Journal of Alzheimer's Disease xx (2022) x–xx
DOI 10.3233/JAD-215164
IOS Press

Discriminating Aging Cognitive Decline Spectrum Using PET and Magnetic Resonance Image Features

Caroline Machado Dartora[a,*,1], Luiś Vinicius de Moura[b], Michel Koole[c],
Ana Maria Marques da Silva[a,b,d] and for the Alzheimer's Disease Neuroimaging Initiative[2]
[a]*PUCRS, School of Medicine, Porto Alegre, Brazil*
[b]*PUCRS, School of Technology, Porto Alegre, Brazil*
[c]*KU Leuven, Nuclear Medicine and Molecular Imaging, Department of Imaging and Pathology, Medical Imaging Research Center, Leuven, Belgium*
[d]*PUCRS, Brain Institute of Rio Grande do Sul (BraIns), Porto Alegre, Brazil*

Handling Associate Editor: Anette Hall

Accepted 18 July 2022
Pre-press 16 August 2022

Abstract.

Background: The population aging increased the prevalence of brain diseases, like Alzheimer's disease (AD), and early identification of individuals with higher odds of cognitive decline is essential to maintain quality of life. Imaging evaluation of individuals at risk of cognitive decline includes biomarkers extracted from brain positron emission tomography (PET) and structural magnetic resonance imaging (MRI).

Objective: We propose investigating ensemble models to classify groups in the aging cognitive decline spectrum by combining features extracted from single imaging modalities and combinations of imaging modalities (FDG+AMY+MRI, and a PET ensemble).

Methods: We group imaging data of 131 individuals into four classes related to the individuals' cognitive assessment in baseline and follow-up: stable cognitive non-impaired; individuals converting to mild cognitive impairment (MCI) syndrome; stable MCI; and Alzheimer's clinical syndrome. We assess the performance of four algorithms using leave-one-out cross-validation: decision tree classifier, random forest (RF), light gradient boosting machine (LGBM), and categorical boosting (CAT). The performance analysis of models is evaluated using balanced accuracy before and after using Shapley Additive exPlanations with recursive feature elimination (SHAP-RFECV) method.

Results: Our results show that feature selection with CAT or RF algorithms have the best overall performance in discriminating early cognitive decline spectrum mainly using MRI imaging features.

Conclusion: Use of CAT or RF algorithms with SHAP-RFECV shows good discrimination of early stages of aging cognitive decline, mainly using MRI image features. Further work is required to analyze the impact of selected brain regions and their correlation with cognitive decline spectrum.

Keywords: Aging, amyloid, atrophy, fluorodeoxyglucose F18, machine learning, multimodal imaging

[*] Correspondence to: Caroline Machado Dartora, Av. Ipiranga 6681, Prédio 96A, Sala 220, Porto Alegre, Brazil. E-mail: caroline.dartora@acad.pucrs.br.

[1] Present address: Division of Clinical Geriatrics, Center for Alzheimer Research, Department of Neurobiology, Care Sciences and Society, Karolinska Institutet, Stockholm, Sweden.

[2] Data used in 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 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

INTRODUCTION

Aging is a complex process that evolves deleterious changes in molecular and morphological levels leading to cognitive decline and increased risk of diseases and death. The population aging increases the prevalence of age-related brain diseases and syndromes, like dementia [1]. The main cause of dementia in the elderly population worldwide is Alzheimer's disease (AD), a multifactorial progressive and irreversible neurodegenerative disease [2].

AD was first defined as a clinical-pathologic entity based on clinical history, neurological examinations, cognitive testing, and neuroimaging [3], with definitive diagnosis by autopsy [4]. In 2011, the National Institute on Aging and Alzheimer's Association created separate diagnostic recommendations for the preclinical, mild cognitive impairment (MCI), and dementia stages of AD. The definition of AD in living people is biologically identified by an ensemble of neuropathological changes, like amyloid-β (Aβ) and tau in abnormal levels, determined by *in vivo* biomarkers and postmortem evaluation without considering the clinical symptoms in a research framework. In clinical practice, clinical symptoms are still the main diagnosis of dementia. However, in the absence of clear threshold values to define abnormal levels of Aβ and tau, clinical-pathological evaluation is still used, dividing the cognitive continuum into three traditional categories, healthy cognitive non-impaired individuals (CNI), MCI, and dementia, with dementia further subdivided into mild, moderate, and severe stages [4]. Neuropathological AD changes begin several decades before cognitive impairment. Drugs can temporarily relieve symptoms but do not stop or slow down the pathological damage, leading to the idea that preventive and treatments may be more effective in the early phases [1, 5].

Several neuroimaging modalities have been used to investigate, diagnose, and predict early dementia. Magnetic resonance imaging (MRI) identifies neuronal/synapse loss and atrophy. Positron emission tomography (PET) using ^{18}F-fluorodeoxyglucose (FDG PET) enables glucose metabolism assessment, and amyloid-β tracers quantify protein burden (AMY PET). The combination of neuroimaging and artificial intelligence techniques, like machine learning (ML), has been increasing in the last years, aiming to predict dementia development and classify individuals based on image features and neuropsychological test scores. The neuroimaging technique more present in the literature associated with ML methods is the MRI, followed by PET images, achieving mean classification accuracies of 74.5%, for MRI alone, 76.9% for PET images, and 77.5% when combined both modalities [6]. Despite recent developments in classification and prediction models in cognitive decline progression using image features, current literature focuses on comparing CNI versus MCI, MCI versus AD, and CNI versus AD [2, 7–15]. Investigating early conversion using image features is still challenging and requires further investigation.

In this study, we propose to investigate tree-based ensemble models to classify individuals in the cognitive decline spectrum by using features extracted from single imaging modalities (FDG PET, AMY PET, and MRI) and combinations of imaging modalities (FDG PET+AMY PET+MRI, and a PET ensemble) to verify which combination of features and algorithm performs better. We evaluate the performance of four algorithms before and after feature selection using Shapley Additive Explanations with recursive feature evaluation and cross-validation (SHAP-RFECV) to classify four groups: stable CNI, healthy individuals who just ended up with MCI referred to as converters (CONV), stable MCI, and those with Alzheimer's clinical syndrome (ACS). Our results showed that combining SHAP-RFECV with the categorical boosting, and the random forest algorithms showed good performance discriminating early cognitive decline. Features extracted from MRI achieve higher accuracy in the discrimination of CNI from all other groups. The classification using the multimodal combination of all images achieves higher accuracies than the PET ensemble.

MATERIALS AND METHODS

Image dataset

We use FDG PET, AMY PET (acquired with ^{11}C-PiB or ^{18}F-AV45), and structural T1-weighted MRI retrieved from the Alzheimer's Disease Neuroimaging Initiative (ADNI, http://adni.loni.usc.edu) database to train and evaluate our models. ADNI was launched in 2003 as a public-private partnership led by Principal Investigator Michael W. Weiner, MD. Inclusion and exclusion ADNI criteria can be found in their general procedure manual (http://adni.loni.usc.edu/methods/documents/). FDG PET and MRI were acquired on the same day, while AMY PET was acquired on different days or visits. PET and MRI acquisition protocols can be found on the ADNI website.

For our study, data from individuals are grouped into four classes (CNI, CONV, MCI, and ACS) related to their cognitive assessment in the baseline and follow-up, using the criteria described in the following paragraphs.

CNI individuals have no memory complaints, normal memory function documented by scoring at specific cutoffs described in ADNI protocol. In addition, our sample remains cognitively healthy for more than 5 years in the follow-up.

CONV individuals are characterized as CNI in the baseline, converting to MCI in the follow-up years, based on their cognitive scores according to ADNI protocol. Image inclusion criteria include images that were acquired between six months before conversion to MCI and one year after conversion to avoid fluctuations with subjects that are stable in their diagnosis as CNI or MCI.

MCI are patients with memory complaints and abnormal memory function documented by scoring below the adjusted education cutoff described in the ADNI protocol. Our MCI individuals are stable for at least 5 years follow-up.

ADNI protocol classified ACS individuals as "probable AD" because they have memory complaints, abnormal memory function, and NINCDS/ADRDA (National Institute of Neurological and Communicative Diseases and Stroke/ Alzheimer Disease and Related Disorders Association) criteria for probable AD.

All stable individuals (CNI, MCI, and ACS) were randomly chosen in the ADNI dataset if they attended the inclusion criteria of at least 5 years of stability in their diagnosis.

Table 1 shows the number of individuals in our sample, with all three imaging modalities (FDG PET, AMY PET, and MRI) and those with only FDG PET and MRI and demographic information.

MRI was acquired on the same day as FDG PET images. MRI acquired on the same day of AMY PET images was used for processing purposes but was not included in the analysis. Individuals with images of three modalities were the same as those included in the only FDG PET and MR images.

We checked each PET to assure scattering and attenuation correction. We selected only MRI acquired on the same day or the nearest date to PET. Image quality was visually inspected after download. Images with poor quality, missing brain parts (usually the cerebellum), and non-standardized PET time frames (for FDG PET 6 frames or 30 min, and AMY PET 4 frames or 20 min) were excluded.

There is a statistically significant difference between age, demonstrated by one-way ANOVA (for FDG PET/MRI $F = 17.451$, $p < 0.05$; for AMY PET $F = 13.049$; $p < 0.05$). Tukey's post hoc test showed that CNI and CONV are statistically older than MCI and ACS ($p < 0.05$) in FDG/MRI. There is no significant difference between CNI and CONV ($p > 0.05$).

There is a slight gender difference, with $\chi^2 = 7.711$, $p = 0.052$, for FDG PET/MRI, primarily due to the small number of females in the CONV group. For AMY PET, the χ^2 test does not show a significant statistical difference between gender in CNI, CONV, MCI, and ACS ($\chi^2 = 6.609$, $p = 0.085$).

According to the one-way ANOVA, there was no statistically significant difference between groups in years of education (for FDG PET/MRI $F = 0.385$, $p = 0.764$; for AMY PET $F = 1.958$; $p = 0.125$).

Image preprocessing

We processed all images in a pipeline using PMOD® (https://www.pmod.com/web/) version 4.0 and SPM12 (https://www.fil.ion.ucl.ac.uk/spm/software/spm12/) software. Pixel interpolation (1 mm^3) is applied in all images before processing to harmonize the data extracted from different matrix sizes. A flowchart overview of the applied methodology used in this work is presented in Supplementary Figure 1.

PET processing

Initially, motion correction is applied using normalized mutual information in PMOD®, with the

Table 1
Demographics

	Sample size		Age (y)		Gender (M/F)		Education (y)	
Group/ Modality	All modalities	FDG PET/MRI	FDG PET/MRI	AMY PET	FDG PET/MRI	AMY PET	FDG PET/MRI	AMY PET
CNI	22	36	79.6 ± 5.5	80.5 ± 4.4	18/18	11/11	16.0 ± 3.6	17.3 ± 2.6
CONV	16	24	81.7 ± 4.4	81.8 ± 4.9	19/5	13/3	16.4 ± 3.2	16.1 ± 3.4
MCI	40	40	71.6 ± 6.8	71.8 ± 7.1	19/21	19/21	16.1 ± 2.5	16.1 ± 2.5
ACS	29	31	73.3 ± 8.3	75.6 ± 7.9	20/11	19/10	15.54 ± 2.79	15.4 ± 2.7

first frame (5 min) as reference. Then, the average PET image is calculated in the last 15 min for FDG PET and the last 20 min for AMY PET.

In SPM12, the image origin is manually positioned in the anterior commissure-posterior commissure brain line. PET and MRI co-registration is made with trilinear interpolation. Individual MRI segmentation of white matter (WM), gray matter (GM), and cerebrospinal fluid (CSF) are realized in the MNI space. Subsequently, PET is normalized to the MNI space. Finally, a whole-brain mask based on WM, GM, and CSF MRI segmentation is applied to the PET image smoothed with a gaussian filter of 8 mm kernel. In the end, all PET images have 91 x 109 x 91 pixels, with a 2 mm isotropic voxel size.

MRI processing

MRI is processed using the Computational Anatomy Toolbox (CAT, http://www.neuro.uni-jena.de/cat/) for volume estimation in the GM brain regions after cropping to remove extra tissues, as the neck and shoulders. Images are initially denoised with a spatial adaptive non-local means denoising filter, bias-corrected, affine-registered to template space, and segmented in GM, WM, and CSF. Then, a skull-stripping is realized, and brain parcellation in right and left hemispheres, subcortical areas, and the cerebellum. Subsequently, a local intensity transformation of all tissue classes and adaptive maximum a posteriori (AMAP) segmentation is performed. Finally, the AMAP segmentation is refined by applying partial volume correction, and tissues are spatially normalized to a common reference space using DARTEL (Diffeomorphic Anatomical Registration Through Exponentiated Lie Algebra). Further details can be found in the CAT12 toolbox Manual (http://www.neuro.uni-jena.de/cat12/CAT12-Manual.pdf). In the end, all MR images have 91 x 109 x 91 pixels, with a 2.0 mm isotropic voxel size, and are smoothed with a gaussian filter of 6.0 mm kernel.

Classification algorithms

We evaluate the performance of four classification models using scikit-learn [16], LightGBM [17], and CatBoost [18] libraries, with Python version 3.6.5. The classifier algorithms are ensemble and tree-based and have an increased level of complexity, described in the following sub-sections. These algorithms were chosen based on the applicability of SHapley Additive exPlanations with recursive feature elimination (SHAP-RFECV, described on section "Feature Selection") method, which allows interpretability of the selected features, and because they are powerful tools that have been used to provide easy-to-interpret predictive results based on decisions trees.

Decision tree classifier

A decision tree classifier (DTC) is a non-parametric supervised learning method that produces a classification model by splitting data using simple decisional rules. It is extensively applied in many pattern recognition problems such as remotely sensed multisource data classification, medical diagnosis, speech, and character recognition. Some issues are created using DTC, as pointed out by Safavian and Landgrebe [19]. However, a truly optimal solution concerning the choice of the decision tree structure, feature subsets, and decision rule strategies is yet far from realization [19, 20]. Our study uses the classification implemented in scikit-learn (https://scikit-learn.org/stable/modules/tree.html#tree) with the best split strategy, optimizing the criterion for information gain between Gini impurity and entropy and the maximum number of features for the best split.

Random forest

Random forest (RF) is a classifier that aims to avoid overfitting mainly by adding two sources of randomness in the training stage. The first source is that each tree in the forest is made from a sample of the original training data. The second one is that when splitting a tree node, the algorithm uses only a random subset of all the features. After training all the trees, the model chose the prediction based on the most selected features or average prediction probabilities [21]. We use the scikit-learn implementation of RF, using the average prediction probabilities approach. We maintain the maximum number of features to consider when seeking for best split set as automatic. The parameters used for RF optimization are the number of estimators, the criterion (Gini impurity or entropy), the need for bootstrap, and where to use out-of-bag samples to estimate the generalization score.

Light gradient boosting machine

Light gradient boosting machine (LGBM) is an ensemble model of decision trees aiming to reduce the complexity of histogram building by reducing the data. Two main techniques are used and have more efficiency and less memory usage. The first one is the gradient-based one side sampling technique, which uses only the instances with the most significant gradients to maximize the information gain and randomly drop the instances with small gradients. Thus, the technique reduces the dimensionality in the dataset and then reduces the training and prediction time. The second technique uses exclusive feature bundling to reduce the problem's dimensionality using graphs and solve the problem with a constant approximation ratio [17]. Our study uses the gradient boosting decision tree and binary learning task with the following hyperparameters: the number of estimators, the number of leaves, minimum child weight, and samples.

Categorical boosting

Categorical boosting (CAT), also known as CatBoost, is a gradient boosting algorithm that handles categorical features during the training phase, different from others that need to be addressed during the preprocessing step. Although CAT is designed mainly to deal with categorical features, it is possible to run over a dataset with continuous features. The primary motivation of CAT is to avoid the prediction shift of traditional gradient boosting models. Instead, it uses ordered boosting, which creates a given number of sub-datasets based on the permutation of the original data to train the model. CAT also differs in the use of oblivion trees with a more robust regularization due to the restriction in the building processes and better computational performance due to limitations in the feature's splits per tree level [18]. In our study, we used the CAT as an ordered gradient boosting on decisions trees with loss function, learning rate, bagging aggressivity for Bayesian bootstrap, the coefficient at the L2 regularization term of the cost function, depth of the tree, overfitting detector type, and threshold as parameters for model tunning.

Feature extraction

Imaging features are vectorized, with rows representing the individuals, and columns the imaging features extracted from the following brain regions: amygdala, brainstem, caudate nucleus, cerebellum, cingulate gyri, corpus callosum, frontal lobe, hippocampus, insula, nucleus accumbens, occipital lobe, occipital lobe cuneus, pallidum, parietal lobe, putamen, temporal lobe, thalamus, and ventricles.

PET imaging features are composed of the mean uptake of the previous brain regions normalized by the ratio between each voxel and the whole-brain mean uptake, extracted from Hammers N30R83 atlas [22] overlapped in PET using an in-house MATLAB code to produce a brain region-based analysis. The normalization avoids the variability of PET images acquired in different institutions or equipment.

MR imaging features are the volumes of the previous brain regions normalized by the total intracranial volume using the Hammers N20R67 atlas [6].

Feature selection

We use Shapley additive explanations (SHAP) combined with the recursive feature elimination with cross-validation (RFECV) for imaging feature selection.

SHAP is an additive feature attribution method based on the Shapley values from the game theory that assigns an "importance value" for each feature for a particular prediction. The method calculates the contribution of each feature individually, allowing comparison between different models and analyzing the feature influence against the feature value. Unlike other explainable methods, SHAP perturbs all subsets of features, dealing with the interaction between features [16, 17].

The RFECV is a dimensionality reduction algorithm that recursively constructs the model, chooses the least important variable, removes the feature with the lowest importance until the desired number of features or the set of features gives the best performance. RFECV method uses the impurity index (Gini impurity) for tree-based models to select features, handling with nonlinear relation between features [18, 19]. However, the impurity shows only the features' frequency and magnitude in the tree-based model and not its importance. Thus, features with atypical values have more chance to be considered the most important feature, increasing bias in the selection. In our work, we used the combination of SHAP and RFECV to avoid bias in feature selection.

The feature selection uses 10-folds cross-validation, eliminating 10% of image features with the smallest SHAP values in each fold. We use the set of features that achieves the highest area under the curve (AUC) of the receiver operating characteristic (ROC) curve in a training dataset with 80% of the whole dataset after the 10-folds cross-validation.

Evaluation strategy

The algorithms presented in section "Classification algorithms" are evaluated before and after feature selection. They are tuned and evaluated with the best parameters. More details are presented in the following sub-sections:

Hyperparameter tuning

Each model was tuned using a randomized search with cross-validation from the sci - kit learn library to optimize the classification. The method uses a range of values of set parameters randomly to optimize the model seeking the parameters that give the highest sensitivity between all tested parameters combinations. We chose to run 100 iterations for each model, using leave-one-out cross-validation (LOOCV).

Performance metrics

The performance metrics used for the classifiers' comparison are the balanced accuracy, accuracy, and the number of selected features. We chose to mainly use balanced accuracy to compare our results due to its joint representation of sensitivity and specificity than accuracy itself. Accuracy is only used to compare our results with the literature.

The balanced accuracy of a model is calculated as follows:

$$Balanced\ accuracy\ \ (Bacc)$$

$$= \frac{1}{2}\,(sensitivity + specificity)$$

$$= \frac{1}{2}\left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP}\right),$$

where TP = true positive, TN = true negative, FP = false positive, and FN = false negative.

Interpretation of selected features

Interpretation of selected features by the SHAP-RFECV model is obtained with the SHAP interpreter trained in 80% of data, with hyperparameter tuning with randomized search strategy with 10-folds cross-validation seeking for the highest area under the curve of the receiving operating curve. To an unbiased interpretation of the selected features, the trained model is evaluated in the test dataset (20%).

Classification experiments

We investigate three binary problems to classify individuals in the cognitive decline spectrum: CNI versus CONV, CNI versus MCI, and CNI versus ACS. For each binary task, we tested features extracted from FDG PET, AMY PET, and MRI modalities separately, a multimodality approach using features extracted from all images, and features extracted a combination of both FDG and AMY PET images. All imaging features are concatenated in a vector for the same individual.

We evaluate the performance of four classification models (DTC, RF, LGBM, and CAT) before and after the feature selection using SHAP-RFECV. Additionally, we perform a randomized search with LOOCV for hyperparameter tuning in the models before and after feature selection.

RESULTS

Results reveal feature selection using SHAP-RFECV method improved the balanced accuracy of the classification models. However, exceptions occurred mainly for DTC and LGBM algorithms. The highest balanced accuracy difference between before and after feature selection was 26%.

Figure 1 shows the number of features selected by imaging modality for each pairwise comparison using the combination of SHAP and RFECV.

Figures 2–4 show the balanced accuracy, confidence interval values, and the p-value of the two groups non-parametric Wilcoxon test for paired data for the classification models before feature selection (DTC-1, RF-1, LGBM-1, CAT-1) and after feature selection (DTC-2, RF-2, LGBM-2, CAT-2) for each binary classification task (CNI versus CONV, CNI versus MCI, and CNI versus ACS).

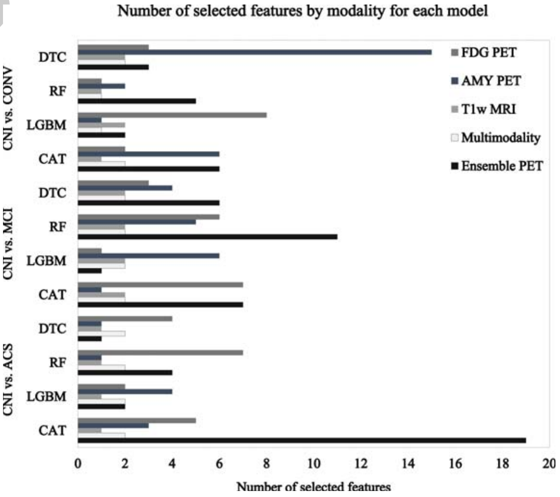

Fig. 1. Number of selected features for each pairwise comparison in single and multimodality imaging approaches for all classification models.

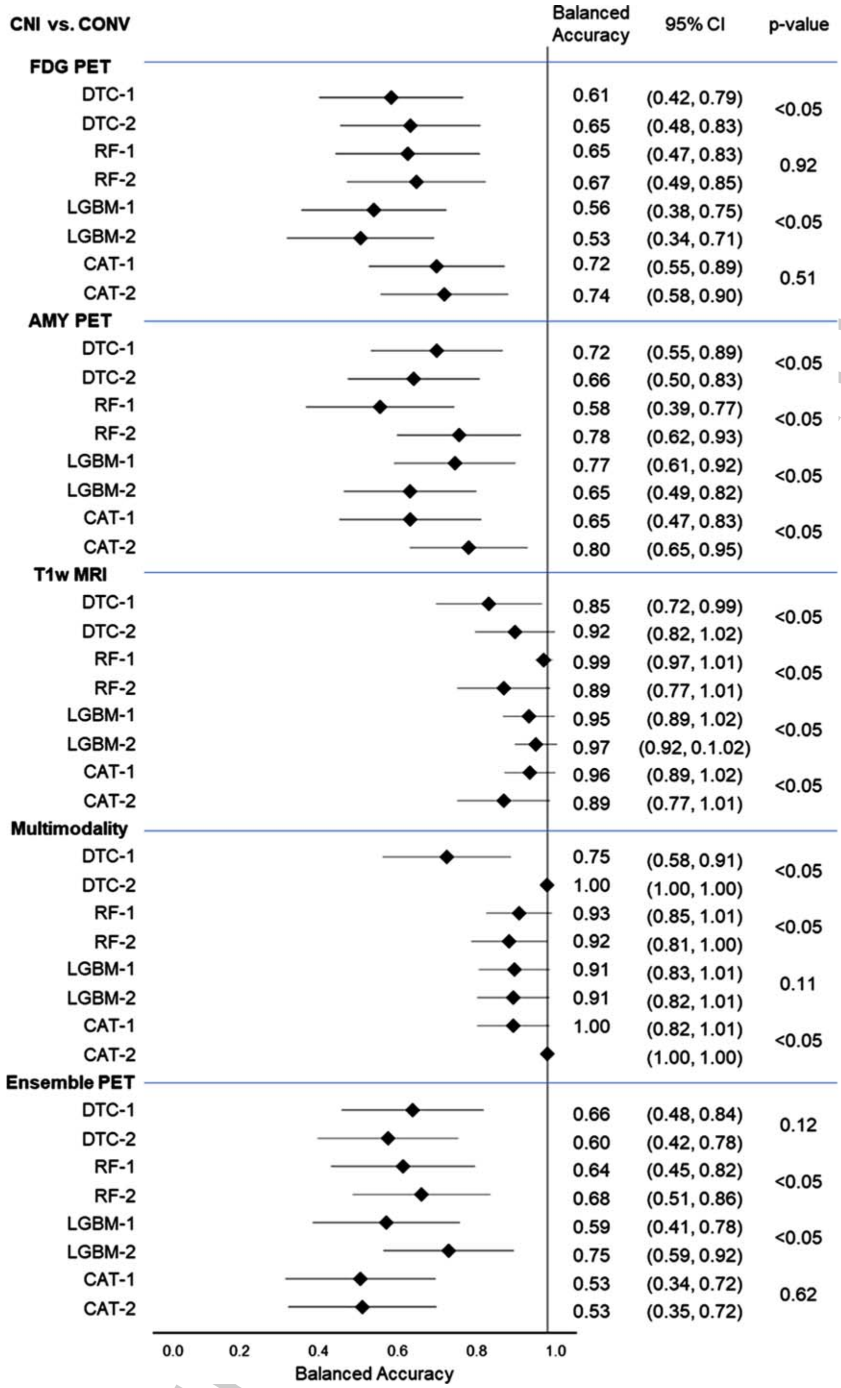

Fig. 2. Balanced accuracy with variance, and 95% confidence interval (CI) for each classification model before feature selection (Model-1) and after feature selection (Model-2), for the binary classification task CNI versus CONV.

DISCUSSION

This study investigates ensemble with tree-based algorithms to classify individuals in the cognitive decline spectrum by using features extracted from single imaging modalities (FDG PET, AMY PET, and MRI) and combinations of imaging modalities (FDG PET+AMY PET+MRI, and a PET ensemble). We study the effect of feature selection in the classification of healthy cognitive non-impaired

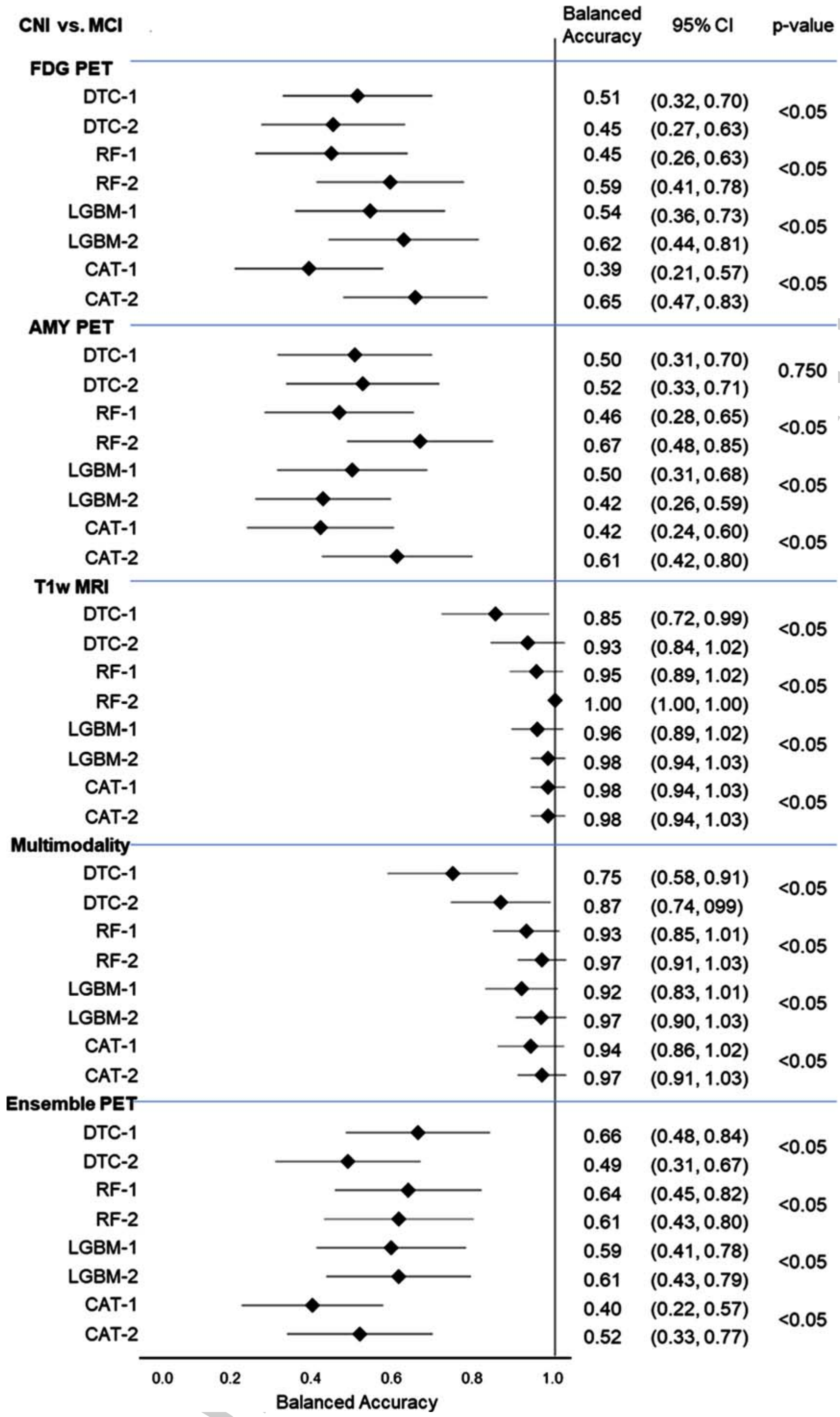

Fig. 3. Balanced accuracy with variance, and 95% confidence interval (CI) for each classification model before feature selection (Model-1) and after feature selection (Model-2), for the binary classification task CNI versus MCI.

individuals (CNI) in a pairwise comparison with converters (CONV), MCI, and ACS.

Estimating the features' importance for classification in neuroimaging is valuable because it allows assessing the features contributing to the classifier. It can potentially identify, for example, regions or structures with a biologically plausible connection to the pathology. The feature selection is particularly inter-

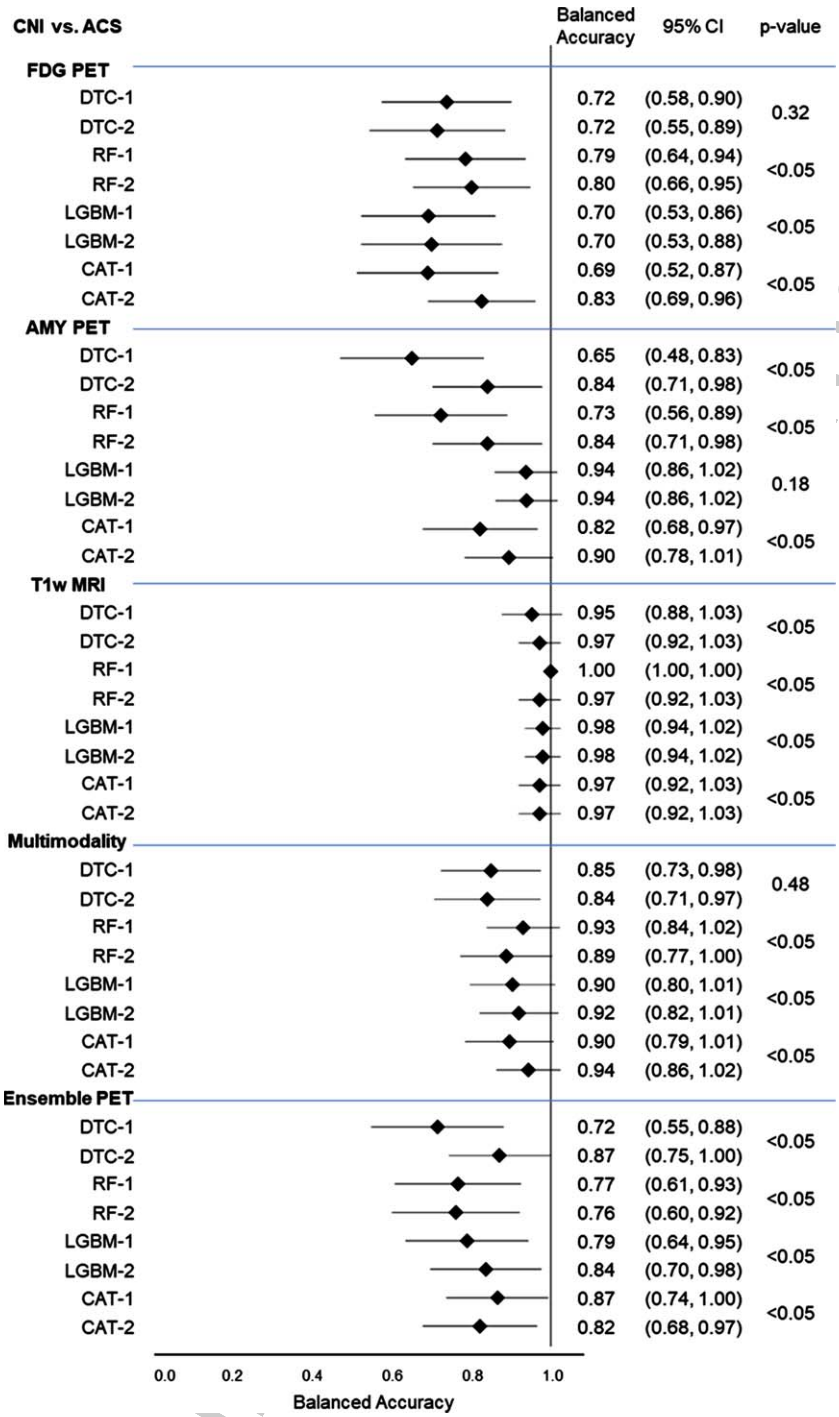

Fig. 4. Balanced accuracy with variance, and 95% confidence interval (CI) for each classification model before feature selection (Model-1) and after feature selection (Model-2), for the binary classification task CNI versus ACS.

esting in studying cognitive decline using imaging features to connect the disease evolution and radiomic features.

Several methods and algorithms are already implemented to select features in ML models based on univariate group-level statistical tests, filtering, and wrapper methods, like SHAP-RFECV, used in our study. Each method has its particularities, advantages, and disadvantages. Feature reduction methods are excellent and usually provide higher accuracies because they use all the variance of feature information in a small feature space, like the principal component analysis. However, the information about the importance of each feature is lost in the process. Statistic-based features have the advantage of being independent of model performance. However, they are sensitive to the group mean, leading to the loss of discriminatory information due to exclusion [23]. Like Pearson's correlation, filtering methods are independent of the algorithm performance, but most methods treat the features independently, ignoring their relationships [23]. Wrapper methods consider the feature selection as a search problem and eliminate features based on features weights assigned by the best performance on an external estimator. SHAP feature importance, used in our study, is a way to get each feature influence in the prediction, even more for a tree-based model, due to the lack of information when using only the impurity as a measurement for the feature importance.

Our sample size in the groups varied from 16 to 40 subjects, a small number compared to the number of imaging features. In some cases, a ratio of a sample size to features was almost 1 : 1 (i.e., CNI versus CONV, with 38 subjects for 36 image features in the PET ensemble approach). According to Vabalas et al. [24], if the ratio of features to sample size is high, the classification model tends to fit the noise of data instead of the underlying pattern and overfitting. Our results showed an overall improvement in the classification models' balanced accuracy with the feature selection. The SHAP-RFECV ensures to avoid bias in feature selection, and its use shows to reduce problems of fitting to noise [25].

Our results show that the features extracted from the MRI approach produce the highest performance for all models in all binary classification tasks. Our MRI features are the mean volume of cortical GM brain regions normalized by the estimated intracranial volume based on Hammers' atlas. Measures of cortical thickness and subcortical volumes are the most used biomarkers related to structural neurodegeneration in AD and cognitive decline [8]. For the four different algorithms, one MRI imaging feature was consistently selected in all binary classification tasks: the parietal lobe (Supplementary Table 1). The parietal lobe comprises the precuneus and regions of the somatosensory and visuospatial cortex and is involved in higher cognitive functions [28]. Previous works showed the volume of parietal structures is predictive of conversion from MCI to AD [20, 21]. In our sample, CNI individuals presents higher parietal volumes than the other three groups (data not shown) being possible to verify that this region could be used as an early marker of neurodegeneration, considering that the CONV group is in the same age group as the HC, and that MCI, and AD groups are about 10 years younger. Following the literature, the MRI feature selected together with the parietal lobe in the binary tasks (CNI versus CONV and CNI versus MCI) was the frontal lobe, which plays a part in monitoring and controlling processes that support memory [29], language, and visuoconstructive abilities [30]. Moreover, the frontal theory of cognitive aging suggests that the frontal lobe is responsible for the decline in memory, attention, and cognitive flexibility that accompany healthy aging [31], supporting our results. In our sample, frontal lobe of COVN, MCI, and ACS groups overlap themselves, while CNI individuals presents smaller volume compared to them (data not shown). It is important to note that CNI and CONV groups are about 10 years older than MCI and ACS groups, and smaller volumes of this region is expected even in non-impaired individuals. We hypothesize that in the presence of all four groups with the same average age, the frontal lobe was going to show smaller volumes in the MCI and ACS groups, related to cognitive decline in these subjects. However, more data is necessary to conduct this analysis.

Our study shows AMY PET usually outperforms FDG PET in all binary classification tasks when the features are extracted in a single PET modality approach. Trzepacz et al. [32] studied FDG PET, AMY PET, and MRI image features to predict MCI conversion to AD using the features individually and combined. They found that AMY PET and MRI features were more accurate in predicting a two-years conversion from MCI to AD. However, Xu et al. and Nozadi and Kadouri [7, 9] findings go on the contrary way. In a single modality analysis of FDG and ^{18}F-AV45 (Aβ tracer) PET, FDG PET features slightly improved discriminating MCI from AD and CNI.

Combining both PET traces in an ensemble has maintained the mean overall accuracy in the classification tasks compared to single PET modalities. The combination of FDG and AMY PET in classification experiments is unusual because both modalities are not acquired together in clinical practice [32]. However, FDG and AMY PET provide valuable and complementary information [35], as shown in our results. Usually, the classification studies associate FDG PET and MRI imaging features with CSF biomarkers, including the $A\beta_{42}$:$A\beta_{40}$ ratio, total tau, phosphorylated tau, and even genetic markers [8–11, 36]. However, CSF sampling is an invasive procedure, requiring lumbar puncture and does not present location and extension of the pathology, which is valuable information in the earliest stages of Aβ accumulation [35]. Therefore, Chételat et al. [35] defend AMY PET as a first-line diagnostic procedure, avoiding several visits and unnecessary invasive interventions.

The classification model's performance was close to the MRI approach in the multimodality imaging analysis because feature selection was resumed to the MRI features. MRI volume of the parietal and frontal lobe was selected in all models in the multimodality approach. Furthermore, for the CNI versus CONV and CNI versus ACS, only parietal image features were selected alone for the MRI single modality features, showing the importance of these brain regions in the cognitive decline (Supplementary Table 1).

In our work, SHAP-RFECV was used as a feature selector for each model with all imaging features, seeking not to exclude image features that generate the highest AUC. However, MRI features had the highest balanced accuracy and AUC for all models, like a single modality. Therefore, it was expected that it has more weight in the selection when combined with PET features. Xu et al. [9] used the weighted multimodality sparse representation-based classification to integrate FDG PET, AMY PET, and MRI features. They found that the imaging modalities contributed differently depending on the classification problem for different pairwise comparisons.

Tables 2 and 3 compare our best classification models' (RF and CAT) results with similar publications, using single imaging modalities and a multimodality approach. Accuracy is used for direct comparison (Supplementary Table 2). We did not find studies classifying between CNI and converters in the early stage of MCI or using a PET ensemble of FDG and AMY images to classify CNI versus CONV, MCI, or ACS individuals.

Our results using AMY PET, MR single modalities show similar performance in the classification when compared to the literature. We did not find studies using ML models to classify between CNI and converters using PET and MRI. Although direct comparison is not entirely appropriate due to different datasets (even different subjects in the same dataset) and different algorithms (SVM, RF, CAT, SRC, ELM), our results show good agreement with the performance reported in the literature.

Our FDG PET approach resulted in lower accuracies, even for CNI versus ACS binary classification task. Several aspects can explain the limited performance. Our FDG PET data was averaged between 45 to 60 min post-injection, which is less usual because usually PET images are averaged from 30 to 60 min post-injection. Furthermore, PET images were acquired from several PET scanners, which can lead to variations in the image quantification, affecting the imaging features calculated as the mean value of the normalized voxel intensity in the brain regions. No direct corrections for these differences were performed.

Moreover, we hypothesize that the use of large volumes in brain parcellation may have obscured metabolic FDG PET differences in smaller brain regions. In our study, the parcellated brain volume was an adaptation of Hammers atlas with 18 brain regions, a low number compared to other studies. Our option was supported by Samper-González et al. [37]. They analyzed the influence of different atlases consisting of 56 to 345 regions for brain parcellation on the classification using MRI and FDG PET. None provided differences in classification performance for CNI versus AD, CNI versus progressive MCI, and stable MCI versus progressive MCI. In our work, the low performance in the classification using FDG PET features can be attributed to the unspecific FDG uptake in brain regions. The average uptake over a brain region can obscure differences in hyper- or hypometabolism detection. Likely, a brain parcellation could highlight minor differences in FDG uptake between groups, especially in early decline. We believe brain parcellation will not significantly affect the classification performance using AMY PET and MRI because both markers are more specifically related to brain regions affected by the disease.

Some limitations are present in this study. Our datasets are smaller compared with the literature and get smaller in multimodality approaches because we included only individuals with all three imaging modalities. Moreover, our image features are normalized mean values of brain regions, determined by a modified Hammers' atlas in both the right and left hemispheres, potentially obscuring laterality differences and differences in smaller regions such as the cingulate cortex.

Another limitation of this study was the used sample size. In total, we had 131 individuals, distributed into four diagnosis groups. The inclusion and exclu-

Table 2
Comparison between the literature and our work results in single imaging modality approach (FDG PET, AMY PET, and MRI)

	Single Imaging Modality					
				Accuracy (%)		
Method	Imaging Modality	Algorithm	*n* of each study	CNI versus CONV	CNI versus MCI	CNI versus ACS
Nozadi and Kadouri [7]	FDG PET	SVM	208 CNI, 164 Early MCI, 189 Late MCI, 99 ACS	–	63.3 / 63.5 [1]	91.7
Nozadi and Kadouri [7]		RF	–	56.7 / 65.4 [1]	91.2	
Garali et al. [13]		RF	61 CNI, 29 MCI, 91 ACS	–	76.6	91.5
Xu et al. [9]		SRC	117 CNI, 110 MCI, 113 ACS	–	71.8	90.9
Gray et al. [12]		RF	35 CNI, 41 stable MCI, 34 progressive MCI, 37 ACS	–	60.2	86.5
Gray et al. [34]		SVM	54 CNI, 64 stable MCI, 53 progressive MCI, 50 ACS	–	70.7 [2]	80.9
Lin et al. [32]		ELM	200 CNI, 205 stable MCI, 110 progressive MCI, 102 ACS	–	–	76.7
Zhang et al. [11]		SVM	52 CNI, 99 MCI, 51 ACS	–	71.4	86.5
Pan et al. [2]		SVM	90 CNI, 88 MCI, 94 ACS	–	83.2	91.9
Our study		**RF**	**36 CNI, 24 CONV, 40 MCI, 31 ACS**	**66.9**	**59.1**	**80.3**
Our study		**CAT**	**74.0**	**65.3**	**82.8**	
Nozadi and Kadouri [7]	AMY PET	SVM	208 CN, 164 EMCI, 189 LMCI, 99 ACS	–	57.7 / 61.2 [1]	90.8
Nozadi and Kadouri [7]		RF	–	59.7 / 55.7 [1]	87.9	
Xu et al. [9]		SRC	117 CNI, 110 MCI, 113 ACS	–	70.5	83.7
Our study		**RF**	**22 CNI, 16 CONV, 40 MCI, 29 ACS**	**77.7**	**66.5**	**84.2**
Our study		**CAT**	**80.1**	**60.8**	**89.6**	
Xu et al. [9]	MRI	SRC	117 CNI, 110 MCI, 113 ACS	–	68.7	89.6
Gray et al. [12]		RF	35 CNI, 41 stable MCI, 34 progressive MCI, 37 ACS	–	69.1	82.1
Lin et al. [32]		ELM	200 CNI, 205 stable MCI, 110 progressive MCI, 102 ACS	–	–	74.5
Zhang et al. [11]		SVM	52 CNI, 99 MCI, 51 ACS	–	72	86.2
Toshkhujaev et al. [8]		SVM	28 CNI, 32 ACS	–	–	91.7 [3]
Our study		**RF**	**36 CNI, 24 CONV, 40 MCI, 31 ACS**	**89.1**	**100**	**97.2**
Our study		**CAT**	**89.1**	**98.2**	**97.2**	

[1] Classification between CNI and early MCI/late MCI; [2] Classification between CNI and prodromal MCI. [3] Considering only ADNI dataset results. CAT, categorical boosting; ELM, extreme learning machine; RF, random forest; SRC, sparse representation-based classification; SVM, support vector machine.

Table 3
Comparison between the literature and our work results in multimodality approach

	Imaging multimodality					
				Accuracy (%)		
Method	Model	*n* of each study	Algorithm	CNI versus CONV	CNI versus MCI	CNI versus ACS
Liu et al. [14]	FDG+MRI	52 CNI, 99 MCI, 51 ACS	SVM	–	78.8	94.4
Xu et al. [9]	FDG+AMY+MRI	117 CNI, 110 MCI, 113 ACS	wmSRC	–	74.5	94.8
Lei et al. [10]	PET+MRI+CSF	186 CNI, 393 MCI, 226 ACS	SVM	–	80.3	94.7
Zhang et al. [11]	FDG+MRI+CSF	52 CNI, 99 MCI, 51 ACS	SVM	–	76.4	93.2
Gray et al. [12]	FDG+MRI+CSF+Genetic	35 CNI, 41 stable MCI, 34 progressive MCI, 37 ACS	RF [1]	–	72.7	89
Gray et al. [12]	FDG+MRI+CSF+Genetic		RF [2]	–	65.3	87.1
Lin et al. [32]	FDG+MRI+CSF+Genetic	200 CNI, 205 stable MCI, 110 progressive MCI, 102 ACS	ELM	–	–	84.7
Tong et al. [15]	FDG+MRI+CSF+Genetic	35 CNI, 75 MCI, 37 ACS	RF [2]	–	73.1	86.2
Tong et al. [15]	FDG+MRI+CSF+Genetic		RF [3]	–	79.5	91.8
Our study	**FDG+AMY+MRI**	**22 CNI, 16 CONV, 40 MCI, 29 ACS**	**RF**	**90.3**	**96.7**	**88.9**
Our study	**FDG+AMY+MRI**		**CAT**	**100**	**96.7**	**94.4**

[1]Combined embedding features; [2] Concatenated features; [3]Non-linear fusion graphs. CAT, categorical boosting; CSF, cerebrospinal fluid; ELM, extreme learning machine; RF, random forest; SVM, support vector machine; wmSRC, weighted multimodality sparse representation-based classification.

sion criteria for the CONV group, aggregated with the possibility to have at least one PET and one MRI image in the determined interval between 6 months before and 12 months after clinical progression from CNI to MCI, has reduced our sample significantly. Further data are required to overcome these limitations and generalize our results.

It is important to state that FDG and AMY PET are rarely used in clinical practice for a joint analysis. Even AMY PET being more used in the suspect of dementia, mainly in the clinical signs of Alzheimer's pathology, and for differentiation between neurological disorders, FDG PET is still the most available radiotracer and is used in the absence of AMY PET. In this works, the authors wanted to show the contribution and potentialities of the use of both modalities together. Another point to be considered was the use of only biomarkers based on image data in this work. Future work will include clinical variables, e.g., age, sex, presence of *APOE* ε4, and CSF tau, in the model to improve the classification results.

CONCLUSION

Our work investigates ensemble tree-based classification models in early cognitive decline studies using features extracted from single and multimodality imaging approaches. In addition, our analysis includes the use of SHAP-RFECV as an unbiased feature selection, and early stages of aging cognitive decline, looking for subtle imaging differences indicating neurodegeneration.

The feature selection implemented with the Shapley additive explanations combined with the recursive feature elimination with cross-validation showed improvement in the classification models' accuracy. Among the studied models, the categorical boosting model and the random forest produced the best overall performance for classifying cognitively non-impaired individuals from early stages of cognitive decline, mild cognitive decline, and Alzheimer's clinical syndrome. Further work is required to analyze the impact on feature selection on the right and left-brain sides using an atlas with a higher number of regions to brain parcellation. Ongoing work includes a detailed evaluation of the selected brain regions and correlation with the cognitive decline spectrum in stable individuals and those that progress in the cognitive impairment.

ACKNOWLEDGMENTS

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nivel Superior – Brasil (CAPES) – Finance Code 001 and by PUCRS in the form of research scholarships.

We thank Jade Centeno Amandio for helping and supporting the data curation in this work.

Data collection and sharing were funded 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 funded by the National Institute on Aging, the

National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: 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 (http://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.

Authors' disclosures available online (https://www.j-alz.com/manuscript-disclosures/21-5164r3).

SUPPLEMENTARY MATERIAL

The supplementary material is available in the electronic version of this article: https://dx.doi.org/10.3233/JAD-215164.

REFERENCES

[1] Liu X, Chen K, Wu T, Weidman D, Lure F, Li J (2018) Use of multimodality imaging and artificial intelligence for diagnosis and prognosis of early stages of Alzheimer's disease. *Transl Res* **194**, 56-67.

[2] Pan X, Adel M, Fossati C, Gaidon T, Guedj E (2019) Multilevel feature representation of FDG-PET brain images for diagnosing Alzheimer's disease. *IEEE J Biomed Health Inform* **23**, 1499-1506.

[3] Berti V, Osorio RS, Mosconi L, Li Y, De Santi S, de Leon MJ (2010) Early detection of Alzheimer's disease with PET imaging. *Neurodegener Dis* **7**, 131-135.

[4] Jack CR, Bennett DA, Blennow K, Carrillo MC, Dunn B, Haeberlein SB, Holtzman DM, Jagust W, Jessen F, Karlawish J, Liu E, Molinuevo JL, Montine T, Phelps C, Rankin KP, Rowe CC, Scheltens P, Siemers E, Snyder HM, Sperling R, Elliott C, Masliah E, Ryan L, Silverberg N (2018) NIA-AA Research Framework: Toward a biological definition of Alzheimer's disease. *Alzheimers Dement* **14**, 535-562.

[5] Gómez-Ramírez J, Ávila-Villanueva M, Fernández-Blázquez MÁ (2020) Selecting the most important self-assessed features for predicting conversion to mild cognitive impairment with random forest and permutation-based methods. *Sci Rep* **10**, 20630.

[6] Grueso S, Viejo-Sobera R (2021) Machine learning methods for predicting progression from mild cognitive impairment to Alzheimer's disease dementia: A systematic review. *Alzheimers Res Ther* **13**, 162.

[7] Nozadi SH, Kadoury S, Alzheimer's Disease Neuroimaging Initiative (2018) Classification of Alzheimer's and MCI patients from semantically parcelled PET images: A comparison between AV45 and FDG-PET. *Int J Biomed Imaging* **2018**, 1247430.

[8] Toshkhujaev S, Lee KH, Choi KY, Lee JJ, Kwon G-R, Gupta Y, Lama RK (2020) Classification of Alzheimer's disease and mild cognitive impairment based on cortical and subcortical features from MRI T1 brain images utilizing four different types of datasets. *J Healthc Eng* **2020**, 3743171.

[9] Xu L, Wu X, Chen K, Yao L (2015) Multi-modality sparse representation-based classification for Alzheimer's disease and mild cognitive impairment. *Comput Methods Programs Biomed* **122**, 182-190.

[10] Lei B, Yang P, Wang T, Chen S, Ni D (2017) Relational-regularized discriminative sparse learning for Alzheimer's disease diagnosis. *IEEE Trans Cybernetics* **47**, 1102-1113.

[11] Zhang D, Wang Y, Zhou L, Yuan H, Shen D (2011) Multimodal classification of Alzheimer's disease and mild cognitive impairment. *Neuroimage* **55**, 856-867.

[12] Gray KR, Aljabar P, Heckemann RA, Hammers A, Rueckert D (2013) Random forest-based similarity measures for multi-modal classification of Alzheimer's disease. *Neuroimage* **65**, 167-175.

[13] Garali I, Adel M, Bourennane S, Guedj E (2018) Histogram-based features selection and volume of interest ranking for brain PET image classification. *IEEE J Transl Eng Health Med* **6**, 1-12.

[14] Liu F, Wee C-Y, Chen H, Shen D (2014) Inter-modality relationship constrained multi-modality multi-task feature selection for Alzheimer's disease and mild cognitive impairment identification. *Neuroimage* **84**, 466-475.

[15] Tong T, Gray K, Gao Q, Chen L, Rueckert D (2017) Multi-modal classification of Alzheimer's disease using nonlinear graph fusion. *Pattern Recognit* **63**, 171-181.

[16] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V (2011) Scikit-learn: Machine learning in Python. *J Mach Learn Res* **12**, 2825-2830.

[17] Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q, Liu T-Y (2017) Lightgbm: A highly efficient gradient boosting decision tree. *Adv Neural Inf Process Syst* **30**.

[18] Dorogush AV, Ershov V, Gulin A (2018) CatBoost: Gradient boosting with categorical features support. *arXiv preprint arXiv:1810.11363*.

[19] Safavian SR, Landgrebe D (1991) A survey of decision tree classifier methodology. *IEEE Trans Syst Man Cybern* **21**, 660-674.

[20] Vibha L, Harshavardhan GM, Pranaw K, Shenoy PD, Venugopal KR, Patnaik LM (2006) Classification of mammograms using decision trees. In *2006 10th Interna-*

tional Database Engineering and Applications Symposium (IDEAS'06), pp. 263-266.

[21] Breiman L (2001) Random forests. *Mach Learn* **45**, 5-32.

[22] Hammers A, Allom R, Koepp MJ, Free SL, Myers R, Lemieux L, Mitchell TN, Brooks DJ, Duncan JS (2003) Three-dimensional maximum probability atlas of the human brain, with particular reference to the temporal lobe. *Hum Brain Mapp* **19**, 224-247.

[23] Arbabshirani MR, Plis S, Sui J, Calhoun VD (2017) Single subject prediction of brain disorders in neuroimaging: Promises and pitfalls. *Neuroimage* **145**, 137-165.

[24] Vabalas A, Gowen E, Poliakoff E, Casson AJ (2019) Machine learning algorithm validation with a limited sample size. *PLoS One* **14**, e0224365.

[25] Bugaj M, Wrobel K, Iwaniec J (2021) Model explainability using SHAP values for LightGBM predictions. In *2021 IEEE XVIIth International Conference on the Perspective Technologies and Methods in MEMS Design (MEMSTECH)* IEEE, pp. 102–106.

[26] Ottoy J, Niemantsverdriet E, Verhaeghe J, De Roeck E, Struyfs H, Somers C, wyffels L, Ceyssens S, Van Mossevelde S, Van den Bossche T, Van Broeckhoven C, Ribbens A, Bjerke M, Stroobants S, Engelborghs S, Staelens S (2019) Association of short-term cognitive decline and MCI-to-AD dementia conversion with CSF, MRI, amyloid- and 18F-FDG-PET imaging. *Neuroimage Clin* **22**, 101771.

[27] Walhovd KB, Fjell AM, Dale AM, McEvoy LK, Brewer J, Karow DS, Salmon DP, Fennema-Notestine C (2010) Multimodal imaging predicts memory performance in normal aging and cognitive decline. *Neurobiol Aging* **31**, 1107-1121.

[28] Wang K, Liang M, Wang L, Tian L, Zhang X, Li K, Jiang T (2007) Altered functional connectivity in early Alzheimer's disease: A resting-state fMRI study. *Hum Brain Mapp* **28**, 967-978.

[29] Pergher V, Demaerel P, Soenen O, Saarela C, Tournoy J, Schoenmakers B, Karrasch M, Van Hulle MM (2019) Identifying brain changes related to cognitive aging using VBM and visual rating scales. *Neuroimage Clin* **22**, 101697.

[30] Trzepacz PT, Yu P, Sun J, Schuh K, Case M, Witte MM, Hochstetler H, Hake A (2014) Comparison of neuroimaging modalities for the prediction of conversion from mild cognitive impairment to Alzheimer's dementia. *Neurobiol Aging* **35**, 143-151.

[31] Chételat G, Arbizu J, Barthel H, Garibotto V, Law I, Morbelli S, van de Giessen E, Agosta F, Barkhof F, Brooks DJ, Carrillo MC, Dubois B, Fjell AM, Frisoni GB, Hansson O, Herholz K, Hutton BF, Jack CR, Jr., Lammertsma AA, Landau SM, Minoshima S, Nobili F, Nordberg A, Ossenkoppele R, Oyen WJG, Perani D, Rabinovici GD, Scheltens P, Villemagne VL, Zetterberg H, Drzezga A (2020) Amyloid-PET and ^{18}F-FDG-PET in the diagnostic investigation of Alzheimer's disease and other dementias. *Lancet Neurol* **19**, 951-962.

[32] Lin W, Gao Q, Yuan J, Chen Z, Feng C, Chen W, Du M, Tong T (2020) Predicting Alzheimer's disease conversion from mild cognitive impairment using an extreme learning machine-based grading method with multimodal data. *Front Aging Neurosci* **12**, 77.

[33] Samper-González J, Burgos N, Bottani S, Fontanella S, Lu P, Marcoux A, Routier A, Guillon J, Bacci M, Wen J, Bertrand A, Bertin H, Habert M-O, Durrleman S, Evgeniou T, Colliot O (2018) Reproducible evaluation of classification methods in Alzheimer's disease: Framework and application to MRI and PET data. *Neuroimage* **183**, 504-521.

[34] Gray KR, Wolz R, Heckemann RA, Aljabar P, Hammers A, Rueckert D (2012) Multi-region analysis of longitudinal FDG-PET for the classification of Alzheimer's disease. *Neuroimage* **60**, 221-229.