Reproducible evaluation of methods for predicting progression to Alzheimer's disease from clinical and neuroimaging data

Jorge Samper-Gonzalez, Ninon Burgos, Simona Bottani, Marie-Odile Habert, Theodoros Evgeniou, Stephane Epelbaum, Olivier Colliot

Reproducible evaluation of methods for predicting progression to Alzheimer's disease from clinical and neuroimaging data

Jorge Samper-González[a,b,c,d,e], Ninon Burgos[b,a,c,d,e], Simona Bottani[a,b,c,d,e], Marie-Odile Habert[g,i], Theodoros Evgeniou[j], Stéphane Epelbaum[b,a,c,d,e,f], Olivier Colliot[b,a,c,d,e,f,g], and the Alzheimer's Disease Neuroimaging Initiative

[a]Inria, Aramis project-team, Paris, France
[b]Institut du Cerveau et de la Moelle épinière, ICM, F-75013, Paris, France
[c]Inserm, U 1127, F-75013, Paris, France
[d]CNRS, UMR 7225, F-75013, Paris, France
[e]Sorbonne Université, F-75013, Paris, France
[f]AP-HP, Hôpital Pitié-Salpêtrière, Department of Neurology, IM2A, Paris, France
[g]AP-HP, Hôpital Pitié-Salpêtrière, Department of Neuroradiology, Paris, France
[h]LIB, Sorbonne Université, Inserm U 1146, CNRS UMR 7371, Paris, France
[i]AP-HP, Hôpital Pitié-Salpêtrière, Department of Nuclear Medicine, Paris, France
[j]INSEAD, Bd de Constance, 77305 Fontainebleau, France

ABSTRACT

Various machine learning methods have been proposed for predicting progression of patients with mild cognitive impairment (MCI) to Alzheimer's disease (AD) using neuroimaging data. Even though the vast majority of these works use the public dataset ADNI, reproducing their results is complicated because they often do not make available elements that are essential for reproducibility, such as selected participants and input data, image preprocessing and cross-validation procedures. Comparability is also an issue. Specially, the influence of different components like preprocessing, feature extraction or classification algorithms on the performance is difficult to evaluate. Finally, these studies rarely compare their results to models built from clinical data only, a critical aspect to demonstrate the utility of neuroimaging. In our previous work,[1,2] we presented a framework for reproducible and objective classification experiments in AD, that included automatic conversion of ADNI database into the BIDS community standard, image preprocessing pipelines and machine learning evaluation. We applied this framework to perform unimodal classifications of T1 MRI and FDG-PET images. In the present paper, we extend this work to the combination of multimodal clinical and neuroimaging data. All experiments are based on standard approaches (namely SVM and random forests). In particular, we assess the added value of neuroimaging over using only clinical data. We first demonstrate that using only demographic and clinical data (gender, education level, MMSE, CDR sum of boxes, ADASCog) results in a balanced accuracy of 75% (AUC of 0.84). This performance is higher than that of standard neuroimaging-based classifiers. We then propose a simple trick to improve the performance of neuroimaging-based classifiers: training from AD patients and controls (rather than from MCI patients) improves the performance of FDG-PET classification by 5 percent points, reaching the level of the clinical classifier. Finally, combining clinical and neuroimaging data, prediction results further improved to 80% balanced accuracy and an AUC of 0.88). These prediction accuracies, obtained in a reproducible way, provide a base to develop on top of it and, to compare against, more sophisticated methods. All the code of the framework and the experiments is publicly available at https://gitlab.icm-institute.org/aramislab/AD-ML.

Keywords: Alzheimer's disease, machine learning, random forests, SVM, multimodal data

Further author information: (Send correspondence to O.C.)
O.C.: E-mail: olivier.colliot@upmc.fr
J.S.G.: E-mail: jorge.samper-gonzalez@inria.fr

1. INTRODUCTION

Alzheimer's disease (AD) is the first cause of dementia worldwide, affecting over 20 million people. Identifying AD at an early stage is essential to ensure a proper care of patients and also to develop and test novel treatments. AD progression can be characterized using different measurements. Neuropsychological tests can measure the cognitive decline of a subject in areas such as learning and memory, executive functioning, processing speed, attention, and semantic knowledge.[3] Neuroimaging can provide measures of atrophy due to gray matter loss with anatomical magnetic resonance imaging (MRI), of hypometabolism with 18F-fluorodeoxyglucose positron emission tomography (FDG PET) and of accumulation of amyloid-beta and tau proteins with amyloid-PET and tau-PET imaging.[4] There is an interest in exploring the predicting capabilities of these markers, that reflect different aspects of the disease, from an early stage. A large body of research on the early stages of AD has focused on patients with mild cognitive impairment (MCI), who have objective cognitive deficits but do not yet have dementia. Some of these patients will subsequently develop dementia while others will remain stable. Identifying those who will develop AD is major challenge.

Numerous machine learning (ML) and deep learning[5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,26] approaches have been proposed to predict progression to AD among patients with MCI from neuroimaging data (see e.g 27, 28, 29 for reviews on that topic). To compare these approaches in an objective way is practically impossible, given their differences in: i) subsets of patients; ii) image preprocessing pipelines; iii) feature extraction and selection; iv) machine learning algorithms; v) cross-validation procedures and vi) reported evaluation metrics. This makes it difficult to establish if a method outperforms another or to measure the contribution of different components (preprocessing, features, ML algorithm), limiting the practical impact of these studies. Additionally, these studies rarely compare their results to models built from clinical/cognitive data only. This is an important point to demonstrate the utility of sophisticated neuroimaging-based methods. Indeed, cognitive assessments are cheaper to perform and do not require sophisticated equipment, compared to neuroimaging or other biomarkers. Furthermore, these different components are often not made publicly available by the authors. Reproducibility, the ability to reproduce results based on the same data and experimental procedures, can be a first step in the direction of making the evaluation of machine learning approaches more objective. In that respect, data sharing, storing of data using community standards, fully automatic data manipulation and sharing of code are essential to enable reproducible research.

In our previous work,[1,2] we proposed a framework for the reproducible evaluation of machine learning algorithms in AD. The framework comprised the following components. Tools for fully automatic conversion into the BIDS (Brain Imaging Data Structure) community standard[30] of public datasets including the Alzheimer's Disease Neuroimaging Initiative (ADNI). This saves other researchers a large amount of time and allows them to use or to reproduce experiments using this data. We proposed standard preprocessing and feature extraction pipelines for different imaging modalities that are made available in a modular way. Tools for classification using standard machine learning algorithms (support vector machine, random forest, logistic regression), following rigorous validation and providing extensive reporting, were developed. This set of tools allows the objective evaluation of the influence of specific elements, given that they can be straightforwardly replaced. This framework was then used for an extensive evaluation of different parameters, features, and classification algorithms on classification tasks using unimodal neuroimaging data (T1 MRI and FDG PET).

In this paper, we extend our previous work to the combination of multimodal clinical and neuroimaging data. The present study is focused on the prediction of progression of subjects with mild cognitive impairment (MCI) to AD, a clinically important task. Compared to our previous work, the contributions of the present paper are the following. First, we compare the performance of neuroimaging-based models to that of models using only clinical data. Indeed, given that clinical data is more widely available, it would be a more natural choice as input data for baseline models. Second, we propose a simple trick to improve the performance of neuroimaging-based models: training on AD patients and controls (rather than on progressive and stable MCI patients) and applying the resulting model to prediction of progression to AD. Third, we assess the performance of the combination of multiple modalities (clinical, neuroimaging and APOE genotype). Finally, while the previous paper was restricted to the prediction of progression to AD at 36 months, we study the performance for various dates (from 12 to 36 months).

All the code of the framework and the experiments is publicly available: general-purpose tools have been integrated into Clinica*,[31] an open-source software platform for neuroimaging studies, and the paper-specific code is available at: https://gitlab.icm-institute.org/aramislab/AD-ML.

2. MATERIALS AND METHODS

2.1 Data

All the data used in the preparation of this article were obtained from the ADNI database. The same group of subjects as in 1 was used, except three subjects that were excluded because of missing neuropsychological tests. It consists of 748 subjects for whom a T1w MRI and an FDG PET scan, with a known effective resolution, were available at baseline. Our definition for stable and progressing mild cognitive impairment subjects was:

- $sMCI_N$: subjects who were diagnosed as MCI, EMCI or LMCI at baseline, were followed during at least N months and did not progress to AD between their first visit and the visit at N months;
- $pMCI_N$: subjects who were diagnosed as MCI, EMCI or LMCI at baseline, were followed during at least N months and progressed to AD between their first visit and the visit at N months.

Even though not the main focus of this work, CN $A\beta$- (cognitively normal subjects with a negative amyloid status) and AD $A\beta$+ subjects (AD patients with a positive amyloid status) were used for some of the experiments (section 2.5.2). They were diagnosed at baseline and had a known amyloid status, determined from a PiB or an AV45 PET scan using standard cutoff values of 1.47 and 1.10, respectively.[32]

Population details can be observed in Table 1. Subject lists were obtained automatically using our publicly-available code.

Table 1. Studied populations. Summary of participant demographics, mini-mental state examination (MMSE) and global clinical dementia rating (CDR) scores

	N	Age*	Gender	MMSE*	CDR
$sMCI_{36}$	340	71.8±7.5 [55.0, 88.6]	201 M / 139 F	28.1±1.6 [23, 30]	0.5: 340
$pMCI_{36}$	167	74.9±6.9 [55.0, 88.3]	98 M / 69 F	27.0±1.7 [24, 30]	0.5: 166; 1: 1
CN $A\beta$-	115	72.2±6.1 [56.2, 89.0]	59 M / 56 F	29.0±1.3 [24,30]	0: 115
AD $A\beta$+	126	74.1±8.1 [55.1, 90.3]	65 M / 61 F	22.9±2.1 [19, 26]	0.5: 54; 1: 71; 2: 16

* Values are presented as mean±SD [range]. M: male, F: female

Additionally, for some of the experiments, we had to use different subsets of the population, because some of the features were not available for all the subjects. Specifically, this was the case for experiments using amyloid status and for those using volumetric MRI and regional PET measures available in the ADNIMERGE tabular file. The two tables describing each used subset are included in the Results section (Tables 6 and 4). For these two subsets, we verified that the characteristics of age, gender, MMSE and CDR of these subgroups followed the same distribution as that of the study population.

2.2 Data conversion

ADNI is a complex multimodal dataset with plenty of incomplete data, multiple instances of a given modality and complex metadata. To allow reproducibility, as in our previous work, ADNI was fully automatically converted into BIDS format, a community standard30. We performed conversion of T1w MRI and FDG PET imaging modalities, and of selected clinical/cognitive data. For T1 scans, gradwarp and B1-inhomogeneity corrected images were selected when available, otherwise the original image was selected. When several T1 scans were available for a single session, the preferred scan, if available, or higher quality scan, was chosen. For FDG PET scans, the images co-registered and averaged across time frames were selected. Images in DICOM format were converted

* http://clinica.run

to NIFTI format. All images are organized in a folder hierarchy following the BIDS specifications. Regarding the clinical data, scores of interest were extracted from the csv files provided by ADNI and gathered in tsv files located in the BIDS folder hierarchy. From the clinical data, we used socio-demographic data (gender, education level), APOE genotype, and five neuropsychological tests results: mini-mental state examination (MMSE) score, the sum of boxes of clinical dementia rating (CDR-SB) test, the scores of Alzheimer's Disease Assessment Scale cognitive sub-scale (ADASCog) separated into four categories (memory, language, concentration and praxis), the Logical Memory (immediate and delayed recall) test, and the Rey Auditory Verbal Learning Test (RAVLT). Also, some volumetric and regional neuroimaging measures provided by ADNI (available in the ADNIMERGE csv file) were gathered: volumetric measures for different regions (ventricles, hippocampus, entorhinal cortex, fusiform gyrus, mid-temporal gyrus) computed from MRI and the average FDG-PET of angular, temporal, and posterior cingulate regions. Volumetric MRI measures were normalized according to the intracranial volume of each subject. The converter that we developed was integrated into the Clinica software (see 1). Note that the downloaded files must be kept exactly as they were downloaded. The different steps are then performed by the automatic converter (no user intervention is required).

2.3 Preprocessing and feature extraction

T1w MR and FDG PET images were preprocessed as in 1. For anatomical T1w MRI, the t1-volume-new-template pipeline from Clinica was applied. Based on SPM12, it applies the Unified Segmentation,[33] DARTEL[33] and DARTEL to MNI[34] procedures. As a result, we obtain tissue maps in a common space, providing a voxel-wise correspondence across subjects. FDG PET preprocessing was done using pet-volume pipeline from Clinica, which is also based on SPM12. Making use of the T1w preprocessing pipeline, PET image was registered to the T1w native space and then to the common space. Intensity normalization using the pons region as reference and brain masking were applied. The resulting standardized uptake value ratio (SUVR) maps are also in a common space providing voxel-wise correspondence across subjects.

2.4 Age correction

Age correction of T1w MR and FDG PET images was done, separately, as in 35. For each voxel, a linear regression was performed between the age and the value (GM density or PET SUVR) at this voxel, using CN amyloid negative subjects. Then images were corrected according to the expected values for the subject's age.

2.5 Classification approaches

To predict the progression of MCI subjects to AD, we trained classifiers for the task sMCI vs pMCI, based on different data modalities. We first assessed prediction using clinical/cognitive data alone. We then studied the use of T1w MRI and FDG-PET imaging data, either alone or in combination with clinical/cognitive data.

All the classifications were done using Clinica software tools which wraps different tools from scikit-learn[†]. We relied on standard classifiers, namely support vector machines (SVM) and random forests (RF).

2.5.1 Classification using clinical data

First, we considered only demographic and clinical data. The first task used as features a combination of gender, education level, MMSE score and the sum of boxes of CDR test (we will refer to this set of features as $\text{Clinical}_{\text{base}}$). MMSE and CDR-SB tests can be performed in standard clinical centers. We then tested the added value of two memory tests: RAVLT and Logical Memory (LogMem). We also evaluated the added value of ADAS-Cog, a test that is usually performed only in more specialized centers. Note that the ADAS-Cog was separated into four domains, as explained in Section 2.2. Finally, we assessed the added value of the APOE4 genotype.

[†] http://scikit-learn.org

2.5.2 Image-based classification

We then assessed the performance of neuroimaging data, namely T1w MRI and FDG PET modalities taken separately. For this purpose, we used SVM classifiers trained on all voxels of preprocessed and age corrected images. First, we used a standard approach in which the classifier was trained on the population of sMCI and pMCI subjects. We then assessed another approach in which the classifier was trained to distinguish between CN $A\beta$- and AD $A\beta$+ groups, and the resulting classifier was applied to sMCI and pMCI subjects to predict disease progression. Indeed, if we see the evolution of Alzheimer's disease from subjects being cognitive normal progressing in time to demented patients, we can define CN $A\beta$-, sMCI, pMCI and AD $A\beta$+ as an ordered list of possible states of a subject. Training the classifier on the simpler task of differentiating the CN $A\beta$- and AD $A\beta$+ states, could allow it to learn a disease pattern that would be more difficult to obtain if training directly on sMCI and pMCI subjects. We tested whether the information contained in this learned classifier is directly transferable to the problem of predicting disease progression.

2.5.3 Integrating clinical and imaging data

Finally, we assessed the combination of clinical and neuroimaging data. For each neuroimaging modality, we constructed a score from the SVM classifier. Indeed, for each image, a score can be obtained from an SVM as $\hat{y} = w * x + b$. For each subject, two scores are computed, one for T1w MRI and one for FDG PET scans (Scores $_{\text{T1, FDG}}$). These scores can be seen as markers of AD-like spatial pattern of neurodegeneration: gray matter atrophy pattern in the case of anatomical T1w MRI and hypometabolism pattern in the case of FDG PET. We then combined demographic and clinical data with these two scores (containing information from imaging data) into a random forest classifier. Namely, we first used Clinical$_{\text{base}}$ features, and Scores $_{\text{T1, FDG}}$. We then added RAVLT and ADASCog tests.

Moreover, we compared the performance of the neuroimaging SVM scores (Scores $_{\text{T1, FDG}}$) to that of volumetric MRI measures and regional FDG-PET value (as available in ADNIMERGE). For this purpose, the same experiments, using Clinical$_{\text{base}}$ features, RAVLT and ADASCog tests, and volumetric and FDG-PET data were performed on the subpopulation containing all the required values (Table 4).

2.5.4 Integrating amyloid status

In addition, we explored the predictive value of amyloid status, either in isolation or combined with the other studied variables (clinical, T1 and FDG-PET neuroimaging). The status was determined from a PiB or an AV45 PET scan using standard cutoff values of 1.47 and 1.10, respectively.[32] These experiments were performed on the subpopulation for which amyloid status was available (Table 6).

2.5.5 Prediction at different time-points

We also wanted to assess the influence of using different time spans for MCI subjects progressing to AD. We obtained lists of subjects who progressed to AD before 12, 18, 24 and 30 months from the baseline. We assessed the performance of models using: i) Clinical$_{\text{base}}$ features and ADASCog; ii) Clinical$_{\text{base}}$ features, ADASCog and Scores $_{\text{T1, FDG}}$.

2.6 Validation

Cross validation, following strict guidelines as presented in 36, was applied to all the experiments: results are the mean of 250 iterations of stratified random splits with 80% of samples used for training and remaining 20% for testing. RF classifiers were trained using fixed hyperparameters: 100 trees, tree depth limited to 5 levels and only the square root of the total number of features is considered when looking for a split. For linear SVM classifiers, the hyperparameter controlling regularization was optimized using an inner 10-fold cross validation.

As output of the classification, we report the balanced accuracy, area under the ROC curve (AUC), accuracy, sensitivity, specificity and, in addition, the predicted class for each subject, so the user can calculate other desired metrics with this information.

3. RESULTS

3.1 Classification using clinical data

Classification results using only clinical/cognitive data are presented in Table 2.

Classifications obtained using sociodemographical and simpler neuropsychological tests (which can be performed in a routine clinical environment), namely MMSE and CDR-SB, provided a balanced accuracy of only 68% and an AUC of 0.75. The addition of the RAVLT led to a strong improvement (balanced accuracy of 74%, AUC of 0.82). This was also the case for the addition of the ADAS-Cog features (balanced accuracy of 75%, AUC of 0.84). Compared to the RAVLT, the other memory test LogMem, resulted in a much lower improvement (balanced accuracy of 70%, AUC of 0.79) and the combination of both memory tests (RAVLT and LogMem) did not improve the results. Finally, the combination of ADAS-Cog and RAVLT provided a very small improvement (balanced accuracy of 76%, AUC of 0.85). On the other hand, the addition of APOE4 did not improve the performance.

Based on these results, the APOE was not considered in the subsequent experiments and the RAVLT was preferred to the LogMem test.

Table 2. Results for models based on clinical data only

Classifier - Features	Bal. acc.	AUC	Acc.	Sens.	Spec.
RF - $\text{Clinical}_{\text{base}}$	0.660	0.726	0.684	0.587	0.734
RF - $\text{Clinical}_{\text{base}}$ + LogMem	0.702	0.792	0.728	0.624	0.78
RF - $\text{Clinical}_{\text{base}}$ + RAVLT	**0.742**	0.823	0.75	0.717	0.767
RF - $\text{Clinical}_{\text{base}}$ + LogMem + RAVLT	0.745	0.836	0.755	0.712	0.777
RF - $\text{Clinical}_{\text{base}}$ + ADAS	0.754	0.836	0.760	0.736	0.772
RF - $\text{Clinical}_{\text{base}}$ + RAVLT + ADAS	**0.762**	0.852	0.768	0.743	0.781
RF - $\text{Clinical}_{\text{base}}$ + RAVLT + APOE4	**0.756**	0.838	0.759	0.750	0.763
RF - $\text{Clinical}_{\text{base}}$ + ADAS + APOE4	0.757	0.842	0.766	0.731	0.784
RF - $\text{Clinical}_{\text{base}}$ + RAVLT + ADAS + APOE4	**0.765**	0.857	0.772	0.746	0.785

$\text{Clinical}_{\text{base}}$: gender, education level, MMSE score, sum of boxes of CDR test

3.2 Integration of imaging and clinical data

Classification results using either neuroimaging alone or in combination with clinical/cognitive data are presented in Table 3.

When trained on sMCI vs pMCI, the performance of T1w MRI and FDG PET data alone was substantially lower than that of clinical data (including ADAS-Cog or RAVLT) and comparable to that of $\text{Clinical}_{\text{base}}$. Still, the performance of FDG PET was superior to that of MRI. Interestingly, training SVM classifiers on the CN $A\beta$- vs AD $A\beta$+ task and evaluating them on sMCI vs pMCI, improved the performance for FDG PET modality (balanced accuracy of 76% and AUC of 0.82) compared to training and testing on sMCI and pMCI classes (balanced accuracy of 71% and AUC of 0.78). Using this approach, FDG PET alone reached a performance similar to that of clinical data (including ADAS-Cog or RAVLT).

The combination of clinical and imaging data further improved the results. When using T1w MRI and FDG PET scores, socio-demographics, and neuropsychological tests, we reached a balanced accuracy of 80% and the AUC was 0.88.

Classification results using volumetric MRI features and a regional FDG PET measure (obtained from ADNIMERGE file) are shown in Table 5. The studied subpopulation is presented in Table 4. The performances were slightly lower than that obtained using the scores for T1 and FDG PET obtained from SVMs. In this subpopulation, only ADNIMERGE features provided a balanced accuracy of 73% and an AUC of 0.83, while only SVM scores provided a balanced accuracy of 78% and an AUC of 0.81. In the case where Clinical base and ADAS are also added, the use of ADNIMERGE gave a balanced accuracy of 78% and an AUC of 0.88, while SVM scores produced a balanced accuracy of 80% and an AUC of 0.89.

Table 3. Results for models based on imaging data only and on the combination of imaging and clinical data

Classifier - Features	Bal. acc.	AUC	Acc.	Sens.	Spec.
SVM - T1w MRI	0.670	0.736	0.698	0.586	0.754
SVM (trained on CN $A\beta$- vs AD $A\beta$+) - T1w MRI	0.679	0.764	0.708	0.547	0.811
SVM - FDG PET	0.708	0.777	0.732	0.633	0.782
SVM (trained on CN $A\beta$- vs AD $A\beta$+) - FDG PET	**0.761**	**0.818**	**0.788**	0.666	0.856
RF - $\text{Clinical}_{\text{base}}$ + Score $_{\text{T1}}$	0.717	0.792	0.732	0.671	0.763
RF - $\text{Clinical}_{\text{base}}$ + Score $_{\text{FDG}}$	0.760	0.831	0.791	0.669	0.852
RF - $\text{Clinical}_{\text{base}}$ + Scores $_{\text{T1,FDG}}$	0.769	0.855	0.796	0.685	0.852
RF - $\text{Clinical}_{\text{base}}$ + RAVLT + Scores $_{\text{T1,FDG}}$	**0.791**	0.881	0.809	0.735	0.846
RF - $\text{Clinical}_{\text{base}}$ + ADAS + Scores $_{\text{T1,FDG}}$	**0.790**	0.873	0.810	0.729	0.851
RF - $\text{Clinical}_{\text{base}}$ + RAVLT + ADAS + Scores $_{\text{T1,FDG}}$	**0.792**	0.888	0.811	0.736	0.849

$\text{Clinical}_{\text{base}}$: gender, education level, MMSE score, sum of boxes of CDR test

Table 4. Subpopulation used for the experiments using volumetric MRI and a regional FDG-PET feature (as available in ADNIMERGE)

	N	Age*	Gender	MMSE*	CDR
sMCI_{36}	267	71.4±7.5 [55.0, 88.6]	158 M / 109 F	28.2±1.6 [23, 30]	0.5: 267
pMCI_{36}	135	73.2±6.9 [55.0, 88.3]	80 M / 55 F	27.1±1.8 [24, 30]	0.5: 134; 1: 1

* Values are presented as mean±SD [range]. M: male, F: female

Table 5. Results using volumetric MRI and a regional FDG-PET feature (as available in ADNIMERGE). The studied subpopulation is described in Table 4.

Classifier - Features	Bal. acc.	AUC	Acc.	Sens.	Spec.
RF - $\text{Clinical}_{\text{base}}$	0.646	0.711	0.680	0.545	0.747
RF - $\text{Clinical}_{\text{base}}$ + RAVLT	0.716	0.816	0.726	0.683	0.748
RF - $\text{Clinical}_{\text{base}}$ + ADAS	0.769	0.850	0.779	0.737	0.800
RF - $\text{Clinical}_{\text{base}}$ + RAVLT + ADAS	0.767	0.860	0.776	0.740	0.795
RF - ADNI $_{\text{T1}}$	0.699	0.773	0.734	0.594	0.804
RF - ADNI $_{\text{FDG}}$	0.696	0.764	0.719	0.628	0.765
RF - ADNI $_{\text{T1,FDG}}$	0.733	0.828	0.756	0.663	0.802
RF - $\text{Clinical}_{\text{base}}$ + RAVLT + ADNI $_{\text{T1,FDG}}$	0.782	0.869	0.795	0.74	0.823
RF - $\text{Clinical}_{\text{base}}$ + RAVLT + ADAS + ADNI $_{\text{T1,FDG}}$	0.796	0.885	0.809	0.755	0.836
Scores $_{\text{T1}}$	0.661	0.722	0.665	0.649	0.673
Scores $_{\text{FDG}}$	0.755	0.805	0.791	0.649	0.862
Scores $_{\text{T1,FDG}}$	0.776	0.814	0.806	0.686	0.866
RF - $\text{Clinical}_{\text{base}}$ + RAVLT + Scores $_{\text{T1,FDG}}$	0.799	0.883	0.818	0.740	0.857
RF - $\text{Clinical}_{\text{base}}$ + RAVLT + ADAS + Scores $_{\text{T1,FDG}}$	0.803	0.896	0.822	0.746	0.860

3.3 Integration of amyloid status

Classification results using amyloid status are shown in Table 7. The studied subpopulation is presented in Table 6. Results show that classification accuracy improves when amyloid status is added to the Clinical base features and to different combinations of ADAS-Cog and T1 and FDG PET scores, but it does not provide a superior performance than just using a combination of the other features.

3.4 Prediction at different time-points

Results for prediction at different time-points are presented in Table 8. The performance improved along with the follow up time. This may be due to the reduced number of progressing MCI subjects for shorter follow up

Table 6. Subpopulation used for the experiments using the amyloid status

	N	Age*	Gender	MMSE*	CDR
$sMCI_{36}$	265	71.0±7.3 [55.0, 88.6]	148 M / 117 F	28.3±1.6 [23, 30]	0.5: 265
$pMCI_{36}$	94	72.9±7.0 [55.0, 85.9]	52 M / 42 F	27.2±1.8 [24, 30]	0.5: 93; 1.0: 1

* Values are presented as mean±SD [range]. M: male, F: female

Table 7. Results using the amyloid status. The studied subpopulation is described in Table 6.

Classifier - Features	Bal. acc.	AUC	Acc.	Sens.	Spec.
RF - $Clinical_{base}$	0.667	0.75	0.695	0.591	0.742
RF - $Clinical_{base}$ + Score $_{T1}$	0.741	0.819	0.760	0.688	0.793
RF - $Clinical_{base}$ + Score $_{FDG}$	0.773	0.854	0.808	0.680	0.867
RF - $Clinical_{base}$ + RAVLT	0.725	0.828	0.759	0.653	0.797
RF - $Clinical_{base}$ + ADAS	0.761	0.855	0.782	0.703	0.819
RF - $Clinical_{base}$ + RAVLT + ADAS	0.744	0.870	0.793	0.638	0.849
RF - $Clinical_{base}$ + RAVLT + Scores $_{T1,FDG}$	0.797	0.889	0.843	0.699	0.895
RF - $Clinical_{base}$ + ADAS + Scores $_{T1,FDG}$	0.803	0.888	0.830	0.730	0.876
RF - $Clinical_{base}$ + RAVLT + ADAS + Scores $_{T1,FDG}$	0.798	0.898	0.850	0.688	0.908
RF - $A\beta$					
RF - $Clinical_{base}$ + $A\beta$	0.700	0.786	0.706	0.685	0.716
RF - $Clinical_{base}$ + Score $_{T1}$ + $A\beta$	0.747	0.837	0.763	0.704	0.789
RF - $Clinical_{base}$ + Score $_{FDG}$ + $A\beta$	0.782	0.862	0.816	0.691	0.873
RF - $Clinical_{base}$ + RAVLT + $A\beta$	0.782	0.876	0.799	0.745	0.819
RF - $Clinical_{base}$ + ADAS + $A\beta$	0.765	0.860	0.784	0.713	0.816
RF - $Clinical_{base}$ + RAVLT + ADAS + $A\beta$	0.796	0.900	0.829	0.725	0.866
RF - $Clinical_{base}$ + RAVLT + Scores $_{T1,FDG}$ + $A\beta$	0.799	0.906	0.837	0.719	0.879
RF - $Clinical_{base}$ + ADAS + Scores $_{T1,FDG}$ + $A\beta$	0.805	0.890	0.830	0.737	0.872
RF - $Clinical_{base}$ + RAVLT + ADAS + Scores $_{T1,FDG}$ + $A\beta$	0.800	0.911	0.848	0.697	0.902

Table 8. Balanced accuracy for sMCI vs pMCI task for different follow up times. Number of subjects in each class.

Features	12 m	18 m	24 m	30 m	36 m
$Clinical_{base}$ + ADAS	0.630	0.654	0.707	0.714	0.754
$Clinical_{base}$ + ADAS + Scores $_{T1,FDG}$	0.611	0.679	0.724	0.728	0.790
Number of subjects	**12 m**	**18 m**	**24 m**	**30 m**	**36 m**
sMCI	467	448	415	407	340
pMCI	39	55	87	87	167

Table 9. Balanced accuracy for sMCI vs pMCI task for different follow up times. Same number of subjects was used for each time point (n(sMCI) = 78, n(pMCI) = 39)

Features	12 m	18 m	24 m	30 m	36 m
$Clinical_{base}$ + ADAS	0.722	0.775	0.695		0.775
$Clinical_{base}$ + ADAS + Scores $_{T1,FDG}$	0.704	0.768	0.697		0.798

times. Therefore, we also tested with a fixed number of participants at each time-point. To that purpose, for each time-point, we randomly chose a set of participants which number of pMCI matches that of those available for 12 months. The number of sMCI was chosen to be the double of that of pMCI in order to avoid a strong imbalance between classes. Indeed, we previously showed[1] that strong class imbalances (typically 1:6) have a negative effect on performances but that moderate imbalances (such as 1:2) have no impact. Results are shown in Table 9. We observe that ...

4. CONCLUSIONS

In this paper, we extend our previous work by proposing a reproducible evaluation of methods to predict progression of MCI subjects to AD, based on multimodal clinical and neuroimaging data. Importantly, all the tools (including automatic data conversion, standardized imaging preprocessing pipelines and machine learning evaluation framework) are made publicly available.

Our experimental results, based on rigorous and transparent evaluation procedures, led to several interesting conclusions.

First, we found that when using only socio-demographics and neuropsychological tests as input, it is already possible to achieve decent performances. Also, we can observe that the use of other cognitive tests (here the RAVLT or ADAS-Cog) led to substantially higher performances. Note that this is not a tautology, since only clinical data at baseline (MCI diagnosed subjects) is used to predict diagnosis at a future point in time. Importantly, the performance of such models was superior to that of standard classifiers based on neuroimaging data only. We believe that it is an important message for the medical imaging community, in which performance of imaging-based classification methods is rarely compared to that of clinical data only.

Second, we proposed a simple trick that allows a substantial improvement in the performance of a standard neuroimaging-based classifier. The trick consists in training the model on a simpler task (CN $A\beta$- vs AD $A\beta$+) and applying it to a more difficult task (prediction of progression in MCI patients). This can be seen as a very simple form of transfer learning, a widely used approach in machine learning.

Finally, the combination of clinical and imaging data further improved the results. The balanced accuracy of such model was 80%. Such performance is comparable to the majority of state-of-the-art machine learning results, as summarized in 27, 28, 29, which present classification accuracies generally ranging from 60% to 80%, in some cases up to 86% (for prediction of conversion in a short period). The performance is also comparable to deep learning results,[24, 25, 26] whose classification accuracies range between 60% and 84%. It is interesting to note that our results were obtained using standard classification techniques (SVM and random forest) but were comparable to those obtained using more sophisticated techniques.

In conclusion, we proposed a reproducible framework for evaluation of methods for predicting progression to AD. Results obtained using this approach could serve as baseline for comparison of more sophisticated approaches.

REFERENCES

[1] Samper-Gonzlez, 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., and Colliot, O., "Reproducible evaluation of classification methods in Alzheimer's disease: Framework and application to MRI and PET data," *NeuroImage* **183**, 504–521 (Dec. 2018).

[2] Samper-Gonzalez, J., Burgos, N., Fontanella, S., Bertin, H., Habert, M.-O., Durrleman, S., Evgeniou, T., and Colliot, O., "Yet Another ADNI Machine Learning Paper? Paving The Way Towards Fully-reproducible Research on Classification of Alzheimer's Disease," 8 (Sept. 2017).

[3] Bondi, M. W., Jak, A. J., Delano-Wood, L., Jacobson, M. W., Delis, D. C., and Salmon, D. P., "Neuropsychological contributions to the early identification of Alzheimer's disease," *Neuropsychology Review* **18**, 73–90 (Mar. 2008).

[4] Ewers, M., Sperling, R. A., Klunk, W. E., Weiner, M. W., and Hampel, H., "Neuroimaging markers for the prediction and early diagnosis of Alzheimer's disease dementia," *Trends in Neurosciences* **34**, 430–442 (Aug. 2011).

[5] Coup, P., Eskildsen, S. F., Manjn, J. V., Fonov, V. S., Pruessner, J. C., Allard, M., and Collins, D. L., "Scoring by nonlocal image patch estimator for early detection of Alzheimer's disease," *NeuroImage: Clinical* **1**, 141–152 (Jan. 2012). 00058.

[6] Hett, K., Ta, V.-T., Manjn, J. V., Coup, P., and Alzheimer's Disease Neuroimaging Initiative, "Adaptive fusion of texture-based grading for Alzheimer's disease classification," *Computerized Medical Imaging and Graphics: The Official Journal of the Computerized Medical Imaging Society* **70**, 8–16 (Dec. 2018).

[7] Cuingnet, R., Gerardin, E., Tessieras, J., Auzias, G., Lehricy, S., Habert, M.-O., Chupin, M., Benali, H., and Colliot, O., "Automatic classification of patients with Alzheimer's disease from structural MRI: A comparison of ten methods using the ADNI database," *NeuroImage* **56**, 766–781 (May 2011).

[8] Suk, H.-I., Lee, S.-W., Shen, D., and Alzheimers Disease Neuroimaging Initiative, "Deep ensemble learning of sparse regression models for brain disease diagnosis," *Medical Image Analysis* **37**, 101–113 (Apr. 2017). 00015.

[9] Misra, C., Fan, Y., and Davatzikos, C., "Baseline and longitudinal patterns of brain atrophy in MCI patients, and their use in prediction of short-term conversion to AD: results from ADNI," *NeuroImage* **44**, 1415–1422 (Feb. 2009).

[10] Adaszewski, S., Dukart, J., Kherif, F., Frackowiak, R., Draganski, B., and Alzheimer's Disease Neuroimaging Initiative, "How early can we predict Alzheimer's disease using computational anatomy?," *Neurobiology of Aging* **34**, 2815–2826 (Dec. 2013).

[11] Liu, X., Tosun, D., Weiner, M. W., Schuff, N., and Alzheimer's Disease Neuroimaging Initiative, "Locally linear embedding (LLE) for MRI based Alzheimer's disease classification," *NeuroImage* **83**, 148–157 (Dec. 2013).

[12] Eskildsen, S. F., Coup, P., Garca-Lorenzo, D., Fonov, V., Pruessner, J. C., Collins, D. L., and Alzheimer's Disease Neuroimaging Initiative, "Prediction of Alzheimer's disease in subjects with mild cognitive impairment from the ADNI cohort using patterns of cortical thinning," *NeuroImage* **65**, 511–521 (Jan. 2013). 00122.

[13] Costafreda, S. G., Dinov, I. D., Tu, Z., Shi, Y., Liu, C.-Y., Kloszewska, I., Mecocci, P., Soininen, H., Tsolaki, M., Vellas, B., Wahlund, L.-O., Spenger, C., Toga, A. W., Lovestone, S., and Simmons, A., "Automated hippocampal shape analysis predicts the onset of dementia in Mild Cognitive Impairment," *NeuroImage* **56**, 212–219 (May 2011).

[14] Gray, K. R., Wolz, R., Heckemann, R. A., Aljabar, P., Hammers, A., Rueckert, D., and Alzheimer's Disease Neuroimaging Initiative, "Multi-region analysis of longitudinal FDG-PET for the classification of Alzheimer's disease," *NeuroImage* **60**, 221–229 (Mar. 2012).

[15] Cabral, C., Morgado, P. M., Campos Costa, D., Silveira, M., and Alzheimers Disease Neuroimaging Initiative, "Predicting conversion from MCI to AD with FDG-PET brain images at different prodromal stages," *Computers in Biology and Medicine* **58**, 101–109 (Mar. 2015). 00022.

[16] Davatzikos, C., Bhatt, P., Shaw, L. M., Batmanghelich, K. N., and Trojanowski, J. Q., "Prediction of MCI to AD conversion, via MRI, CSF biomarkers, and pattern classification," *Neurobiology of Aging* **32**, 2322.e19–27 (Dec. 2011).

[17] Tang, X., Holland, D., Dale, A. M., Younes, L., and Miller, M. I., "Baseline Shape Diffeomorphometry Patterns of Subcortical and Ventricular Structures in Predicting Conversion of Mild Cognitive Impairment to Alzheimers Disease," *Journal of Alzheimer's disease : JAD* **44**(2), 599–611 (2015).

[18] Cui, Y., Liu, B., Luo, S., Zhen, X., Fan, M., Liu, T., Zhu, W., Park, M., Jiang, T., Jin, J. S., and Alzheimer's Disease Neuroimaging Initiative, "Identification of conversion from mild cognitive impairment to Alzheimer's disease using multivariate predictors," *PloS One* **6**(7), e21896 (2011).

[19] Srensen, L., Igel, C., Liv Hansen, N., Osler, M., Lauritzen, M., Rostrup, E., Nielsen, M., and Alzheimer's Disease Neuroimaging Initiative and the Australian Imaging Biomarkers and Lifestyle Flagship Study of Ageing, "Early detection of Alzheimer's disease using MRI hippocampal texture," *Human Brain Mapping* **37**, 1148–1161 (Mar. 2016).

[20] Chincarini, A., Bosco, P., Calvini, P., Gemme, G., Esposito, M., Olivieri, C., Rei, L., Squarcia, S., Rodriguez, G., Bellotti, R., Cerello, P., De Mitri, I., Retico, A., Nobili, F., and Alzheimer's Disease Neuroimaging Initiative, "Local MRI analysis approach in the diagnosis of early and prodromal Alzheimer's disease," *NeuroImage* **58**, 469–480 (Sept. 2011).

[21] Moradi, E., Pepe, A., Gaser, C., Huttunen, H., Tohka, J., and Alzheimer's Disease Neuroimaging Initiative, "Machine learning framework for early MRI-based Alzheimer's conversion prediction in MCI subjects," *NeuroImage* **104**, 398–412 (Jan. 2015). 00119.

[22] Cheng, B., Liu, M., Zhang, D., Munsell, B. C., and Shen, D., "Domain Transfer Learning for MCI Conversion Prediction," *IEEE Transactions on Biomedical Engineering* **62**, 1805–1817 (July 2015).

[23] Young, J., Modat, M., Cardoso, M. J., Mendelson, A., Cash, D., and Ourselin, S., "Accurate multimodal probabilistic prediction of conversion to Alzheimer's disease in patients with mild cognitive impairment," *NeuroImage: Clinical* **2**, 735–745 (Jan. 2013).

[24] Suk, H.-I., Lee, S.-W., and Shen, D., "Hierarchical feature representation and multimodal fusion with deep learning for AD/MCI diagnosis," *NeuroImage* **101**, 569–582 (Nov. 2014).

[25] Li, F., Tran, L., Thung, K., Ji, S., Shen, D., and Li, J., "A Robust Deep Model for Improved Classification of AD/MCI Patients," *IEEE Journal of Biomedical and Health Informatics* **19**, 1610–1616 (Sept. 2015).

[26] Choi, H. and Jin, K. H., "Predicting cognitive decline with deep learning of brain metabolism and amyloid imaging," *Behavioural Brain Research* **344**, 103–109 (May 2018).

[27] Arbabshirani, M. R., Plis, S., Sui, J., and Calhoun, V. D., "Single subject prediction of brain disorders in neuroimaging: Promises and pitfalls," *NeuroImage* **145**, 137–165 (Jan. 2017).

[28] Falahati, F., Westman, E., and Simmons, A., "Multivariate data analysis and machine learning in Alzheimer's disease with a focus on structural magnetic resonance imaging," *Journal of Alzheimer's disease: JAD* **41**(3), 685–708 (2014).

[29] Rathore, S., Habes, M., Iftikhar, M. A., Shacklett, A., and Davatzikos, C., "A review on neuroimaging-based classification studies and associated feature extraction methods for Alzheimer's disease and its prodromal stages," *NeuroImage* **155**, 530–548 (July 2017).

[30] Gorgolewski, K. J., Auer, T., Calhoun, V. D., Craddock, R. C., Das, S., Duff, E. P., Flandin, G., Ghosh, S. S., Glatard, T., Halchenko, Y. O., Handwerker, D. A., Hanke, M., Keator, D., Li, X., Michael, Z., Maumet, C., Nichols, B. N., Nichols, T. E., Pellman, J., Poline, J.-B., Rokem, A., Schaefer, G., Sochat, V., Triplett, W., Turner, J. A., Varoquaux, G., and Poldrack, R. A., "The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments," *Scientific Data* **3**, 160044 (June 2016).

[31] Routier, A., Guillon, J., Burgos, N., Samper-Gonzalez, J., Wen, J., Fontanella, S., Bottani, S., Jacquemont, T., Marcoux, A., Gori, P., Lu, P., Moreau, T., Bacci, M., Durrleman, S., and Colliot, O., "Clinica: an open source software platform for reproducible clinical neuroscience studies," (June 2018).

[32] Landau, S. M., Breault, C., Joshi, A. D., Pontecorvo, M., Mathis, C. A., Jagust, W. J., Mintun, M. A., and Alzheimers Disease Neuroimaging Initiative, "Amyloid- imaging with Pittsburgh compound B and florbetapir: comparing radiotracers and quantification methods," *Journal of Nuclear Medicine: Official Publication, Society of Nuclear Medicine* **54**, 70–77 (Jan. 2013).

[33] Ashburner, J., "A fast diffeomorphic image registration algorithm," *NeuroImage* **38**, 95–113 (Oct. 2007).

[34] Ashburner, J. and Friston, K. J., "Unified segmentation," *NeuroImage* **26**, 839–851 (July 2005).

[35] Dukart, J., Schroeter, M. L., Mueller, K., and Alzheimer's Disease Neuroimaging Initiative, "Age correction in dementia–matching to a healthy brain," *PloS One* **6**(7), e22193 (2011).

[36] Varoquaux, G., Raamana, P. R., Engemann, D. A., Hoyos-Idrobo, A., Schwartz, Y., and Thirion, B., "Assessing and tuning brain decoders: Cross-validation, caveats, and guidelines," *NeuroImage* **145**, 166–179 (Jan. 2017).