Gómez-Sancho, M., Tohka, J. & Gómez-Verdejo, V. (2018). Comparison of feature representations in MRI-based MCI-to-AD conversion prediction. *Magnetic Resonance Imaging, 50*, 84–95.

DOI: 10.1016/j.mri.2018.03.003

Comparison of feature representations in MRI-based MCI-to-AD conversion prediction

Marta Gómez-Sancho[a], Jussi Tohka[b,*], Vanessa Gómez-Verdejo[a,*], for the Alzheimer's Disease Neuroimaging Initiative[c]

[a] *Department of Signal Processing and Communications, Universidad Carlos III de Madrid, Leganes, Spain*

[b] *University of Eastern Finland, AI Virtanen Institute for Molecular Sciences, Kuopio, Finland*

[c] *Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (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*

Abstract

Alzheimer's Disease (AD) is a progressive neurological disorder in which the death of brain cells causes memory loss and cognitive decline. The identification of at-risk subjects yet showing no dementia symptoms but who will later convert to AD can be crucial for the effective treatment of AD. For this, Magnetic Resonance Imaging (MRI) is expected to play a crucial role. During recent years, several Machine Learning (ML) approaches to AD-conversion prediction have been proposed using different types of MRI features. However, few studies comparing these different feature representations exist, and the existing ones do not allow to make definite conclusions. We evaluated the performance of various types of MRI features for the conversion prediction: voxel-based features extracted based on voxel-based morphometry, hippocampus volumes, volumes of the entorhinal cortex, and a set of regional volumetric, surface area, and cortical thickness measures across the brain. Regional features consistently yielded the best performance over two classifiers (Support Vector Machines and Regularized Logistic Regression), and two datasets studied. However, the performance

*These two authors share the senior authorship.

**Corresponding author:vanessa@tsc.uc3m.es.

November 29, 2021

difference to other features was not statistically significant. There was a consistent trend of age correction improving the classification performance, but the improvement reached statistical significance only rarely.

Keywords: Alzheimer's Disease, Magnetic Resonance Imaging, Brain, Machine Learning, Feature Representations

1. Introduction

Alzheimer's Disease (AD) is a progressive neurological disorder in which the death of brain cells causes memory loss and cognitive decline. The progression of the neuropathology in AD starts long before clinical symptoms of the disease become apparent [1, 2, 3, 4, 5]. Also, the symptoms become progressively worse, and much effort has been placed on the early diagnosis of the AD. Related to this, Mild Cognitive Impairment (MCI), defined as a transitional phase from cognitive changes of normal aging to those typically found in dementia, is an important construct [6]. Subjects with MCI present a high risk of developing AD, but still, most people with MCI will not progress to dementia (or AD) even after 10 years of follow-up [7, 8]. Thus, identifying MCI subjects who convert to AD can be crucial for the effective treatment of AD.

Neuroimaging techniques have shown promise as tools for presymptomatic AD detection [9, 10]. Much research has been focused on T1-weighted Magnetic Resonance Imaging (MRI). It is one of the most widely studied imaging techniques [11] because it is completely non-invasive, highly available, inexpensive compared to positron emission tomography and has an excellent contrast between different soft tissues. Over the past few years, many potential MRI markers, such as the whole brain, hippocampal, and entorhinal cortex atrophy, have been shown to have diagnostic value [12]. Also, these markers have been used as the features for Machine Learning (ML) algorithms trying to predict MCI-to-AD conversion.

Indeed, there has been a surge of proposed ML algorithms for automatically predicting the future conversion from MCI to AD based on MRI (e.g.,

[13, 14, 15, 16]). This is partly driven by the free availability of large, high-quality datasets such as ADNI [1]. However, the principal focus has been in the development of new ML techniques, and their comparative evaluation has received much less attention. In particular, ML algorithms have used different types of feature sets extracted from MRI, including hippocampal volumes, volumes of the entorhinal cortex, cortical thickness measures, as well as voxel-based morphometry (VBM) features (e.g., [17, 18, 19, 20, 21] and [22] for a recent review). Despite that, systematic studies of advantages/disadvantages of various feature sets have been limited so far, and existing studies do not allow to make definite conclusions. To add to the confusion, high dimensional feature sets, such as cortical thickness or voxel-based morphometry, must be coupled with dimensionality reduction technique such as averaging the values within a brain region, Principal Component Analysis (PCA) or feature selection (see [23] for a review).

Existing comparisons between different feature representations do not provide a clear answer to the question we are interested in: "Is there a preferred representation of MRI for AD-conversion prediction?". There are multiple reasons for this. The comparisons have been geared to the AD vs. control classification problem ([24, 25, 26]), they have not included voxel-based representations [24, 27], they have utilized very short follow-up (18-months [27, 28]), they have been based on a single learning algorithm [27, 29] and/or have had highly unbalanced pMCI and sMCI classes (in [29] 149 of 165 MCI subjects converted during the 4-year follow-up that is in stark contrast to conversion rates reported in other analyses [8]). An early and important study [28], which we want to highlight, compared various feature representations including hippocampal volumes, cortical thickness, and VBM with and without regional averaging. No feature representation in this study managed to perform significantly better than chance. This somewhat disappointing result could be because 1) the methods were early ones, mostly geared to the much easier normal control vs. AD sub-

[1] Information and data can be found at `adni.loni.usc.edu`

ject classification problem, 2) the dataset was smaller than the one currently available, and 3) the MCI non-converter was somewhat arbitrarily defined as a subject who did not convert in 18 months period. Moradi et al. [30] evaluated their method over the same dataset as [28] managing to obtain significantly better performance than the chance level, pointing to the reason 1) as the most significant cause of the improvement.

Since [28], we can a find few studies of different feature representations presenting partially conflicting results. As an example, [21] found that the prognostic efficacy of hippocampus volumetry was better than combined regional volumetrics in two commercially available brain volumetric software packages for MCI conversion prediction. On the other hand, Gaser et al. [17] have demonstrated the superiority of their voxel-based brainAGE approach over the hippocampus volume biomarker and Westman et al. have emphasized the importance of having a complete set of regional features [27, 24]. Some researchers have opted to study feature selection, either supporting [20, 31] or opposing [32] data-driven feature selection. The comparisons of different automatic algorithms for hippocampal [33] and entorhinal cortex volumetry [34] have indicated that the algorithm-choice did not affect the classification accuracy. Intracranial volume adjustment to regional volumetry appears to only have subtle effects to the conversion prediction accuracy [35, 27]. Finally, it has been demonstrated that the neurophychological test scores are the best predictors of conversion, but combining them with MRI information leads to improved prediction accuracy [30, 36].

To close this information gap, we asked what type of feature representations are the best for the MRI-based AD-conversion prediction. We used follow-up period of 3 years to define the AD-conversion, twice longer than in [27, 28]. We evaluated the performance of various MRI features, including VBM-style, voxel-based features [17], coupled with feature preselection [30] or PCA-based dimensionality reduction, hippocampus volumes, volumes of the entorhinal cortex, and a complete set of regional volumetry, surface area, and cortical thickness measures extracted by FreeSurfer. This complements earlier studies which did

not include voxels -based representations [27, 21]. We additionally evaluated age removal [30, 37], which have been found to improve the prognostic efficacy of ML-based MRI biomarkers. Moreover, we used two different classifiers (Support Vector Machines, SVM, and Regularized Logistic Regression, RLR) for reducing the classifier specificity of the conclusions and trained them applying a repeated 10-fold cross-Validation (CV) with a sound statistical inference to compare the methods, which can be seen as an improvement of separate training and test sets in [28].

2. Material and methods

ADNI data

Data is collected from the the Alzheimers Disease Neuroimaging Initiative (ADNI) public database[2]. The ADNI initiative was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimers disease (AD)[3].

ADNI material considered in this work included all subjects from ADNI1 for whom baseline MRI data (T1-weighted MP-RAGE sequence at 1.5 Tesla, typically 256 x 256 x 170 voxels with the voxel size of approximately 1 mm x 1 mm x 1.2 mm) and sufficient follow-up information were available. We focused on the classification of MCI individuals based on their future diagnosis (AD or not AD) and, therefore, MRI scans were all obtained at the baseline visit.

Two flavors of this dataset were evaluated. The first dataset, Quality Control (QC) dataset, included 183 MCI subjects whose FreeSurfer 4.3 MRI segmentations had passed the complete quality control. The second one, Non QC dataset,

[2]Available at `adni.loni.usc.edu`.

[3] For up-to-date information, see `www.adni-info.org`.

included the complete dataset of 264 MCI subjects without any quality control. The reason for evaluating the two different sets was to study if the quality control yielded an improvement in the data analysis. Note that the QC dataset was a subset of the non-QC dataset.

Following [30], a subject was considered as a progressive MCI (pMCI) if diagnosed as MCI at baseline and the diagnosis changed to AD during the 3-year follow-up period. The subject was considered as a stable MCI (sMCI) if diagnosed as MCI at baseline and the diagnosis remained as MCI during the follow-up. The minimum length of follow-up was 3 years and the subject was excluded from the study if she converted after the 3 year follow-up, the diagnosis fluctuated after the 3-year follow-up period, or less than 3-years of follow-up information was available. Table 1 lists the main characteristics of the subjects of each dataset and the list of Roster IDs of the included subjects and their diagnostic categories are available in the supplement.

Table 1: Demographics of the two flavors of the dataset (QC and Non-QC) used in this work. The NC and AD subjects' data were not used directly in the learning algorithms. The NC subjects were used for the age-correction. The AD and NC subjects were used in Moradi-method for feature selection.

	QC dataset				**Non-QC dataset**			
	sMCI	**pMCI**	**AD**	**NC**	**sMCI**	**pMCI**	**AD**	**NC**
No. subjects	73	110	126	182	100	164	200	231
Males / Females	46/27	58/52	65/61	91/91	66/34	97/67	103/97	119/112
Age range	59-88	55-89	55-91	60-90	57-88	55-89	55-91	60-90

2.1. Image preprocessing

Table 2 details the feature representations we investigated and their respective number of features. Hippocampus volumes consisted of left and right hippocampal volumes. Hippocampus + Entorhinal volumes consisted of left and right volumes of hippocampus and left and right volumes of entorhinal cortex. We considered both raw volumes as well as volumes normalized by the intracranial volume (ICV) as it is still unclear if the normalization by ICV is beneficial

for the prediction task [35, 27]. Region based features included a complete set of 257 regional cortical thickness, surface area and volume measures provided by FreeSurfer [4] , [5]. We note that this set of features included also ICV.

Freesurfer 4.3 software was employed for the extraction of hippocampus and entorhinal cortex volumes as well as region features. Particularly, the FreeSurfer 4.3 processing results available at the ADNI website were used (UCSF Cross-sectional Freesurfer version 4.3), and the description of the pipeline and the QC procedure can be found at [6]. The rationales for using the processing results provided by ADNI was to ensure that the processing pipeline was a standard one, the processing results are readily available to other researchers, and the quality control, independent from the authors of this study, has been performed. We note that albeit different versions of FreeSurfer can result in different segmentations, the classification results based on different software versions have been found to be the same [34].

Voxel-Based Morphometry (VBM) based features consisted of 29852 gray matter density values from the VBM style preprocessing by the VBM8 software. In brief, the MRIs were preprocessed into gray matter tissue images in the stereotactic space as described in [17, 30], smoothed with the 8-mm FWHM Gaussian kernel, resampled to 4 mm spatial resolution, and masked into 29852 voxels. In the Moradi set of features, VBM features were further processed through the feature selection method of [30]. This method applies MRIs of AD and NC subjects to select features for MCI classification through a repeated application of the elastic net penalized linear regression. We applied the ADNI data from 231 (182) normal controls and 200 (126) AD subjects for this feature

[4] `http://surfer.nmr.mgh.harvard.edu/`

[5] Originally, this set included 274 measures. We selected a subset of 256 regions from the aforementioned 274 measures discarding the regions that presented missed data. A more detailed description of the 256 features is provided in `https://github.com/MartaGomez/Regions-list-/wiki/Regions-list`.

[6] `https://adni.bitbucket.io/reference/docs/UCSFFRESFR/UCSFFreeSurferMethodsSummary.pdf`

selection with the non-QC (QC dataset). We reduced the number of VBM features also using principal component analysis (PCA). For this, we retained the PCA components that explained 90 % of the variance (see Tables 9, 10 and 11 of Supplementary Material for a performance comparison with different variance thresholds).

Table 2: Summary of the sets of features considered in this study. Note that the number of Moradi and PCA voxel features was dataset dependent.

Feature Set	**Number of features**
Hippocampus volumes	2
Hippocampus + Entorhinal volumes	4
Region	257
Voxel	29852
Moradi	525 (non-QC); 431 (QC)
PCA Voxel	225 (non-QC); 157 (QC)

We further evaluated the representations with and without the age correction. The age correction may be important as the effects of normal aging on the brain structure partially overlap with the effects of AD [38, 37]. We applied the age correction procedure of [30]. This method estimates the age effect by a linear regression for each feature separately based on the MRIs of normal controls (231 normal controls with the age range from 55 to 90 years of ADNI) and then adjusts the features of the MCI subjects based on the estimated model.

2.2. Validation and test procedure

For the implementation and evaluation of the classification methods, we performed a repeated and nested 10-fold Cross Validation (CV). In the outer CV loop, data was split in 10 different folds from which one fold at time was designated as the test fold (for performance evaluation) and the nine remaining folds were used for classifier training. The train/test cycle was repeated with each fold once as the test fold. In the inner CV loop, each train fold was, itself,

split into 10 validation folds from which one part was used to select the classifier hyperparameters. The optimal hyperparameters were selected evaluating either the classification accuracy (ACC, number of correctly classified samples over the total number of samples) or the Area Under the receiving operating Curve (AUC) [39]. It has been suggested that AUC has key advantages over ACC as a model selection criterion [40]. The nested CV was repeated 10 times, each with a different randomly selected folding scheme, to minimize the effect of a particular folding scheme to the results. Also, the hypothesis test we used to compare different representations requires the repeated use of CV.

To study the classifier performance, we considered several metrics: AUC, ACC, Sensitivity (SEN, number of correctly classified pMCI subjects divided by the total number of pMCI subjects) and Specificity (SPE, number of correctly classified sMCI subjects divided by the total number of sMCI subjects). We selected AUC as our principal performance measure as it is insensitive to the class-imbalance whereas ACC can be strongly affected by the class-imbalance.

2.3. Classifiers

To evaluate each feature set, we considered two types of widely used supervised learning classifiers: Support Vector Machine (SVM) [41] and elastic-net Regularized Logistic Regression (RLR) [42]. Accessible description of these learning methods can be found in [43]. For the SVM implementation, we used the Python open source library Scikit-learn[7], which is based on the LIBSVM implementation[8]. For the RLR classifiers, we applied the GLMNET Python library[9], which solves the resulting penalized optimization problem by a coordinate descent algorithm. We note that both of these learning algorithms tolerate high-dimensional data via regularization and are therefore suited for the cases where the number of features is higher than the number of subjects. Especially, elastic-net includes an L1-penalty, which leads to feature selection embedded to

[7] `http://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html`
[8] `https://www.csie.ntu.edu.tw/~cjlin/libsvm/`
[9] `https://web.stanford.edu/~hastie/glmnet_python/`

the classifier learning [44]. A large majority of supervised learning techniques have utilized these learning algorithms [22] and a comparison of different classification algorithms MCI-to-AD prediction is available in [45]. We note that we did not include Random Forests [46] as these are not straight-forwardly suitable for high-dimensional small-sample problems and the computation time and memory requirements for nearly all implementations would be prohibiting for the voxel-based features (however, see [47]).

For the case of the SVM classifier, we decided to use the linear SVM (we have also analyzed the possibility of using a RBF (Radial Basis Function) kernel, however, experimental results showed similar performances). In this way, we had to select only the soft margin parameter, C, whose value was explored among the set $\{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1},\ 1, 10, 10^2, 10^3\}$ (see [41] for notation). Despite considering the linear SVM, its implementation was carried out in the dual space, precomputing a linear kernel; in this way, we simplified the calculations and reduced the computation time with the high dimensional feature representations, such as VBM ones.

For the RLR classifier, using the notation of [42], we set the parameter of the elastic net α to 0.5, just in between lasso (α=1) and ridge ($\alpha = 0$) regularization. The principal regularization parameter of the RLR (λ), which sets the balance between the regularization and the data terms, was chosen among the set of values $\{10^{-10}, 10^{-4}, 10^{-3}, 5 \cdot 10^{-3}, 10^{-2}, 5 \cdot 10^{-2}, 10^{-1}, 5 \cdot 10^{-1}\}$. We previously demonstrated that selecting also the parameter α by cross-validation did not yield advantages over fixing its value to 0.5 [31]. However, we confirmed that this is the case with the setup of this paper by experimenting with different values of α (see Table 12 of the Supplementary Material).

Finally, as a step prior to training the classifiers, we normalized the data by removing its mean and scaling it to the unit variance.

2.4. Statistical test

To compare the AUC values provided by different approaches, we applied the corrected resampled t-test [48]. The problem in applying standard statis-

tical methodology, such as uncorrected t-test to assess the differences between AUCs is that $r \times k$ AUC values from in a k-fold CV repeated r times are not statistically independent. Instead, the corrected resampled t-test assumes dependency among the AUCs in a k-fold CV repeated r times and, therefore, it allows to statistically compare two mean AUC values by correcting the variance estimation. The corrected resampled t-test can be seen as an improvement over the 5x2 CV of [49] and McNemar's test for the classification accuracy [48]. Although the test was developed for the classification accuracy, it is as well applicable for testing the differences between AUCs.

To describe the test formally, let n_1 and n_2, respectively, denote the number of instances used for training and testing in each fold, and a_{ij} and b_{ij} represent the AUCs of the i-th fold and j-th run of the method A and B with $i = 1, ..., k$ and $j = 1, ..., r$. Denoting the estimated mean and variance values of the differences between methods A and B by $\hat{m}$ and $\hat{\sigma}^2$ i.e.,

$$\hat{m} = \frac{1}{kr} \sum_{i=1}^{k} \sum_{j=1}^{r} a_{ij} - b_{ij} \tag{1}$$

$$\hat{\sigma}^2 = \frac{1}{kr - 1} \sum_{i=1}^{k} \sum_{j=1}^{r} (a_{ij} - b_{ij} - \hat{m})^2 \tag{2}$$

we can estimate the statistic of the test, t, as:

$$t = \frac{\hat{m}}{\sqrt{\left(\frac{1}{kr} + \frac{n_2}{n_1}\right) \hat{\sigma}^2}} \tag{3}$$

The statistic t follows a student's t-distribution with $kr - 1$ degrees of freedom. In our case, $r = k = 10$.

3. Results

Tables 3 – 8 show the results for the QC dataset and non-QC datasets using the AUC for model selection and the results of the non-QC dataset when the best model was selected using ACC. In particular, each table includes for the SVM and RLR classifiers the values of AUC, ACC, SEN, SPE, as well as three

Table 3: Cross-validated SVM performance measures with the QC dataset using AUC as the model selection criterion. Hippocampus (Hippo. and entor. vol.) volumes ICV refers to hippocampus (hippocampus + entorhinal) volumes normalized by the ICV.

Feature Set	**Age Removal**	**AUC (%)**	**ACC (%)**	**SEN (%)**	**SPE (%)**	p_{Age}	p_{Hippo}	p_{Class}
Hippocampus	No	73.50 ± 12.61	63.15 ± 7.27	93.00 ±10.90	18.16 ±25.35			0.873
Hippocampus	Yes	77.31 ± 11.22	66.42 ± 9.67	90.54 ±11.78	30.20 ±31.18	0.052		0.819
Hippocampus vol. ICV	No	69.61 ± 12.60	67.21 ± 9.88	81.36 ±9.91	45.82 ±14.42			0.690
Hippocampus vol. ICV	Yes	72.67 ± 11.53	68.31 ± 9.60	82.36 ±9.43	47.11 ±14.48	0.124	0.169	0.416
Hippo. and entor.	No	**75.91** ± 11.84	63.69 ± 7.91	94.55 ±9.09	17.14 ±25.78			0.314
Hippo. and entor.	Yes	**78.59%** ± 10.84	61.77 ± 5.64	97.82 ±7.02	7.29 ±21.19	0.055	0.492	0.285
Hippo. & entor. vol. ICV	No	72.66 ± 11.26	68.26 ± 10.27	83.82 ±10.17	44.79 ±16.58			0.850
Hippo. & entor. vol. ICV	Yes	74.50 ± 10.65	68.86 ± 9.84	81.64 ±9.70	49.66 ± 15.61	0.325	0.399	0.505
Voxel features	No	63.19 ± 12.21	60.51 ± 5.66	91.36 ± 14.74	13.98 ± 21.14			0.649
Voxel features	Yes	66.67 ± 11.71	61.65 ± 5.69	91.82 ±12.23	16.20 ±17.19	0.175	0.042	0.923
PCA VF 90	No	64.33 ± 11.63	60.11 ± 5.16	90.36 ± 14.48	14.38 ± 20.84			0.791
PCA VF 90	Yes	66.78 ± 11.67	61.09 ± 5.74	91.82 ±14.35	14.64 ±22.43	0.351	0.042	0.569
Moradi features	No	71.92 ± 10.62	63.54 ± 7.69	92.63 ±10.91	19.66 ±23.45			0.025
Moradi features	Yes	75.08 ± 9.80	62.32 ± 5.60	97.10 ±7.03	9.86 ±20.64	0.243	0.610	0.001
Region features	No	74.06 ± 12.21	64.67 ± 7.29	92.27 ± 11.31	22.91 ± 24.08			0.739
Region features	Yes	77.34 ± 10.22	69.29 ± 8.96	88.45 ±10.92	37.27 ±26.62	0.241	0.990	0.759

Table 4: Cross-validated RLR performance measures with the QC dataset using AUC as the model selection criterion. Hippocampus (Hippo. and entor. vol.) volumes ICV refers to hippocampus (hippocampus + entorhinal) volumes normalized by the ICV.

Feature Set	**Age Removal**	**AUC (%)**	**ACC(%)**	**SEN(%)**	**SPE(%)**	p_{Age}	p_{Hippo}	p_{Class}
Hippocampus	No	73.38 ± 12.72	68.41 ± 10.22	82.00 ±11.85	47.86 ± 17.66			
Hippocampus	Yes	77.19 ± 11.15	71.31 ± 9.16	83.91 ±10.82	52.43 ±16.82	0.042		
Hippocampus vol. ICV	No	69.80 ± 12.67	66.57 ± 9.47	75.16 ±11.87	38.88 ±18.27			
Hippocampus vol. ICV	Yes	72.25 ± 11.66	68.32 ± 9.29	84.64 ±11.26	43.68 ±16.23	0.253	0.138	
Hippo. and entor.	No	72.52 ± 12.18	66.85 ± 9.54	80.82 ±10.33	45.73 ±18.48			
Hippo. and entor.	Yes	74.13 ± 11.28 %	68.05 ± 9.62	81.64 ±10.40	47.55 ±18.97	0.085	0.767	
Hippo. & entor. vol. ICV	No	71.72 ± 11.38	68.59 ± 9.89 %	84.47 ± 11.85	42.60 ± 17.11			
Hippo. & entor. vol. ICV	Yes	73.60 ± 10.83 %	68.90 ± 9.23	83.51 ±11.50	45.00 ±18.30	0.332	0.348	
Voxel features	No	64.40 ± 11.95 %	61.91 ± 9.86	76.18 ±13.22	40.43 ±15.45			
Voxel features	Yes	66.94 ± 11.15	63.57 ± 9.48	77.36 ± 13.20	40.43 ± 16.37	0.268	0.029	
PCA VF 90	No	63.63 ± 11.20	60.88 ± 8.33	79.45 ±14.00	33.00 ±18.91			
PCA VF 90	Yes	65.35 ± 12.33	61.28 ± 8.44	81.73 ±14.46	30.43 ±19.14	0.573	0.018	
Moradi features	No	65.83 ± 11.36	63.28 ± 9.68	75.00 ±13.01	42.73 ±16.84			
Moradi features	Yes	67.89 ± 10.43	65.52 ± 9.71	77.54 ±13.69	57.50 ±17.24	0.348	0.038	
Region features	No	**74.57** ± 10.92	70.67 ± 8.92	81.91 ±10.40	53.70 ±17.31			
Region features	Yes	**77.91** ± 9.59	71.93 ± 9.22	81.45 ±10.90%	57.50 ±18.26	0.171	0.839	

Table 5: Cross-validated SVM performance measures with the non QC dataset using AUC as the model selection criterion. Hippocampus (Hippo. and entor. vol.) volumes ICV refers to hippocampus (hippocampus + entorhinal) volumes normalized by the ICV.

Feature Set	**Age Removal**	**AUC (%)**	**ACC (%)**	**SEN (%)**	**SPE (%)**	p_{Age}	p_{Hippo}	p_{Class}
Hippocampus volumes	No	70.29 ± 10.16	63.09 ± 4.18	96.02 ± 8.14	9.10 ±15.82			0.708
Hippocampus volumes	Yes	75.57 ± 8.41	63.94 ± 4.36	95.39 ±8.42	12.30 ±21.06	0.047		0.398
Hippocampus vol. ICV	No	69.56 ± 10.16 %	68.25 ± 6.96	85.60 ±7.25	39.80 ±12.00			0.987
Hippocampus vol. ICV	Yes	72.28 ± 9.97	69.46 ± 7.46	85.3 ±7.48	43.50 ±12.03	0.172	0.181	0.807
Hippo. and entor. vol.	No	73.23 ± 9.37	64.57 ± 1.74	98.16 ±6.08	3.50 ±11.78			0.358
Hippo. and entor. vol.	Yes	76.05 ± 8.76	63.73 ± 5.13	96.86 ±7.34	9.50 ±19.97	0.088	0.744	0.253
Hippo. & entor. vol. ICV	No	71.75 ± 10.00	68.55 ± 7.81	84.36 ±7.87	42.60 ±14.04			0.948
Hippo. & entor. vol. ICV	Yes	73.77 ± 9.94	69.60 ± 8.24	82.95 ±8.26	47.90 ±15.05	0.194	0.495	0.756
Voxel based features	No	68.55 ± 10.17	62.11 ± 3.02	97.10 ±8.99	4.80 ±12.53			0.695
Voxel based features	Yes	69.55 ± 10.67	63.79 ± 4.59	96.55 ±9.73	10.10 ±14.46	0.659	0.169	0.489
PCA VF 90	No	67.94 ± 10.01	62.11 ± 3.81	93.26 ±11.92	11.00 ±18.79			0.923
PCA VF 90	Yes	68.38 ± 11.49	62.50 ± 4.49	92.39 ± 12.56	13.50 ±19.97	0.194	0.117	0.915
Moradi features	No	**73.26** ± 10.35	64.20 ± 8.04	96.04 ±9.58	12.00 ±15.40			0.078
Moradi features	Yes	75.72 ± 10.35	63.40 ± 8.04	96.89 ±9.58	8.50 ±15.40	0.180	0.965	0.288
Region features	No	73.11 ± 10.01	65.29 ± 5.89	90.07 ±11.04	24.60 ±22.95			0.123
Region features	Yes	**76.89** ± 9.12	66.94 ± 7.36	95.30 ±7.86	20.50 ±22.41	0.091	0.680	0.101

Table 6: Cross-validated RLR performance measures with the non QC dataset using AUC as the model selection criterion. Hippocampus (Hippo. and entor. vol.) volumes ICV refers to hippocampus (hippocampus + entorhinal) volumes normalized by the ICV.

Feature Set	**Age Removal**	**AUC (%)**	**ACC (%)**	**SEN (%)**	**SPE (%)**	p_{age}	p_{Hippo}	p_{Class}
Hippocampus volumes	No	70.95 ± 8.89	67.00 ± 7.29	88.04 ±10.43	32.40 ±15.24			
Hippocampus volumes	Yes	74.95 ± 8.25	67.68 ± 7.00	86.34 ±8.98	37.00 ±13.60	0.046		
Hippocampus vol. ICV	No	69.60 ± 10.41	67.94 ± 6.88	87.91 ±8.49%	35.20 ±15.84			
Hippocampus vol. ICV	Yes	72.37 ± 10.03	69.44 ± 7.47	88.76 ±8.12	37.80 ±15.20	0.148	0.272	
Hippo. and entor. vol.	No	72.59 ± 9.23	69.72 ± 7.86	85.88 ±9.59%	43.20 ±15.35			
Hippo. and entor. vol.	Yes	75.31 ± 9.03	70.61 ± 7.47	84.98 ±9.94	47.00 ±14.93	0.094	0.801	
Hippo. & entor. vol. ICV	No	71.72 ± 9.76	68.59 ± 8.08	84.47 ±9.24	42.60 ±16.71			
Hippo. & entor. vol. ICV	Yes	73.60 ± 9.93	68.90 ± 7.63	83.51 ± 8.88	45.00 ±16.03	0.210	0.603	
Voxel based features	No	69.46 ± 9.69	66.63 ± 7.79	83.42 ±8.39	39.10 ±15.88			
Voxel based features	Yes	71.34 ± 10.34	66.99 ± 8.04	82.68 ±9.58	41.30 ±15.40	0.370	0.394	
PCA VF 90	No	67.75 ± 10.10	65.19 ± 8.09	84.09 ±11.33	34.20 ±17.33			
PCA VF 90	Yes	68.63 ± 10.37	65.25 ± 7.02	85.72 ±10.96	31.70 ±15.81	0.713	0.116	
Moradi features	No	69.94 ± 10.05	67.77 ± 7.77	83.81 ±10.11	41.50 ±14.79			
Moradi features	Yes	74.04 ± 9.37	70.84 ± 7.28	86.79 ±8.26	44.70 ±14.66	0.068	0.798	
Region features	No	**76.38** ± 8.67	71.27 ± 7.53	85.97 ±9.57	47.10 ±16.33			
Region features	Yes	**79.58** ± 7.71	71.73 ± 7.56	84.07 ±9.26	51.50 ±15.58	0.120	0.060	

Table 7: Cross-validated SVM performance measures with the non-QC dataset using ACC as the model selection criterion.

Feature Set	**Age Removal**	**AUC (%)**	**ACC (%)**	**SEN (%)**	**SPE (%)**	p_{age}	p_{Hippo}	p_{Class}
Hippocampus volumes	No	70.51 ± 9.00	67.35 ± 7.91	85.06 ±8.59	38.30 ±12.09			0.330
Hippocampus volumes	Yes	74.99 ± 8.28	68.86 ± 7.48	81.91 ± 7.41	47.40 ± 11.71	0.026		0.759
Hippo. and entor. vol.	No	72.43 ± 9.33	69.03 ± 8.29	81.36 ± 8.05	48.8 ± 13.44			0.968
Hippo. and entor. vol.	Yes	75.40 ± 8.55	**71.71** ± 8.35	82.20 ± 7.97	53.50 ± 14.79	0.065	0.777	0.880
Voxel based features	No	67.10 ± 11.11	61.71 ± 7.97	79.86 ± 13.67	32.10 ± 21.23			0.939
Voxel based features	Yes	68.35 ± 10.40	62.46 ± 7.03	80.59 ±14.26	32.90 ± 21.60	0.506	0.106	0.548
PCA VF 90	No	66.93 ± 10.40	63.65 ± 7.03	78.65 ±14.26	39.10 ±21.60			0.669
PCA VF 90	Yes	67.58 ± 10.03	63.93 ± 7.51	78.67 ±10.26	39.80 ±15.16	0.806	0.064	0.356
Moradi features	No	72.85 ± 9.12	68.93 ± 7.46	84.49 ±7.36	43.40 ±12.27			0.356
Moradi features	Yes	75.00 ± 8.63	70.09 ± 7.04	83.99 ±7.47	47.30 ±12.87	0.292	0.997	0.650
Region features	No	72.55 ± 9.76	**69.16** ± 7.87	82.73 ± 9.79	46.90 ± 16.83			0.122
Region features	Yes	75.98 ± 9.35	71.01 ± 7.29	86.94 ± 10.16	44.90 ± 13.23	0.105	0.763	0.076

Table 8: Cross-validated RLR performance measures with the non-QC dataset using ACC as the model selection criterion.

Feature Set	**Age Removal**	**AUC (%)**	**ACC (%)**	**SEN (%)**	**SPE (%)**	p_{age}	p_{Hippo}	p_{Class}
Hippocampus volumes	No	70.96 ± 9.07	66.28 ± 7.21	88.92 ± 10.73	29.10 ± 13.94			
Hippocampus volumes	Yes	74.85 ± 8.39	69.12 ± 7.31	84.10 ±9.25	44.50 ±14.24	0.050		
Hippo. and entor. vol.	No	72.40 ± 9.15	69.55 ± 7.90	85.50 ± 9.18	43.30 ±15.81			
Hippo. and entor. vol.	Yes	75.50 ± 9.11	70.50 ± 7.54	84.68 ± 9.50	47.20 ± 15.69	0.056	0.650	
Voxel based features	No	67.33 ± 10.03	64.81 ± 8.03	80.76 ± 9.45	38.60 ± 16.43			
Voxel based features	Yes	69.95 ± 10.43	66.15 ± 8.26	80.21 ±10.41	43.10 ±16.35	0.235	0.244	
PCA VF 90	No	68.01 ± 10.43	64.95 ± 8.26	86.20 ±10.41	30.10 ±16.35			
PCA VF 90	Yes	69.63 ± 9.55	65.08 ± 7.75	84.72 ±13.08	32.90 ±17.68	0.484	0.180	
Moradi features	No	71.08 ± 9.55	68.60 ± 6.59	85.35 ±8.80	41.10 ±14.83			
Moradi features	Yes	74.10 ± 9.33	70.42 ± 7.38	85.94 ± 9.51	45.00 ± 14.39	0.125	0.836	
Region features	No	75.91 ± 9.12	**71.01** ± 7.89	84.52 ± 9.27	48.80 ± 16.02			
Region features	Yes	79.41 ± 7.98	**72.07** ± 7.81	84.24 ± 9.40	52.10 ± 15.32	0.123	0.717	

p-values from hypothesis tests comparing the AUCs: p_{Age} (comparing age removed features vs. non age removed features), p_{Hippo} (comparing hippocampus features with the remaining features for the age removed case) and p_{Class} (comparing SVM vs. RLR results over the same set of features). The best results of each experimental setup, for each classifier and with/without the age correction process, have been marked in bold face. In addition, the standard deviation of each measure, computed as $\hat{\sigma}$ in Eq. (2), is included below its average value after $\pm$ symbol.

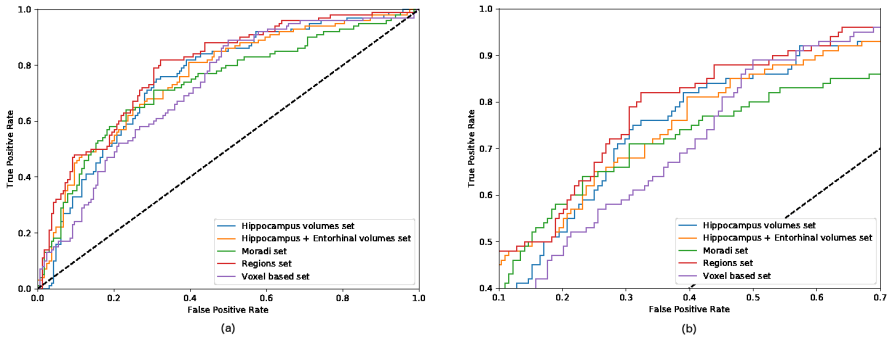

Figure 1: ROC curves corresponding to distinct the features sets used in RLR classification with the non-QC dataset. The age effect was removed.

The AUC values of region features were the highest in all the experiments. However, the performance improvement over the hippocampus feature set, which was our baseline, did not reach the statistical significance and these improved AUCs need to be interpreted with care. In the particular case of the non-QC dataset and the RLR classifier, the regions feature set produced significantly higher AUC than hippocampus volumes.

Figure 1 depicts the ROC curves for the different feature sets under study for the RLR classifier in the non-QC dataset. Focusing on the center of these curves (see the panel 1 b), we can corroborate that the region feature set appeared superior, but the performance differences were small. To avoid crowding, the ROCs of the PCA voxel feature set were not visualized as PCA voxel features always performed worse than the voxel features without PCA. For similar

reason, the figure only displays the ROC curves of raw Hippocampus and Hippocampus + Entorhinal cortex volumes and not the ICV-normalized ones. The same principle will be followed in later figures.

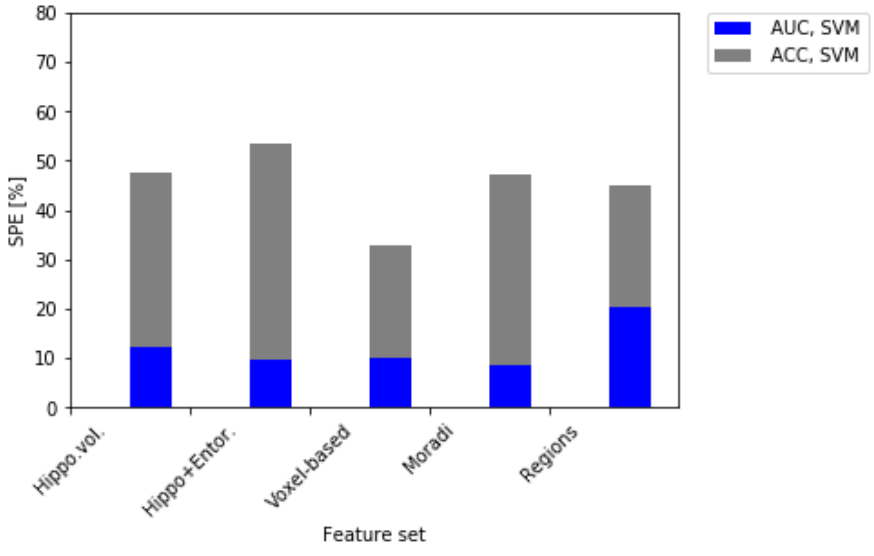

Figure 2: Specificity values of SVM classifiers when AUC and ACC were used for model selection. The models selected with ACC resulted in specificity values close 50 % whereas the models selected with AUC resulted in very low specificity values.

Regarding the use of two different classifiers, differences between AUCs of SVM and RLR were not significant. However, SVM yielded low specificity values and the relation between SPE and SEN was more balanced with the RLR classifier. Because of this we studied whether the use of AUC as the model selection criteria contributed to this imbalance with the SVM classifier. Using ACC as a model selection criterion notably reduced this SPE/SEN imbalance as can be seen in Figure 2 where the specificity values are compared between ACC and AUC based model selection. As the comparison of Tables 5 - 8 reveals, the final AUC values did not markedly differ between the two model selectors. With ACC as the model selection criterion, the sensitivity values were still markedly higher than the specificity values. Some insight to the phenomenon

can be obtained by visual analysis of the Hippocampus feature set, with just 2 features, thereby permitting visual analysis. Figure 3 shows the datapoints along with the decision regions for two classes over 10 CV folds. It can be observed that the data from two classes were highly overlapping, and in these cases, the classification boundary has a tendency to shift more towards the majority class (pMCI in this case) than what might be expected based on a modest class-imbalance.

We evaluated the effects of age removal on the feature sets. For this purpose, Figure 4 shows a detailed analysis of the advantages of removing the age effects. As a result, classification scores improved for every age removed effects feature set (see the panel 4 c). However, as visible in Tables 3-6, significant improvement (p-value < 0.1) was observed only for hippocampus and hippocampus + entorhinal volume feature sets.

The differences between the AUCs of raw and ICV-normalized hippocampus and hippocampus + entorhinal volumes were not significant. Surprisingly, the raw volumes performed slightly better in terms of AUC within each dataset. However, this result agrees with findings in [35, 27] and it is not central for the purposes of this work to analyze the potential reasons for this result.

Finally, Figure 5 shows the differences between QC and non QC datasets when age effects were removed. As expected Hippocampus and Hippocampus plus Entorhinal volumes were benefited from the Quality Control process, whereas remaining features sets resulted in better performances when all the available data were used.

4. Discussion

In this work, we compared six different feature representations of MRI for predicting the AD conversion in MCI subjects. The feature sets we studied varied from high dimensional feature sets produced by VBM via regional cortical thickness, surface area, and volumetry to simple and easily interpretable features such as hippocampus and entorhinal cortex volumes (see Table 2). We addressed

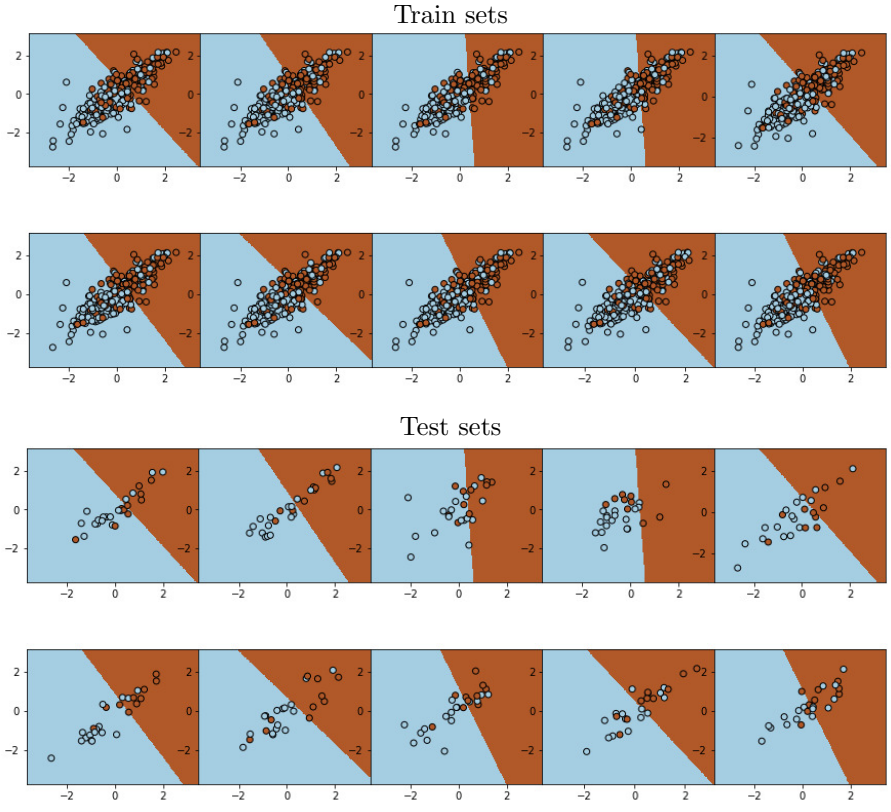

Figure 3: SVM classification boundaries with the age corrected hippocampus non-QC feature set overlaid to the train and test data. ACC was used as the model selection criteria. Each panel depicts the classifier training in a single CV fold (folds from the first CV run are shown). On top of the decision regions, train or test sets of that particular fold are plotted. Red color corresponds to sMCI class and blue corresponds to pMCI class. Note that the decision regions are always based on the training set, and therefore they are the same whether overlaying test or train data. x-axis (y-axis) of the feature space corresponds to right (left) Hippocampus volume. The feature values are normalized as explained in Section 2.3.

the feature representations using two learning algorithms, SVM and RLR, and with several metrics, AUC, ACC, SEN and SPE, that gave a reliable insight into the relative performance of different feature sets. AUC was selected as the principal figure of merit, due to its insensitivity to the class imbalance (note that the datasets contained twice the number of pMCIs (subjects who converted to AD) compared to sMCIs (subjects who remained as MCIs)). The

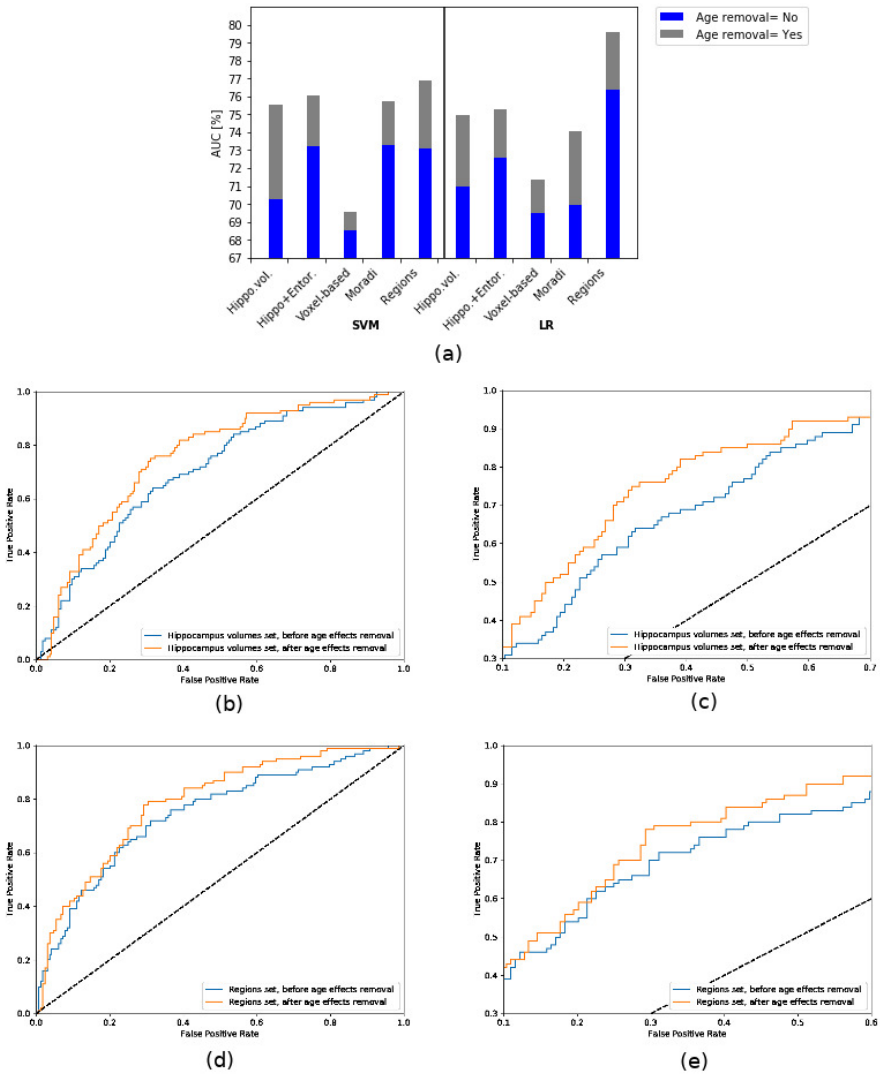

Figure 4: Analysis of age removal effects: (a) AUC comparison for different feature sets and both classifiers; (b) and (c) ROC curves for RLR classifier using hippocampus volumes; (d) and (e) ROC curves for RLR classifier using region features. Age removal improved predictions in all cases.

evaluation process was carried out with a nested 10-fold CV repeated 10 times ensuring the insensitivity of the conclusions to random train/test division of the holdout method used previously [28]. Selecting the parameters of the classifiers

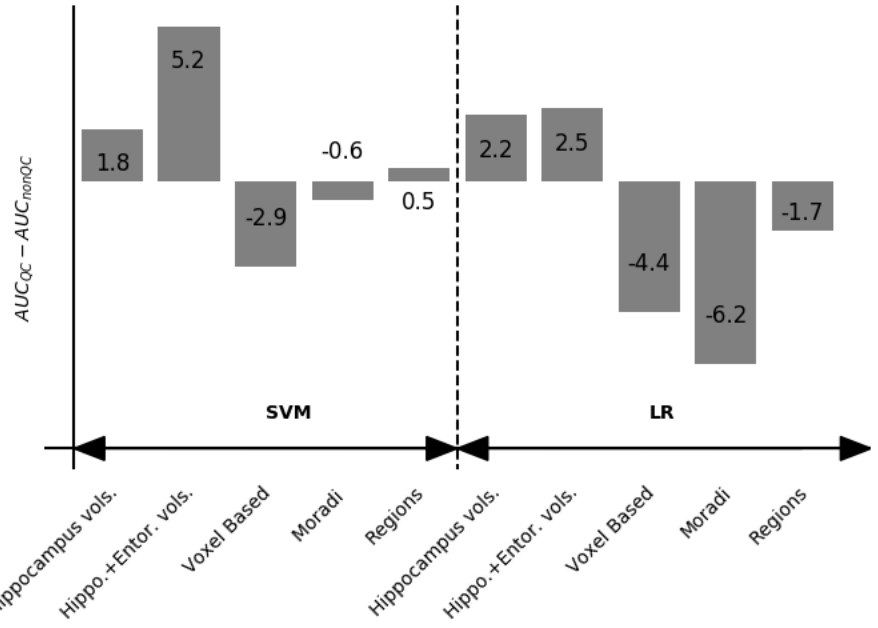

Figure 5: Differences between the AUC values with the QC dataset and the non-QC dataset for SVM (left) and RLR (right).

inside nested CV ensures that there are no biases towards particular feature representations due to arbitrarily selected classifier parameters.

We found that age-corrected *regions feature set* [10] outperformed the remaining feature sets, specifically in AUC, even though the improvement did not reach statistical significance. This result suggests that regions based features were equal or better predictors than the left and right hippocampal volumes (HV) alone (which were included in the region feature set). This is interesting as a recent study [21] concluded that HV had the highest AUC among a set of individual regional volume features and was better in terms of the prognostic efficacy of combining various volumetrics. Their experimental setting was similar to the one analyzed here, however, with three main differences. First, removing

[10] See `https://github.com/MartaGomez/Regions-list-/wiki/Regions-list` for a detailed description

age related effects from MRI data was not considered; second, the set of pMCI patients was about half of ours; and, third, the combined volumetric analysis did not consider measures such as surface area or cortical thickness. This can explain the improvement in the best classification accuracy from 69 % of [21] to 80 % in the present study.

Voxel-based representations did not perform well in this study when coupled with standard feature reduction techniques (elastic-net or PCA). This was in contrast to a recent data-analysis competition, where the goal was to classify subjects into NC, MCI, and AD categories based on MRI [26]. However, as multiple factors have effect to a performance of an approach in a data analysis competition, definite conclusions on feature representations cannot be made based on such competitions. However, also in our own experience, voxel-based methods, coupled with elastic-net feature selection, perform well in classifying between NC and AD or NC and MCI [31]. These discrepancies may suggest that NC vs. MCI (or AD) classification and AD-conversion prediction have different characteristics. Further, we note that feature pre-selection based on AD and NC data suggested by Moradi et al. [30] improved the conversion prediction accuracy markedly.

Retico et al. found that the voxel based VBM features best discriminate between sMCI and pMCI after applying Recursive Feature Elimination (RFE) [20]. However, again, the maximum accuracy in [20] was much lower than the accuracies in the present study and pMCI vs. sMCI classifiers were trained only using AD and NC subjects that may explain this. Additionally, the statistical framework was incomplete as no hypothesis testing was done and the exact definition of stable MCI class remained unclear. Other works, such as [18], concluded that the combination of different feature representations resulted into a better classification accuracy than one representation alone. Again, the classification accuracies were lower than in the present work. Moreover, [18] selected classifier hyperparameters based on test data that may cause upward bias in the reported accuracies [15].

It is important to point out that while our classification accuracies were

better than those in the studies reviewed above, the performance measures are not directly comparable because different definitions of pMCI and sMCI. In fact, this is a problem that complicates the comparison of ML methods for this particular application and it is reviewed in further length in [22]. Namely, the definition of sMCI subject based on a certain cutoff (say 3 years) is problematic as this simple criterion would place a subject who received an AD diagnosis 4 years after the baseline visit into the sMCI category. Our view is that this would create unrealistic heterogeneity into the sMCI class and therefore tracking subjects' status after the cutoff is necessary (if possible). We have populated our sMCI category based on all the information available by ADNI.

Regarding the used ML methods, RLR provided, in general, similar AUC values than SVM, but had an advantage of higher specificity (it classified sMCI cases much better than the SVM did). SVM had a tendency of overpopulating the pMCI class. However, in the case of SVM, low specificity seemed to depend on the using AUC as the criterion for the hyperparameter selection. The values in Tables 7 and 8 reveal how selecting the hyperparameters instead through ACC resulted in an overall improvement of specificity with a small loss of sensitivity. This is an interesting phenomenon, as it seems to be a problem of a specific class of learning algorithms, which invites further research. However, as this issue is not central to the goals of this work, we do not analyze it further. Also with the ACC model selection and with RLR, the specificity values were lower than the sensitivity values. However, as already mentioned (see Fig. 3), this level of SEN/SPE imbalance can be explained by the slight class imbalance (approximately 60 % pMCI and 40 % sMCI) and overlapping feature densities.

There were no significant differences between the classification accuracies or AUCs obtained with non-QC and QC datasets. However, the small differences between the two datasets were as expected as shown in Figure 5. For Hippocampus and Hippocampus and Entorhinal volumes, the QC was moderately useful whereas for the Moradi and Voxel based features it was moderately detrimental. This is as expected since the QC was based on Freesurfer segmentations

(as Hippocampus and Entorhinal volumes) but the voxel-based and Moradi features were not. Interestingly, for region based features (also based on Freesurfer segmentation), the QC seemed not to influence the performance of the classifier.

It is remarkable that the age removal seem to be a key for better performances. As Figure 4 illustrates, age removal always led to better classification performances, although the improvements were not always statistically significant. This agrees with a recent work of [31] which demonstrated the same for NC vs. MCI classification.

5. Conclusion

This paper evaluated the performance of various types of MRI features for the future AD conversion prediction and it also analyzed the performance of each feature set over two classifiers (Support Vector Machines and Regularized Logistic Regression) and with and without applying an age correction process.

Experimental results showed that regional features consistently yielded the best performance, although the performance difference to other features was not statistically significant. Besides, the age removal seemed to be a key for better performances, but the improvement reached statistical significance only rarely.

6. Supplementary Material

Table 9: Cross-validated performance measures of the PCA voxel feature set with the QC dataset using AUC as the model selection criterion.

Classifier	**Feature Set**	**Age Removal**	**AUC**	**ACC**	**SEN**	**SPE**	p_{Age}	p_{Hippo}	p_{Class}
SVM	PCA VF 99	No	63.50 %	59.88 %	92.73 %	10.23 %			0.604
SVM	PCA VF 99	Yes	65.53 %	60.75 %	92.64 %	12.54 %	0.535	0.035	0.509
SVM	PCA VF 95	No	64.22 %	60.75 %	94.27 %	10.11 %			0.826
SVM	PCA VF 95	Yes	66.81 %	60.74 %	93.27 %	11.59 %	0.332	0.045	0.556
SVM	PCA VF 90	No	64.33 %	60.11 %	90.36 %	14.38 %			0.791
SVM	PCA VF 90	Yes	66.78%	61.09 %	91.82 %	14.64 %	0.351	0.042	0.569
SVM	PCA VF 75	No	63.08 %	59.81 %	84.45 %	22.57 %			0.773
SVM	PCA VF 75	Yes	66.89 %	62.25 %	87.45 %	24.45 %	0.229	0.062	0.395
LR	PCA VF 99	No	61.95 %	59.12 %	78.73 %	26.68 %			
LR	PCA VF 99	Yes	63.40 %	61.33 %	80.82 %	31.91 %	0.600	0.006	
LR	PCA VF 95	No	63.59 %	61.19%	80.00 %	32.93 %			
LR	PCA VF 95	Yes	65.20 %	61.22 %	80.64 %	31.91 %	0.549	0.011	
LR	PCA VF 90	No	63.63 %	60.88 %	79.45 %	33.00 %			
LR	PCA VF 90	Yes	65.35 %	61.28 %	81.73 %	30.43 %	0.573	0.018	
LR	PCA VF 75	No	63.83 %	61.08 %	74.91 %	40.30 %			
LR	PCA VF 75	Yes	64.90 %	61.69 %	76.27 %	39.68 %	0.722	0.013	

Table 10: Cross-validated performance measures of the PCA voxel feature set with the non QC dataset using AUC as the model selection criterion

Classifier	**Feature Set**	**Age Removal**	**AUC**	**ACC**	**SEN**	**SPE**	p_{Age}	p_{Hippo}	p_{Class}
SVM	PCA VF 99	No	68.41 %	62.30 %	96.19 %	6.80 %			0.328
SVM	PCA VF 99	Yes	69.13 %	63.90 %	95.12 %	12.70 %	0.754	0.149	0.549
SVM	PCA VF 95	No	67.73 %	62.26%	94.83 %	8.90 %			0.782
SVM	PCA VF 95	Yes	68.42 %	63.22 %	93.97 %	12.80 %	0.172	0.105	0.789
SVM	PCA VF 90	No	67.94 %	62.11 %	93.26 %	11.00 %			0.923
SVM	PCA VF 90	Yes	68.38 %	62.50 %	92.39 %	13.50 %	0.194	0.117	0.915
SVM	PCA VF 75	No	67.48 %	64.09 %	85.82 %	28.50 %			0.808
SVM	PCA VF 75	Yes	68.95 %	64.35 %	89.85 %	22.60 %	0.785	0.145	0.717
LR	PCA VF 99	No	66.02 %	64.50 %	86.28 %	28.80 %			
LR	PCA VF 99	Yes	67.58 %	64.68 %	87.02 %	28.10 %	0.536	0.058	
LR	PCA VF 95	No	67.11 %	65.21%	85.77 %	31.40 %			
LR	PCA VF 95	Yes	67.78 %	65.64 %	85.66%	32.80 %	0.803	0.074	
LR	PCA VF 90	No	67.75 %	65.19 %	84.09 %	34.20 %			
LR	PCA VF 90	Yes	68.63 %	65.25 %	85.72 %	31.70 %	0.713	0.116	
LR	PCA VF 75	No	67.91 %	65.65 %	79.64 %	42.70 %			
LR	PCA VF 75	Yes	69.74 %	66.69 %	83.49 %	39.10 %	0.410	0.186	

Table 11: Cross-validated performance measures of the PCA voxel feature set with the non-QC dataset using ACC as the model selection criterion

Classifier	**Feature Set**	**Age Removal**	**AUC**	**ACC**	**SEN**	**SPE**	p_{age}	p_{Hippo}	p_{Class}
SVM	PCA VF 99	No	66.78 %	62.15 %	79.56 %	33.70 %			0.621
SVM	PCA VF 99	Yes	67.98 %	63.14 %	80.18 %	35.30 %	0.613	0.085	0.784
SVM	PCA VF 95	No	66.72 %	62.66%	80.00 %	34.30 %			0.981
SVM	PCA VF 95	Yes	67.82 %	64.09 %	79.71 %	38.50 %	0.660	0.086	0.628
SVM	PCA VF 90	No	66.93 %	63.65 %	78.65 %	39.10 %			0.669
SVM	PCA VF 90	Yes	67.58 %	63.93 %	78.67 %	39.80 %	0.806	0.064	0.356
SVM	PCA VF 75	No	67.33 %	64.99 %	80.07 %	40.30 %			0.560
SVM	PCA VF 75	Yes	69.29 %	66.39 %	80.82 %	42.80 %	0.371	0.169	0.879
LR	PCA VF 99	No	65.42 %	63.52 %	87.46 %	24.30 %			
LR	PCA VF 99	Yes	67.35 %	64.96 %	87.76 %	27.60 %	0.513	0.059	
LR	PCA VF 95	No	66.79 %	64.81 %	88.34 %	26.20 %			
LR	PCA VF 95	Yes	68.77 %	66.10 %	86.07%	33.30 %	0.492	0.126	
LR	PCA VF 90	No	68.01 %	64.95 %	86.20 %	30.10 %			
LR	PCA VF 90	Yes	69.63 %	65.08 %	84.72 %	32.90 %	0.484	0.180	
LR	PCA VF 75	No	67.33 %	68.22 %	85.26 %	36.50 %			
LR	PCA VF 75	Yes	69.53 %	66.20 %	83.90 %	37.20 %	0.574	0.211	

Table 12: Cross-validated performance measures with the non-QC dataset using AUC as the model selection criterion for different values of the RLR hyperparameter alpha testing high dimensional age removed features.

Classifier	**Feature Set**	**Alpha Value**	**AUC**	**ACC**	**SEN**	**SPE**
LR	Voxel based features	0.25	71.06 ± 10.09	62.88 ± 3.46	95.28 ±9.28	9.80 ± 13.99
LR	Voxel based features	0.5	71.34 ± 10.35	66.99 ± 8.04	82.68 ±9.58	41.30 ±15.40
LR	Voxel based features	0.75	72.37 ± 9.67	66.65 ± 7.36	86.70 ±8.61	33.80 ± 14.27
LR	Moradi features	0.25	72.28 ± 9.28	66.32 ± 6.35	89.14 ±11.18	29.00 ±27.51
LR	Moradi features	0.5	74.04 ± 9.37	70.84 ± 7.28	86.79 ±8.26	44.70 ±14.66
LR	Moradi features	0.75	73.28 ± 9.11	70.51 ± 6.65	87.91 ±9.75	42.00 ±16.25
LR	Region features	0.25	80.18 ± 7.78	71.72 ± 8.07	83.45 ±9.66	52.50 ±16.15
LR	Region features	0.5	79.58 ± 7.71	71.73 ± 7.56 %	84.07 ±9.26	51.50 ±15.58
LR	Region features	0.75	79.59 ± 7.54	71.96 ± 7.97	84.31 ±9.63	51.70 ±15.24

Acknowledgments

Data collection and sharing for this project was 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, Alzheimers Association; Alzheimers Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; 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.

J. Tohka's work was supported by the Academy of Finland and V. Gómez-Verdejo's work has been partly funded by the Spanish MINECO grant TEC2014-52289R, TEC2016-81900-REDT/AEI and TEC2017-83838-R.

References

[1] H. Braak, E. Braak, Development of Alzheimer-related neurofibrillary changes in the neocortex inversely recapitulates cortical myelogenesis, Acta neuropathologica 92 (2) (1996) 197–201.

[2] A. Delacourte, J. David, N. Sergeant, L. Buee, A. Wattez, P. Vermersch, F. Ghozali, C. Fallet-Bianco, F. Pasquier, F. Lebert, et al., The biochemical pathway of neurofibrillary degeneration in aging and Alzheimers Disease, Neurology 52 (6) (1999) 1158–1158.

[3] J. Morris, M. Storandt, D. McKeel, E. Rubin, J. Price, E. Grant, L. Berg, Cerebral amyloid deposition and diffuse plaques in "normal" aging evidence for presymptomatic and very mild Alzheimer's Disease, Neurology 46 (3) (1996) 707–719.

[4] A. Serrano-Pozo, M. P. Frosch, E. Masliah, B. T. Hyman, Neuropathological alterations in Alzheimer Disease, Cold Spring Harbor perspectives in medicine 1 (1) (2011) a006189.

[5] L. Mosconi, M. Brys, L. Glodzik-Sobanska, S. De Santi, H. Rusinek, M. J. De Leon, Early detection of Alzheimers Disease using neuroimaging, Experimental gerontology 42 (1) (2007) 129–138.

[6] R. C. Petersen, B. Caracciolo, C. Brayne, S. Gauthier, V. Jelic, L. Fratiglioni, Mild cognitive impairment: A concept in evolution, Journal of internal medicine 275 (3) (2014) 214–228.

[7] S. J. Vos, F. Verhey, L. Frölich, J. Kornhuber, J. Wiltfang, W. Maier, O. Peters, E. Rüther, F. Nobili, S. Morbelli, et al., Prevalence and prognosis of Alzheimers Disease at the mild cognitive impairment stage, Brain 138 (5) (2015) 1327–1338.

[8] A. J. Mitchell, M. Shiri-Feshki, Rate of progression of mild cognitive impairment to dementia–meta-analysis of 41 robust inception cohort studies, Acta Psychiatrica Scandinavica 119 (4) (2009) 252–265.

[9] S. Teipel, A. Drzezga, M. J. Grothe, H. Barthel, G. Chételat, N. Schuff, P. Skudlarski, E. Cavedo, G. B. Frisoni, W. Hoffmann, et al., Multimodal imaging in alzheimer's disease: validity and usefulness for early detection, The Lancet Neurology 14 (10) (2015) 1037–1053.

[10] M. W. Weiner, D. P. Veitch, P. S. Aisen, L. A. Beckett, N. J. Cairns, R. C. Green, D. Harvey, C. R. Jack, W. Jagust, J. C. Morris, et al., Recent publications from the alzheimer's disease neuroimaging initiative: Reviewing progress toward improved ad clinical trials, Alzheimer's & Dementia.

[11] K. A. Johnson, N. C. Fox, R. A. Sperling, W. E. Klunk, Brain imaging in Alzheimer Disease, Cold Spring Harbor perspectives in medicine 2 (4) (2012) a006213.

[12] G. B. Frisoni, N. C. Fox, C. R. Jack, P. Scheltens, P. M. Thompson, The clinical use of structural MRI in Alzheimer Disease, Nature Reviews Neurology 6 (2) (2010) 67–77.

[13] S. Klöppel, C. M. Stonnington, C. Chu, B. Draganski, R. I. Scahill, J. D. Rohrer, N. C. Fox, C. R. Jack Jr, J. Ashburner, R. S. Frackowiak, Automatic classification of MR scans in Alzheimer's Disease, Brain 131 (3) (2008) 681–689.

[14] S. Adaszewski, J. Dukart, F. Kherif, R. Frackowiak, B. Draganski, A. D. N. Initiative, et al., How early can we predict Alzheimer's Disease using computational anatomy?, Neurobiology of aging 34 (12) (2013) 2815–2826.

[15] S. F. Eskildsen, P. Coupé, D. García-Lorenzo, V. Fonov, J. C. Pruessner, D. L. Collins, Prediction of Alzheimer's disease in subjects with mild cognitive impairment from the adni cohort using patterns of cortical thinning, NeuroImage 65 (2013) 511–521.

[16] R. Casanova, C. T. Whitlow, B. Wagner, J. Williamson, S. A. Shumaker, J. A. Maldjian, M. A. Espeland, High dimensional classification of structural MRI Alzheimers Disease data based on large scale regularization, Frontiers in neuroinformatics 5.

[17] C. Gaser, K. Franke, S. Klöppel, N. Koutsouleris, H. Sauer, A. D. N. Initiative, et al., BrainAGE in mild cognitive impaired patients: predicting the conversion to Alzheimers Disease, PloS ONE 8 (6) (2013) e67346.

[18] R. Wolz, V. Julkunen, J. Koikkalainen, E. Niskanen, D. P. Zhang, D. Rueckert, H. Soininen, J. Lötjönen, A. D. N. Initiative, et al., Multi-method analysis of MRI images in early diagnostics of Alzheimer's Disease, PloS one 6 (10) (2011) e25446.

[19] E. Westman, J.-S. Muehlboeck, A. Simmons, Combining MRI and CSF measures for classification of Alzheimer's Disease and prediction of Mild Cognitive Impairment conversion, Neuroimage 62 (1) (2012) 229–238.

[20] A. Retico, P. Bosco, P. Cerello, E. Fiorina, A. Chincarini, M. E. Fantacci, Predictive models based on Support Vector Machines: Whole-brain versus regional analysis of structural MRI in the Alzheimer's Disease, Journal of Neuroimaging 25 (4) (2015) 552–563.

[21] T. Tanpitukpongse, M. Mazurowski, J. Ikhena, J. Petrella, Predictive utility of marketed volumetric software tools in subjects at risk for Alzheimer Disease: Do regions outside the hippocampus matter?, American Journal of Neuroradiology 38 (3) (2017) 546–552.

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

[23] B. Mwangi, T. S. Tian, J. C. Soares, A review of feature reduction techniques in neuroimaging, Neuroinformatics 12 (2) (2014) 229–244.

[24] E. Westman, A. Simmons, Y. Zhang, J.-S. Muehlboeck, C. Tunnard, Y. Liu, L. Collins, A. Evans, P. Mecocci, B. Vellas, et al., Multivariate analysis of mri data for alzheimer's disease, mild cognitive impairment and healthy controls, Neuroimage 54 (2) (2011) 1178–1187.

[25] C. G. Schwarz, J. L. Gunter, H. J. Wiste, S. A. Przybelski, S. D. Weigand, C. P. Ward, M. L. Senjem, P. Vemuri, M. E. Murray, D. W. Dickson, et al., A large-scale comparison of cortical thickness and volume methods for measuring alzheimer's disease severity, NeuroImage: Clinical 11 (2016) 802–812.

[26] E. E. Bron, M. Smits, W. M. Van Der Flier, H. Vrenken, F. Barkhof, P. Scheltens, J. M. Papma, R. M. Steketee, C. M. Orellana, R. Meijboom,

et al., Standardized evaluation of algorithms for computer-aided diagnosis of dementia based on structural mri: the caddementia challenge, NeuroImage 111 (2015) 562–579.

[27] E. Westman, C. Aguilar, J.-S. Muehlboeck, A. Simmons, Regional magnetic resonance imaging measures for multivariate analysis in alzheimers disease and mild cognitive impairment, Brain topography 26 (1) (2013) 9–23.

[28] R. Cuingnet, E. Gerardin, J. Tessieras, G. Auzias, S. Lehricy, M. O. Habert, M. Chupin, H. Benali, O. Colliot, Automatic classification of patients with Alzheimer's Disease from structural MRI: a comparison of ten methods using the ADNI database, Neuroimage 56 (2) (2011) 766–781.

[29] A. Lebedev, E. Westman, G. Van Westen, M. Kramberger, A. Lundervold, D. Aarsland, H. Soininen, I. Kłoszewska, P. Mecocci, M. Tsolaki, et al., Random forest ensembles for detection and prediction of alzheimer's disease with a good between-cohort robustness, NeuroImage: Clinical 6 (2014) 115–125.

[30] E. Moradi, A. Pepe, C. Gaser, H. Huttunen, J. Tohka, A. D. N. Initiative, et al., Machine learning framework for early MRI-based Alzheimer's conversion prediction in MCI subjects, Neuroimage 104 (2015) 398–412.

[31] J. Tohka, E. Moradi, H. Huttunen, Comparison of feature selection techniques in machine learning for anatomical brain MRI in dementia, Neuroinformatics 14 (2016) 279 – 296.

[32] C. Chu, A.-L. Hsu, K.-H. Chou, P. Bandettini, C. . Lin, A. D. N. Initiative, et al., Does feature selection improve classification accuracy? Impact of sample size and feature selection on classification using anatomical magnetic resonance images, Neuroimage 60 (1) (2012) 59–70.

[33] A. Zandifar, V. Fonov, P. Coupé, J. Pruessner, D. L. Collins, A. D. N. Initiative, et al., A comparison of accurate automatic hippocampal segmentation methods, NeuroImage 155 (2017) 383 – 393.

[34] J.-L. Chepkoech, K. B. Walhovd, H. Grydeland, A. M. Fjell, Effects of change in freesurfer version on classification accuracy of patients with alzheimer's disease and mild cognitive impairment, Human brain mapping 37 (5) (2016) 1831–1841.

[35] O. Voevodskaya, A. Simmons, R. Nordenskjöld, J. Kullberg, H. Ahlström, L. Lind, L.-O. Wahlund, E.-M. Larsson, E. Westman, A. D. N. Initiative, et al., The effects of intracranial volume adjustment approaches on multiple regional mri volumes in healthy aging and alzheimer's disease, Frontiers in aging neuroscience 6.

[36] Y. Cui, B. Liu, S. Luo, X. Zhen, M. Fan, T. Liu, W. Zhu, M. Park, T. Jiang, J. S. Jin, et al., Identification of conversion from mild cognitive impairment to alzheimer's disease using multivariate predictors, PloS one 6 (7) (2011) e21896.

[37] J. Dukart, M. L. Schroeter, K. Mueller, A. D. N. Initiative, et al., Age correction in dementia–matching to a healthy brain, PloS one 6 (7) (2011) e22193.

[38] K. Franke, G. Ziegler, S. Klöppel, C. Gaser, A. D. N. Initiative, et al., Estimating the age of healthy subjects from t1-weighted MRI scans using kernel methods: Exploring the influence of various parameters, Neuroimage 50 (3) (2010) 883–892.

[39] J. A. Hanley, B. J. McNeil, The meaning and use of the area under a receiver operating characteristic (roc) curve., Radiology 143 (1) (1982) 29–36.

[40] S. Rosset, Model selection via the auc, in: Proceedings of the twenty-first international conference on Machine learning, ACM, 2004, p. 89.

[41] C.-C. Chang, C.-J. Lin, LIBSVM: A library for Support Vector Machines, ACM Trans. Intell. Systems Tech. 2 (2011) 27.

[42] J. H. Friedman, T. Hastie, R. Tibshirani, Regularization paths for generalized linear models via coordinate descent, J. Stat. Software 33 (1) (2010) 1–22.

[43] T. Hastie, R. Tibshirani, J. Friedman, The elements of statistical learning - Data Mining, Inference, and Prediction, Second Edition, Springer series in statistics New York, 2009.

[44] H. Zou, T. Hastie, Regularization and variable selection via the elastic net, J. R. Stat. Soc.: Series B 67 (2) (2005) 301–320.

[45] C. Aguilar, E. Westman, J.-S. Muehlboeck, P. Mecocci, B. Vellas, M. Tsolaki, I. Kloszewska, H. Soininen, S. Lovestone, C. Spenger, et al., Different multivariate techniques for automated classification of mri data in alzheimers disease and mild cognitive impairment, Psychiatry Research: Neuroimaging 212 (2) (2013) 89–98.

[46] L. Breiman, Random Forests, Machine Learning 45 (1) (2001) 5–32.

[47] M. N. Wright, A. Ziegler, ranger: A fast implementation of random forests for high dimensional data in c++ and r, Journal of Statistical Software 77.

[48] C. Nadeau, Y. Bengio, Inference for the generalization error, in: Advances in neural information processing systems, 2000, pp. 307–313.

[49] T. G. Dietterich, Approximate statistical tests for comparing supervised classification learning algorithms, Neural computation 10 (7) (1998) 1895–1923.