**0.85**	0.85	**0.86**	0.80
Sensitivity	**0.81**	0.78	0.69	**0.67**	**0.67**	**0.67**	**0.80**	0.40	0.40	0.72	0.72	**0.76**	0.82	**0.82**	0.73
Specificity	0.85	0.85	**0.89**	**0.92**	**0.92**	0.88	0.89	**0.97**	**0.97**	**0.97**	0.94	**0.97**	0.85	**0.92**	**0.92**
Error	±0.08	±0.08	±0.08	±0.12	±0.12	±0.12	±0.09	±0.11	±0.11	±0.09	±0.09	±0.08	±0.11	±0.09	±0.11
Feature selection	**RF**	**GB**	**XGB**	**RF**	**GB**	**XGB**	**RF**	**GB**	**XGB**	**RF**	**GB**	**XGB**	**RF**	**GB**	**XGB**
Accuracy	**0.82**	**0.82**	0.81	0.75	**0.77**	**0.77**	NA	NA	NA	0.84	0.80	**0.85**	0.81	**0.88**	0.81
PPV	0.78	0.79	**0.81**	0.83	0.84	**0.88**	NA	NA	NA	**0.90**	0.84	**0.90**	0.81	**0.83**	0.81
NPV	**0.85**	0.84	0.81	0.70	**0.72**	0.71	NA	NA	NA	0.81	0.79	**0.83**	0.82	**0.92**	0.82
Sensitivity	**0.78**	0.75	0.69	0.62	**0.67**	0.62	NA	NA	NA	0.68	0.64	**0.72**	0.77	**0.91**	0.77
Specificity	0.85	0.87	**0.89**	0.88	0.88	**0.92**	NA	NA	NA	**0.94**	0.92	**0.94**	**0.85**	**0.85**	**0.85**
Error	±0.08	±0.08	±0.08	±0.12	±0.12	±0.12	NA	NA	NA	±0.09	±0.10	±0.09	±0.11	±0.09	±0.11

Abbreviations: GB = Gradient Boosting; sMRI = structural Magnetic Resonance Imaging; NPV = Negative Predictive Value; PPV = Positive Predictive Value; RF = Random Forest; XGB = eXtreme Gradient Boosting.

Table 4. Performance measures following voting technique among RF, GB, and XGB. Bold cells indicate that the values with the voting technique were higher than those obtained by applying RF, GB, or XGB separately.

	Neuropsychological data (n = 587)	*sMRI data (n = 318)*	*AD-related biomarkers (n = 422)*	*Neuropsychological + AD-related biomarkers (n = 422)*	*Neuropsychological + sMRI data (n = 318)*
All features	**Voting**	**Voting**	**Voting**	**Voting**	**Voting**
Accuracy	0.82	0.77	0.75	**0.90**	0.85
PPV	0.81	0.84	0.86	0.95	0.90
NPV	0.83	0.72	0.72	**0.88**	0.83
Sensitivity	0.72	0.67	0.48	**0.80**	0.77
Specificity	0.89	0.88	0.94	0.97	**0.92**
Error	±0.08	±0.12	±0.11	±0.08	±0.10
Feature selection	**Voting**	**Voting**	**Voting**	**Voting**	**Voting**
Accuracy	0.80	0.75	NA	0.85	0.81
PPV	0.77	0.83	NA	**0.94**	0.81
NPV	0.82	0.70	NA	0.81	0.82
Sensitivity	0.72	0.62	NA	0.68	0.77
Specificity	0.85	0.88	NA	**0.97**	0.85
Error	±0.08	±0.12	NA	±0.09	±0.11

Abbreviations: GB = Gradient Boosting; sMRI = structural Magnetic Resonance Imaging; NPV = Negative Predictive Value; PPV = Positive Predictive Value; RF = Random Forest; XGB = eXtreme Gradient Boosting.

RF, GB, and XGB were also applied after RF features selection technique, an algorithm that can handle feature selection problems even in cases with a higher number of variables (Chen et al., 2020). Feature selection simplifies the model by reducing the number of variables, decreasing training time, reducing overfilling by enhancing generalization, and avoiding the course of dimensionality (Chen et al., 2020). Accuracy, positive predictive value (PPV), negative predictive value (NPV), sensitivity, and specificity were calculated for RF, GB, and XGB before and after RF feature selection to evaluate the performance of each ML algorithm.

The same performance estimates were evaluated after the voting technique, an ensemble method combining multiple models' performances to make predictions. Hard voting prediction (Mishra et al., 2021) with the highest frequency was used.

Biomarkers: Prediction threshold

For the secondary outcomes, the SHAP algorithm (Lundberg and Lee, 2017) was applied to the model with the best accuracy –among RF, GB, and XGB– for the most critical variables in the two class combinations.

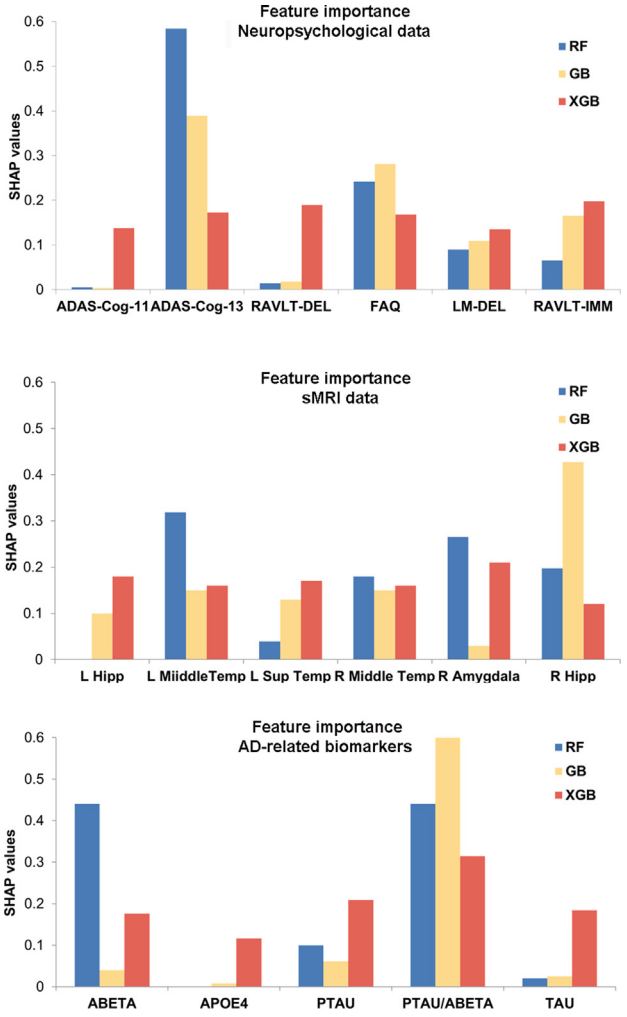

Fig. 1. Feature importance as obtained by the RF features selection technique employed with the most critical variables in single classes. Abbreviations: ABETA = Aβ42 peptide; ADAS-Cog-11 and 13 = Alzheimer's Disease Assessment Scale-Cognitive Subscales-11 items and 13 items; APOE4 = Apolipoprotein ε4 genotype; FAQ = Functional Activities Questionnaire; GB = Gradient Boosting; LM-DEL = Logical Memory-Delayed recall; MRI = Magnetic Resonance Imaging; PTAU = phosphorylated tau protein; RAVLT-DEL and -IMM = Rey Auditory Verbal Learning Test-Delayed and Immediate recall; RF = Random Forest; TAU = total tau; XGB = eXtreme Gradient Boosting.

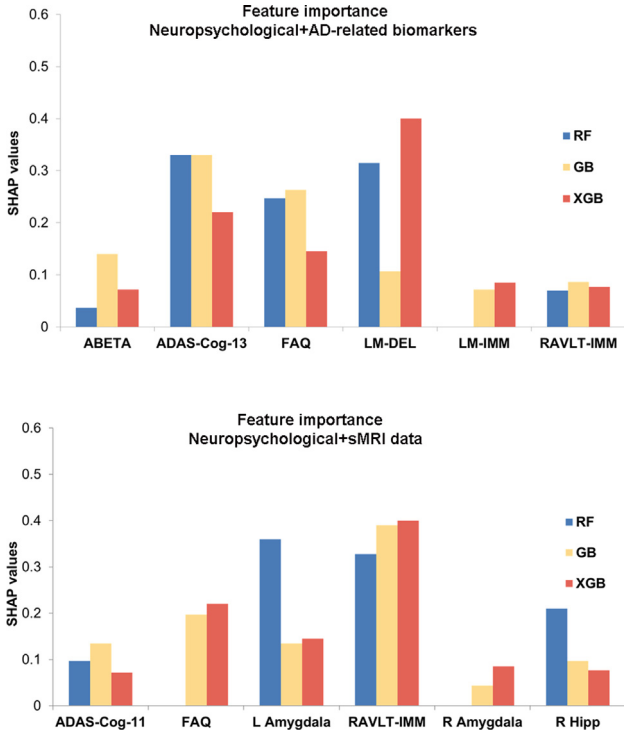

Fig. 2. Feature importance as obtained by the RF features selection technique employed with the most critical variables in class combinations. Abbreviations: ABETA = Aβ42 peptide; ADAS-Cog-11 and 13 = Alzheimer's Disease Assessment Scale-Cognitive Subscales-11 items and 13 items; FAQ = Functional Activities Questionnaire; GB = Gradient Boosting; Hipp = hippocampus; LM-IMM and -DEL = Logical Memory-Immediate and Delayed recall; MRI = Magnetic Resonance Imaging; RAVLT-IMM = Rey Auditory Verbal Learning Test-Immediate recall; RF = Random Forest; XGB = eXtreme Gradient Boosting.

SHAP is an Explainable Artificial Intelligence (XAI) analysis approach useful for understanding ML outputs. SHAP assigns to each feature an "importance" value for a particular prediction. This SHAP value indicates how much each feature helps predict the target. Thus, SHAP is extremely useful for interpreting ML results and predicting complex models. Each SHAP value of the feature can be considered as a force that either increases or decreases the prediction. The prediction starts from the baseline. The baseline for Shapley values is the average of all predictions. Each SHAP value is a force that pushes the prediction to increase (positive value) or decrease (negative value). Thus, the SHAP value for each instant data describes the model's output that does not necessarily represent the real world.

SHAP values for each instant data were calculated for the ML algorithm with the highest accuracy after the RF feature selection technique. Thus, for the most critical biomarker predicting the conversion to AD, biomarkers threshold values were estimated from the SHAP values calculated for each MCI subject.

RESULTS

ML algorithms performance

Performance measures were calculated for single classes and the combinations of neuropsychological features with AD-related biomarkers and sMRI-related data. As shown in Table 3, all ML algorithms applied to each class's variables achieved good accuracy, which was lower than 0.80 for sMRI data only. RF achieved the highest accuracy for neuropsychological features (0.83) and AD-related biomarkers (0.85). GB and XGB reached the best performance revealed by the highest accuracy, PPV, NPV, sensitivity, and specificity for combining neuropsychological features with sMRI data (0.88) and AD-related biomarkers (0.89). For the single classes of variables (i.e., neuropsychological data, sMRI data, and AD-related biomarkers), RF reached the highest value of NPV and sensitivity.

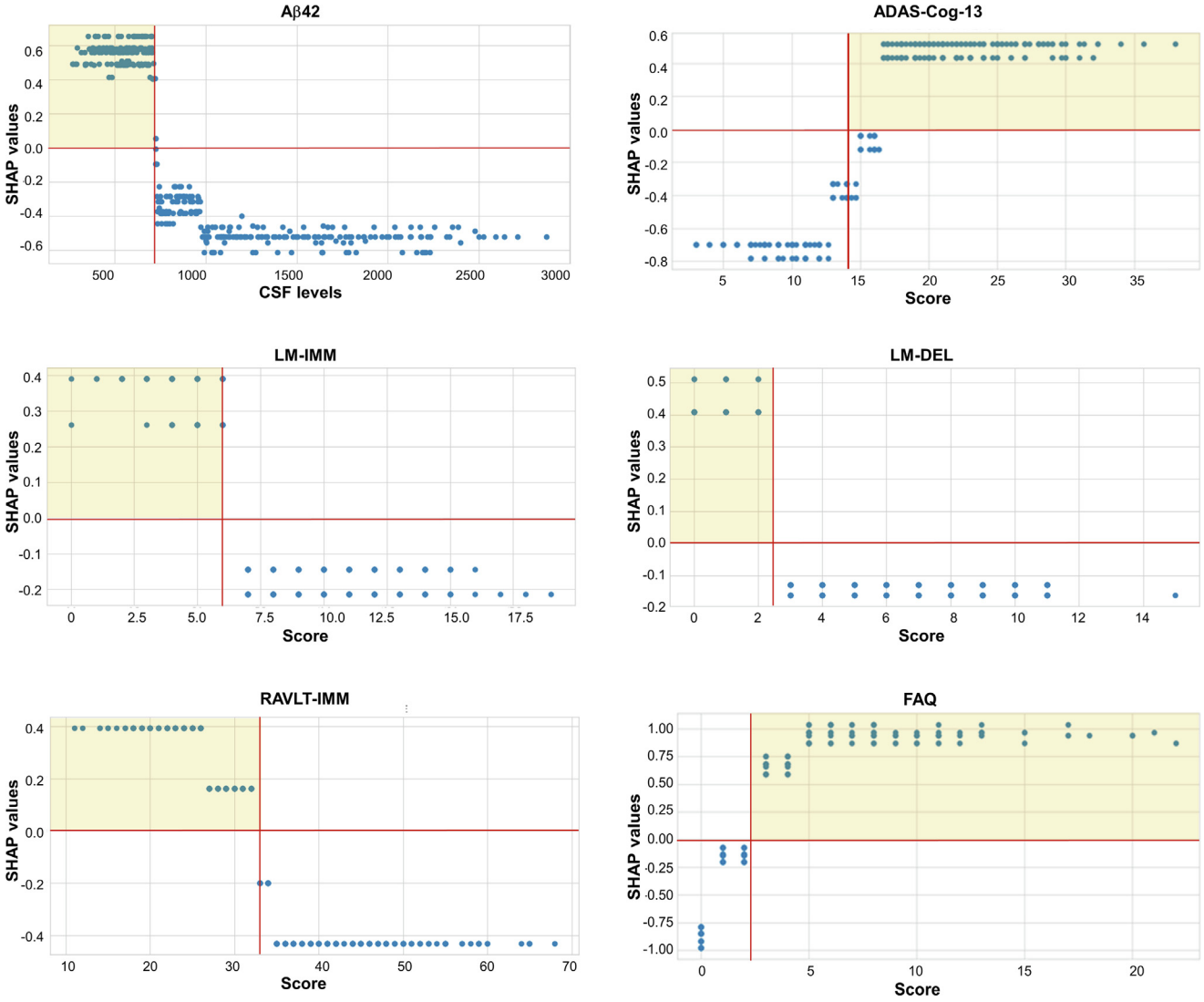

Fig. 3. SHAP values for the different feature values. Positive SHAP values within the yellow quadrant indicate probable conversion to AD within a 36-months follow-up. Values in the other sections indicate a low risk of conversion. The red cross for each variable indicates the decision threshold. Abbreviations: ADAS-Cog-13 = Alzheimer's Disease Assessment Scale-Cognitive Subscales-13 items; FAQ = Functional Activities Questionnaire; LM-IMM and -DEL = Logical Memory-Immediate and Delayed recall; RAVLT-IMM = Rey Auditory Verbal Learning Test-Immediate recall; SHAP = SHapley Additive exPlanations.

When the RF feature selection technique was applied before classification algorithms, the performances did not improve. They decreased in the case of some algorithms and classes (i.e., sMRI data and the combination of neuropsychological tests with AD-related biomarkers).

After applying the voting technique, performance values (shown in Table 4) stayed the same with the exception of a few conditions (i.e., accuracy, NPV, and sensitivity when combining neuropsychological and AD-related biomarkers or specificity in the combination of neuropsychological and sMRI data.

Feature selection: Biomarker's importance for different ML algorithms

RF feature selection on all the ML algorithms and classes showed that six features were the most appropriate quantity. Thus, the number was set to six for all the analyses. The importance of the variables was ranked. The sum of the importance values of the six most relevant features was set to 1 for each class and class combination.

The most relevant neuropsychological variables were related to global cognition (i.e., ADAS-Cog-13), immediate and delayed verbal memory (RAVLT-IMM and -DEL, LM-DEL), and daily functioning (FAQ). In the sMRI dataset, the most relevant measures were the normalized volumes of both hippocampi and right amygdala, cortical thickness of the left and right middle temporal lobe, and the left superior temporal area.

Figs. 1 and 2 show the differences, in terms of variable importance ranking, according to RF, GB, and XGB algorithms for the single classes and the class combinations, respectively. The ADAS-Cog-13 scale was the most critical neuropsychological parameter for single classes according to RF and GB models,

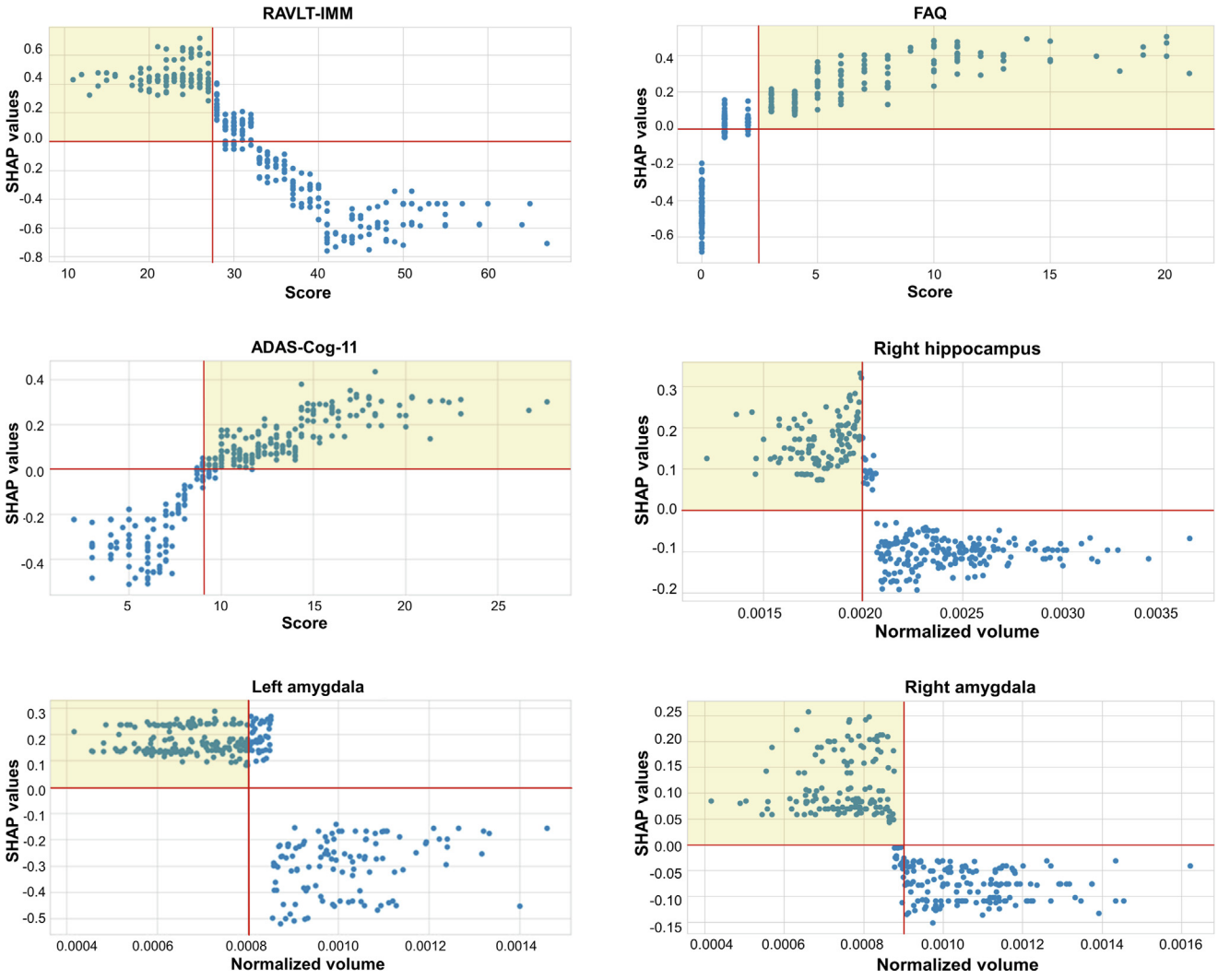

Fig. 4. SHAP values for the different feature values. Positive SHAP values within the yellow quadrant indicate a probable conversion to AD within a 36-months follow-up, while the values in the other sections indicate a low risk of conversion. The red cross for each variable indicates the decision threshold. Normalized volumes are all expressed in cm^3. Abbreviations: ADAS-Cog-11 = Alzheimer's Disease Assessment Scale-Cognitive Subscales-11 items; FAQ = Functional Activities Questionnaire; RAVLT-IMM = Rey Auditory Verbal Learning Test-Immediate recall; SHAP = SHapley Additive exPlanations.

whereas the XGB model classified the six variables as equally important. A similar trend of the XGB model was also observed for sMRI data, and AD-related biomarkers. On the other hand, the cortical thickness of the left middle temporal area and the normalized volume of the right hippocampus were chosen as the essential features to be used by the RF and GB models, respectively. Among the AD-related biomarkers, the three algorithms converged toward the most significant importance for the p-Tau/Aβ42 ratio.

When combining neuropsychological data with AD-related biomarkers, 5 out of the 6 most important variables were neuropsychological variables. ADAS-Cog-13 was confirmed as the key variable in the RF and GB models. In the combination of neuropsychological and sMRI data, the RAVLT-IMM was the most important measure for the GB and XGB models. In contrast, the volume of the left amygdala was the most critical feature for the RF model.

Biomarkers: Decision threshold values

SHAP value analysis was performed on the XGB model applied to the combination of neuropsychological data with AD-related biomarkers (ACC = 0.85) and on the GB model in the combination of neuropsychological and sMRI data (ACC = 0.88).

SHAP analyses to assess the most critical variables showed potential decision thresholds in predicting the conversion of MCI to AD within a 36-months follow-up. Figs. 3 and 4 show the SHAP values for the key variables in the class combinations. For the combination of neuropsychological data with AD-related biomarkers (Fig. 3), the model indicated a higher risk of early (i.e.,

within three years) conversion to AD with the following values: Aβ42 levels < 697 pg/mL, ADAS-Cog-13 score > 14, FAQ score > 3, LM-IMM score < 6, LM-DEL score < 2.5, and RAVLT-IMM score < 33.

For the combination of neuropsychological and MRI data (Fig. 4), the model indicated probable conversion to AD with the following conditions: FAQ score > 3, ADAS-Cog-11 score > 9, RAVL-IMM score < 29, left amygdala normalized volume < 0.8 mm^3, right amygdala normalized volume < 0.9 mm^3, and right hippocampus normalized volume < 2 mm^3.

DISCUSSION

The development of large databases and ML-based techniques provides a powerful tool for clinical studies. However, these approaches must be used in a tailored fashion. To that aim, we compared the eristic performances of three ML algorithms in the automated prediction of MCI conversion to AD. Our data show that RF performed better than GB and XGB when using neuropsychological data (list in Table 1). In contrast, GB and XGB were more accurate than RF when employed with class combinations (Table 3). Possible explanations for these differences relate to the random search of the hyperparameters values (i.e., the number and the depth of each decision tree, the shrinkage or learning rate for GB and XGB), which are fewer and easier to tune for RF than GB and XGB (Elgeldawi et al., 2021). Structural differences within algorithms can also be involved. Using a bagging (bootstrap aggregating) technique, RF reduces the dependence on a single tree by spreading the risk of error across multiple parallel decision trees built by different subsets of the training data. This procedure indirectly decreases the risk of data overfitting. Instead, by applying a boosting technique, GB and XGB generated the decision trees in a sequential manner, thereby learning the mistake from the previous ones (Cox et al., 2001). However, this procedure can lead to overfitting, especially when dealing with a large number of variables.

SMRI data (listed in Table 2) showed lower prediction accuracy (0.79) than other classes, independently from the employed ML algorithms. PPV and specificity values were good (0.89 and 0.92), indicating that brain atrophy is a specific but poorly sensitive (0.67) biomarker to predict the MCI conversion to AD. RF feature selection –when applied before ML algorithms– did not significantly increase performance, especially in the case of sMRI variables (Table 3). The combination of sMRI and neuropsychological variables increased all performance indices compared to each class's individual results. By combining neuropsychological and sMRI data, accuracy reached 0.88, PPV 0.90, and specificity 0.92. This result confirms the holistic nature of the disease. It also indicates that the variables ensemble of different clinical biomarkers (i.e., neuropsychological and anatomical variables) provides a better model estimation for AD prediction as it increased the accuracy to 0.88 when neuropsychological data or sMRI data reached 0.83 and 0.79 respectively.

In the case of AD-related biomarkers, the RF algorithm showed the highest accuracy (0.85), NPV (0.87), and sensitivity (0.80), whereas GB and XGB had the best PPV (0.91) and specificity (0.97). All algorithms showed that the p-Tau/Aβ42 ratio is the most critical parameter of the models. Despite the good accuracy of these standard AD-biomarkers, future studies on AD prediction could include biomarkers linked to synaptic dysfunction, a key pathological feature of AD, which is correlated to cognitive impairment and Tau-amyloid pathogenetic mechanisms (Spires-Jones and Hyman, 2014). Possible biomarkers related to synaptic failure could be presynaptic (Brinkmalm et al., 2014) and dendritic proteins (Casaletto et al., 2017), which are increased in CSF of MCI-AD and AD (Galasko et al., 2019) and Aβ-oligomers which have direct toxic effects at the synapse level (Williams and Serpell 2011). Neuronal and synaptic dysfunction can also be estimated by brain metabolic changes and functional disconnections obtained by 18F fluoro-deoxy-glucose positron emission tomography (FDG-PET) uptake values (Teng et al., 2020) and electroencephalographic (EEG) or magnetoencephalographic (MEG) markers (Poil et al., 2013; Mazaheri et al., 2018).

Other possible biomarkers in AD prediction could be also probe in serum levels of long-chain ceramides which might be predictive of hippocampal volume loss and cognitive decline in MCI subjects (Vozella et al., 2019).

The voting procedure did not increase performance measures (Table 4), suggesting that in some MCI subjects, only a single algorithm outperformed the others in accuracy values. Although the voting system has rarely exceeded the performance of a single algorithm (i.e., in the case of the combination of neuropsychological data with AD-related biomarkers), it is undoubtedly an attractive solution because it makes the system more stable and robust, providing more reliable outputs.

Results of the feature selection procedure indicated that the most important variables for each class were consistent within the three ML algorithms. The ADAS-Cog was a significant predictor of conversion from MCI to AD. SHAP analyses suggested that an ADAS-Cog-13 score > 14 or an ADAS-Cog-11 score > 9 (Figs. 3 and 4) could reliably predict an increased risk of disease progression within 36 months. The normalized volume of the left and right amygdale and right hippocampus were the most predictive sMRI variables, confirming previous studies on the contribution of these areas in the prediction to AD conversion (Apostolova et al., 2006; Poulin et al., 2011). The empirical decision threshold values from SHAP analysis could benefit clinical practice. The approach should be further investigated to confirm its reliability when used alone or in combination.

Comparing performances of the RF, GB, and XGB algorithms indicates that they are valid tools to predict MCI conversion to AD. Our results suggest using different performance indices (i.e., accuracy, PPV, NPV, sensitivity, and specificity) to achieve a comprehensive view of the functioning of the ML algorithms. Indeed,

sensitivity and specificity are independent of disease prevalence, whereas PPV increases when prevalence increases and NPV decreases when prevalence rises. Our results also indicate that using multiple ML algorithms could be helpful in achieving more accurate and reliable predictions. The same multiple ML algorithms could be used to estimate a model for predicting MCI in middle-aged and older adults. In addition, the time of conversion could also be predicted by time Survival Analysis algorithms (i.e., Kaplan Meier Curve, Log Rank Test and Cox Regression), and time series model of DL (i.e., Recurrent Neural Network).

DATA AVAILABILITY STATEMENT

The original data used in this manuscript is available for download in the ADNI.
database (https://adni.loni.usc.edu).

ACKNOWLEDGMENTS

Data collection and sharing were supported by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is supported by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following entities: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE.

Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

FUNDING

This work was partially supported by the project "Innovation Ecosystem: Innovation, digitalisation and sustainability for the diffused economy in Central Italy (Vitality)" funded from the European Union - NextGenerationEU under the National Recovery and Resilience Plan (NRRP), Mission 4, Component 2.

AUTHOR CONTRIBUTIONS

R.F. produced the study's design, supervised data analysis, contributed to data interpretation, and wrote the manuscript. D.N. conducted data analysis and contributed to manuscript writing and reviewing. M.R. contributed to data interpretation and manuscript writing and reviewing. M.O. contributed to critical revision. S.L. S. contributed to data interpretation, writing, reviewing, and final manuscript preparation.

COMPETING INTERESTS

R.F. declares no Competing Financial Interests. She has the following Competing Non-Financial Interest: she serves as associate editor of Frontiers in Human Neuroscience. D.N. and M.R. declare no Competing Financial or Non-Financial Interests. M.O. declares no Competing Non-Financial Interests, but has the following Competing Financial Interests: he served on the scientific advisory boards of GlaxoSmithKline, Novartis, Lundbeck, Eisai, Valeant, Medtronic, and Newron. He received speaker honoraria from Zambon, the World Parkinson Congress, the Movement Disorder Society, and the Atypical Dementias congress. He was also a speaker for Boehringer Ingelheim, GlaxoSmithKline, UCB, and Zambon. He received publishing royalties from Springer. S.L.S. declares the following Competing Non-Financial Interests: he serves as associate editor of Frontiers in Neuroscience, Frontiers in Psychiatry, PlosOne, and Scientific Reports. He reports the following Competing Financial Interests: he is supported by non-profit agencies (the Italian Ministry of Health, the AIRAlzh Onlus [ANCC-COOP], the Alzheimer's Association – Part the Cloud: Translational Research Funding for Alzheimer's Disease [18PTC-19-602325] and the Alzheimer's Association-GAAIN Exploration to Evaluate Novel Alzheimer's Queries [GEENA-Q-19-596,282]).

REFERENCES

Albert MS, DeKosky ST, Dickson D, Dubois B, Feldman HH, Fox NC, Gamst A, Holtzman DM, Jagust WJ, Petersen RC, Snyder PJ, Carrillo MC, Thies B, Phelps CH (2011) 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.

Apostolova LG, Dutton RA, Dinov ID, Hayashi KM, Toga AW, Cummings JL, Thompson PM (2006) Conversion of Mild Cognitive Impairment to Alzheimer Disease Predicted by Hippocampal Atrophy Maps. Arch Neurol 63:693–699.

Bergstra J, Bengio Y (2012) Random search for hyper-parameter optimization. J Mach Learn Res 13:281–305.

Breiman L (2001) Random Forests. Mach Learn 45:5–32.

Brem AK, Sensi SL (2018) Towards Combinatorial Approaches for Preserving Cognitive Fitness in Aging. Trends Neurosci 41:885–897.

Brinkmalm A, Brinkmalm G, Honer WG, Fr€olich L, Hausner L, Minthon L, Hansson O, Wallin A, Zetterberg H, Blennow K, Öhrfelt A (2014) SNAP-25 is a promising novel cerebrospinal fluid

biomarker for synapse degeneration in Alzheimer's disease. Mol Neurodegener 9:53.
Casaletto KB, Elahi FM, Bettcher BM, Neuhaus J, Bendlin BB, Asthana S, et al. (2017) Neurogranin, a synaptic protein, is associated with memory independent of Alzheimer biomarkers. Neurology 89:1782–1788.
Chen RC, Dewi C, Huang SW, Caraka RE (2020) Selecting critical features for data classification based on machine learning methods. J Big Data 7:1–26.
Chen T, Guestrin C (2016) XGBoost. In: in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 785–794 (ACM). https://doi.org/10.1145/2939672.2939785.
Cox TJ, Li F, Darlington P (2001) Extracting room reverberation time from speech using artificial neural networks. J audio Eng. Soc 49:219–230.
Delli Pizzi S, Punzi M, Sensi SL, Initiative ADN (2019) Functional signature of conversion of patients with mild cognitive impairment. Neurobiol Aging 74:21–37.
Desikan RS, Ségonne F, Fischl B, Quinn BT, Dickerson BC, Blacker D, Buckner RL, Dale AM, Maguire R, Hyman BT, Albert MS, Killiany RJ (2006) An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. Neuroimage 31:968–980.
Dinu AJ, Ganesan R (2019) Early detection of Alzheimer's disease using predictive k-NN instance based approach and T-Test Method. Int J Adv Trends Comput Sci Eng 8:29–37.
Elgeldawi E, Sayed A, Galal AR, Zaki AM (2021) Hyperparameter Tuning for Machine Learning Algorithms Used for Arabic Sentiment Analysis. Informatics 8.
Faouri S, AlBashayreh M, Azzeh M (2022) Examining stability of machine learning methods for predicting dementia at early phases of the disease. Decis Sci Lett 333–346. https://doi.org/10.5267/j.dsl.2022.1.005.
Fernández-Delgado M, Cernadas E, Barro S, Amorim D (2014) Do we need hundreds of classifiers to solve real-world classification problems?. J Mach Learn Res 15:3133–3181.
Friedman JH (2001) Greedy function approximation: a gradient boosting machine. Ann Stat:1189–1232.
Galasko D, Xiao M, Xu D, Smirnov D, Salmon DP, Dewit N, Vanbrabant J, Jacobs D, Vanderstichele H, Vanmechelen E, Initiative ADN, (ADNI), Worley P (2019) Synaptic biomarkers in CSF aid in diagnosis, correlate with cognition and predict progression in MCI and Alzheimer's disease. Alzheimers Dement (N Y) 5:871–882.
Janiesch C, Zschech P, Heinrich K (2021) Machine learning and deep learning. Electronic Markets 31:685–695.
Knopman DS, Petersen RC (2014) Mild Cognitive Impairment and Mild Dementia: A Clinical Perspective. Mayo Clin Proc 89:1452–1459.
Lundberg SM, Lee SI (2017) A unified approach to interpreting model predictions. Adv Neural Inf Process Syst 30.
Massetti N, Russo M, Franciotti R, Nardini D, Mandolini GM, Granzotto A, Bomba M, Delli Pizzi S, Mosca A, Scherer R, Onofrj M, Sensi SL, Alzheimer's Disease Neuroimaging Initiative (ADNI); Alzheimer's Disease Metabolomics Consortium (ADMC) (2022) A Machine Learning-Based Holistic Approach to Predict the Clinical Course of Patients within the Alzheimer's Disease Spectrum. J Alzheimer's Dis 85:1639–1655.
Mazaheri A, Segaert K, Olichney J, Yang JC, Niu YQ, Shapiro K, Bowman H (2018) EEG oscillations during word processing predict MCI conversion to Alzheimer's disease. NeuroImage: Clin 17:188–197.
Mishra S, Mallick PK, Tripathy HK, Jena L, Chae GS (2021) Stacked KNN with hard voting predictive approach to assist the hiring process in IT organizations. Int J Electr Eng Educ 0020720921989015.
Natras R, Soja B, Schmidt M (2022) Ensemble Machine Learning of Random Forest, AdaBoost and XGBoost for Vertical Total Electron Content Forecasting. Remote Sens 14:3547.
Petersen RC, Caracciolo B, Brayne C, Gauthier S, Jelic V, Fratiglioni L (2014) Mild cognitive impairment: a concept in evolution. J Intern Med 275:214–228.
Poil SS, de Haan W, van der Flier WM, Mansvelder HD, Scheltens P, Linkenkaer-Hansen K (2013) Integrative EEG biomarkers predict progression to Alzheimer's disease at the MCI stage. Front Aging Neurosci 5:58.
Poulin SP, Dautoff R, Morris JC, Barrett LF, Dickerson BC, Initiative ADN (2011) Amygdala atrophy is prominent in early Alzheimer's disease and relates to symptom severity. Psychiatry Res Neuroimaging 194:7–13.
Qi Y (2012) Random Forest for Bioinformatics. In: Ensemble Machine Learning. US: Springer. p. 307–323. https://doi.org/10.1007/978-1-4419-9326-7_11.
Rohini M, Surendran D (2021) Toward Alzheimer's disease classification through machine learning. Soft Comput 25:2589–2597.
Rossini PM, Miraglia F, Vecchio F (2022) Early dementia diagnosis, MCI-to-dementia risk prediction, and the role of machine learningmethods for feature extraction from integrated biomarkers, in particular for EEG signal analysis. Alzheimer's Dement 18:2699–2706.
Schapire RE, Freund Y (2013) Boosting: Foundations and algorithms. Kybernetes..
Shree RB, Sheshadri HS (2018) Diagnosis of Alzheimer's disease using Naive Bayesian Classifiers. Neural Comput & Applic 29:123–132.
Spires-Jones TL, Hyman BT (2014) The intersection of amyloid beta and tau at synapses in Alzheimer's disease. Neuron 82:756–771.
Syaifullah AH, Shiino A, Kitahara H, Ito R, Ishida M, Tanigaki K (2021) Machine Learning for Diagnosis of AD and Prediction of MCI Progression From Brain MRI Using Brain Anatomical Analysis Using Diffeomorphic Deformation. Front Neurol 11. https://doi.org/10.3389/fneur.2020.576029 576029.
Teng L, Li Y, Zhao Y, Hu T, Zhang Z, Yao Z, Hu B, Alzheimer' s Disease Neuroimaging Initiative (ADNI), (2020) Predicting MCI progression with FDG-PET and cognitive scores: a longitudinal study. BMC Neurol 20:148.
Varoquaux G, Cheplygina V (2022) Machine learning for medical imaging: methodological failures and recommendations for the future. npj Digit Med 5:48.
Vozella V, Basit A, Piras F, Realini N, Armirotti A, Bossù P, Assogna F, Sensi SL, Spalletta G, Piomelli D (2019) Elevated plasma ceramide levels in post-menopausal women: a crosssectional study. Aging (Albany NY) 11:73–88.
Williams TL, Serpell LC (2011) Membrane and surface interactions of the Alzheimer's A! peptide: Insights into the mechanism of cytotoxicity. FEBS J 278:3905–3917.

(Received 25 September 2022, Accepted 24 January 2023)
(Available online 2 February 2023)