ORIGINAL ARTICLE

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

Marta Gómez-Sancho[1] | Jussi Tohka[2,*] | Vanessa Gómez-Verdejo[1,*] | for the Alzheimer's Disease Neuroimaging Initiative[3]

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

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

[3]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`

[*]These two authors share the senior authorship.

Correspondence
Vanessa Gómez-Verdejo, Jussi Tohka
Email: vanessa@tsc.uc3m.es, jussi.tohka@uef.fi

Funding information
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 and TEC2016-81900-REDT/AEI.

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 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 [2, 7, 20, 26, 21]. 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 [24]. Subjects with MCI present a high risk of developing AD, but still, only a fraction of them convert to AD [29]. 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. Much research has been focused on T1-weighted Magnetic Resonance Imaging (MRI). It is one of the most widely studied imaging techniques [17] 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 [13]. 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 detecting the conversion from MCI to AD based on MRI (e.g.,[18, 1, 10, 3]). 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., [14, 31, 30, 25, 27]). 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 [22] for a review).

An early and important study [6] 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 subject 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. [19] evaluated their method over the same dataset as [6] managing to obtain significantly better performance than the chance level, pointing to the reason 1) as the most significant cause of the improvement.

Since [6], we can find few studies of different feature representations presenting partially conflicting results. As an example, [27] found that the prognostic efficacy of hippocampus volumetry was better than combined regional

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

volumetrics in 2 commercially available brain volumetric software packages for MCI conversion prediction. On the other hand, [14] have demonstrated the superiority of their voxel-based brainAGE approach over the hippocampus volume biomarker. Some studies have been directed to feature selection, either supporting [25, 28] or opposing [5] data-driven feature selection.

To close this information gap, we asked what type of feature representations are the best for the MRI-based AD-conversion prediction. For this purpose, we evaluated the performance of various MRI features, including VBM-style, voxel-based features [14], coupled with feature preselection [19] 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. We aimed to classify the subjects into two groups: (1) progressive MCI (pMCI) subjects, who receive the AD diagnosis within three years from imaging and (2) stable MCI (sMCI) subjects, who do not receive the AD diagnosis. We placed a specific care on the definition of stable MCI group by using all the available conversion information for excluding those subjects who receive the AD diagnosis at a later stage. We additionally evaluated age removal [19, 9], 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 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 [6].

2 | MATERIAL

2.1 | ADNI data

Data is collected from the the Alzheimer's Disease Neuroimaging Initiative (ADNI) public database, available at `adni.loni.usc.edu`. 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 Alzheimer's disease (AD). For up-to-date information, see `www.adni-info.org`.

ADNI material considered in this work include 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. Roster ID's of the subjects are available in the supplement.

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

The 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 or less than 3-years of follow-up information was available. Table 1 lists the main characteristics of the subjects of each dataset.

TABLE 1 Demographics of the two datasets (QC and Non-QC) used in this work.

	QC dataset		Non QC dataset	
	sMCI	**pMCI**	**sMCI**	**pMCI**
No. subjects	73	110	100	164
Males / Females	46/27	58/52	66/34	97/67
Age range	59-88	55-89	57-88	55-89

2.2 | 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. Region based features included a complete set of 257 regional cortical thickness, surface area and volume measures provided by FreeSurfer[2]. Freesurfer 4.3 software was employed for the extraction of Hippocampus and entorhinal cortex volumes as well as region features. Particularly, FreeSurfer 4.3 processing results available at the ADNI website were used. 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 [14, 19], 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 [19]. 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 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 99 % of the variance.

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 [11, 9]. We applied the age correction procedure of [19]. 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] 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`.

TABLE 2 Summary of the sets of features considered in this study. Note that the number of Moradi and PCA voxel features is 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)

3 | METHODS

3.1 | 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) [15]. 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.

3.2 | Classifiers

To evaluate each feature set, we considered two types of supervised learning classifiers: Support Vector Machine (SVM) [4] and elastic-net Regularized Logistic Regression (RLR) [12]. Accessible description of these learning methods can be found in [16]. For the SVM implementation, we used the Python open source library Scikit-learn (`http://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html`), which is based on the LIBSVM implementation (`https://www.csie.ntu.edu.tw/~cjlin/libsvm/`). For the RLR classifiers, we applied the GLMNET Python library

(`https://web.stanford.edu/~hastie/glmnet_python/`), 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.

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 have shown 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 [4] 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 [12], 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 $\{10^{-10}, 10^{-4}, 10^{-3}, 5 \cdot 10^{-3}, 10^{-2}, 5 \cdot 10^{-2}, 10^{-1}, 5 \cdot 10^{-1}\}$.

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

3.3 | Statistical test

To compare the AUC values provided by different approaches, we applied the corrected resampled t-test [23]. The problem in applying standard statistical 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 [8] and McNemar's test for the classification accuracy [23]. 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.,

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

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

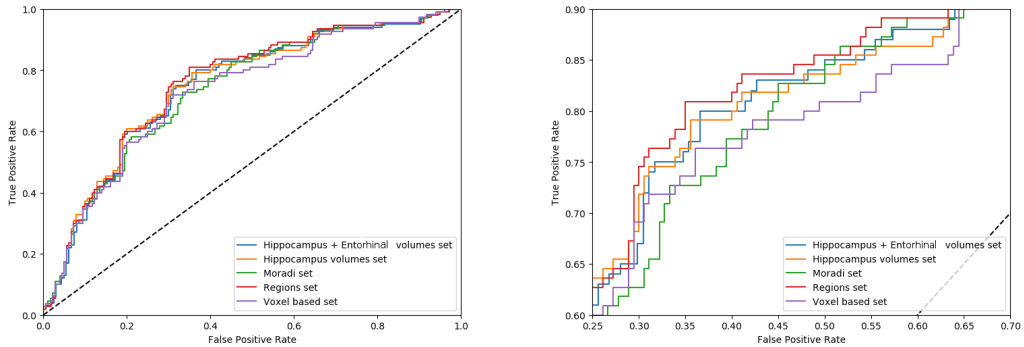
we can estimate the statistic of the test, t, as:

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

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

4 | RESULTS

Tables 3, 4 and 5 show, respectively, 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 (area under the curve), ACC (accuracy), SEN (sensitivity), SPE (specificity), as well as three 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. LR results over the same set of features).



(a) Complete ROC curves.

(b) A magnification when FPR is between 0.25 and 0.70.

FIGURE 1 ROC curves corresponding to distinct the features sets used in RLR classification with the non-QC dataset. 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 1b), 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 was not visualized as it always performed worse than the voxel features without PCA. The same principle will be followed in later figures.

Regarding the use of two different classifiers, differences between AUCs of SVM and RLR were not significant.

TABLE 3 Cross-validated performance measures 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	Hippocampus	No	73.50 %	63.15 %	93.00 %	18.16 %			0.873
SVM	Hippocampus	Yes	77.31 %	66.42 %	90.54 %	30.20 %	0.052		0.819
SVM	Hippo. and entor.	No	75.91 %	63.69 %	94.55 %	17.14 %			0.314
SVM	Hippo. and entor.	Yes	78.59%	61.77 %	97.82 %	7.29 %	0.055	0.492	0.285
SVM	Voxel features	No	63.19 %	60.51 %	91.36 %	13.98 %			0.649
SVM	Voxel features	Yes	66.67 %	61.65 %	91.82 %	16.20 %	0.175	0.042	0.923
SVM	PCA VF	No	63.50 %	59.88 %	92.73 %	10.23 %			0.604
SVM	PCA VF	Yes	65.53 %	60.75 %	92.64 %	12.54 %	0.535	0.035	0.509
SVM	Moradi features	No	71.92%	63.54%	92.63 %	19.66 %			0.025
SVM	Moradi features	Yes	75.08 %	62.32 %	97.10 %	9.86 %	0.243	0.610	0.001
SVM	Region features	No	74.06 %	64.67%	92.27 %	22.91 %			0.739
SVM	Region features	Yes	77.34%	69.29 %	88.45 %	37.27 %	0.241	0.990	0.759
LR	Hippocampus	No	73.38 %	68.41 %	82.00 %	47.86 % %			
LR	Hippocampus	Yes	77.19 %	71.31 %	83.91 %	52.43 %	0.042		
LR	Hippo. and entor.	No	75.13 %	71.19 %	84.54%	51.09 %			
LR	Hippo. and entor.	Yes	77.77 %	73.01 %	84.10 %	56.38 %	0.085	0.767	
LR	Voxel features	No	64.40 %	61.91 %	76.18 %	40.43 %			
LR	Voxel features	Yes	66.94 %	63.57 %	77.36 %	40.43 %	0.268	0.029	
LR	PCA VF	No	61.95 %	59.12 %	78.73 %	26.68 %			
LR	PCA VF	Yes	63.40 %	61.33 %	80.82 %	31.91 %	0.600	0.006	
LR	Moradi features	No	65.83 %	63.28 %	75.00 %	42.73 %			
LR	Moradi features	Yes	67.89 %	65.52 %	77.54 %	57.50 %	0.348	0.038	
LR	Region features	No	74.57 %	70.67 %	81.91 %	53.70 %			
LR	Region features	Yes	77.91 %	71.93 %	81.45 %	57.50 %	0.171	0.839	

TABLE 4 Cross-validated performance measures 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	Hippocampus volumes	No	70.29 %	63.09 %	96.02 %	9.10 %			0.708
SVM	Hippocampus volumes	Yes	75.57 %	63.94 %	95.39 %	12.30 %	0.047		0.398
SVM	Hippo. and entor. vol.	No	73.23 %	64.57 %	98.16 %	3.50 %			0.358
SVM	Hippo. and entor. vol.	Yes	76.05 %	63.73 %	96.86 %	9.50 %	0.088	0.744	0.253
SVM	Voxel based features	No	68.55%	62.11 %	97.10 %	4.80 %			0.695
SVM	Voxel based features	Yes	69.55%	63.79 %	96.55 %	10.10 %	0.659	0.169	0.489
SVM	PCA VF	No	68.41 %	62.30 %	96.19 %	6.80 %			0.328
SVM	PCA VF	Yes	69.13 %	63.90 %	95.12 %	12.70 %	0.754	0.149	0.549
SVM	Moradi features	No	73.26%	64.20%	96.04 %	12.00 %			0.078
SVM	Moradi features	Yes	75.72%	63.40 %	96.89 %	8.50 %	0.180	0.965	0.288
SVM	Region features	No	73.11%	65.29 %	90.07 %	24.60 %			0.123
SVM	Region features	Yes	76.89%	66.94 %	95.30 %	20.50 %	0.091	0.680	0.101
LR	Hippocampus volumes	No	70.95 %	67.00 %	88.04 %	32.40 %			
LR	Hippocampus volumes	Yes	74.95 %	67.68 %	86.34 %	37.00 %	0.046		
LR	Hippo. and entor. vol.	No	72.59 %	69.72 %	85.88 %	43.20 %			
LR	Hippo. and entor. vol.	Yes	75.31 %	70.61 %	84.98%	47.00 %	0.094	0.801	
LR	Voxel based features	No	69.46 %	66.63 %	83.42 %	39.10 %			
LR	Voxel based features	Yes	71.34 %	66.99 %	82.68 %	41.30 %	0.370	0.394	
LR	PCA VF	No	66.02 %	64.50 %	86.28 %	28.80 %			
LR	PCA VF	Yes	67.58 %	64.68 %	87.02 %	28.10 %	0.536	0.058	
LR	Moradi features	No	69.94%	67.77 %	83.81 %	41.50 %			
LR	Moradi features	Yes	74.04 %	70.84 %	86.79 %	44.70 %	0.068	0.798	
LR	Region features	No	76.38 %	71.27 %	85.97 %	47.10 %			
LR	Region features	Yes	79.58 %	71.73 %	84.07 %	51.50 %	0.120	0.060	

TABLE 5 Cross-validated performance measures 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	Hippocampus volumes	No	70.51 %	67.35 %	85.06 %	38.30 %			0.330
SVM	Hippocampus volumes	Yes	74.99 %	68.86 %	81.91 %	47.40 %	0.026		0.759
SVM	Hippo. and entor. vol.	No	72.43 %	69.03 %	81.36 %	48.8 %			0.968
SVM	Hippo. and entor. vol.	Yes	75.40 %	71.71 %	82.20 %	53.50 %	0.065	0.777	0.880
SVM	Voxel based features	No	67.10%	61.71 %	79.86 %	32.10 %			0.939
SVM	Voxel based features	Yes	68.35%	62.46 %	80.59 %	32.90 %	0.506	0.106	0.548
SVM	PCA VF	No	66.78 %	62.15 %	79.56 %	33.70 %			0.621
SVM	PCA VF	Yes	67.98 %	63.14 %	80.18 %	35.30 %	0.613	0.085	0.784
SVM	Moradi features	No	72.85%	68.93%	84.49 %	43.40 %			0.356
SVM	Moradi features	Yes	75.00%	70.09 %	83.99 %	47.30 %	0.292	0.997	0.650
SVM	Region features	No	72.55 %	69.16%	82.73 %	46.90 %			0.122
SVM	Region features	Yes	75.98%	71.01 %	86.94 %	44.90 %	0.105	0.763	0.076
LR	Hippocampus volumes	No	70.96 %	66.28 %	88.92 %	29.10 %			
LR	Hippocampus volumes	Yes	74.85 %	69.12 %	84.10 %	44.50 %	0.050		
LR	Hippo. and entor. vol.	No	72.40 %	69.55 %	85.50 %	43.30 %			
LR	Hippo. and entor. vol.	Yes	75.50 %	70.50 %	84.68 %	47.20 %	0.056	0.650	
LR	Voxel based features	No	67.33 %	64.81 %	80.76 %	38.60 %			
LR	Voxel based features	Yes	69.95 %	66.15 %	80.21 %	43.10 %	0.235	0.244	
LR	PCA VF	No	65.42 %	63.52 %	87.46 %	24.30 %			
LR	PCA VF	Yes	67.35 %	64.96 %	87.76 %	27.60 %	0.513	0.059	
LR	Moradi features	No	71.08 %	68.60 %	85.35 %	41.10 %			
LR	Moradi features	Yes	74.10 %	70.42 %	85.94 %	45.00 %	0.125	0.836	
LR	Region features	No	75.91 %	71.01 %	84.52 %	48.80 %			
LR	Region features	Yes	79.41 %	72.07 %	84.24 %	52.10 %	0.123	0.717	

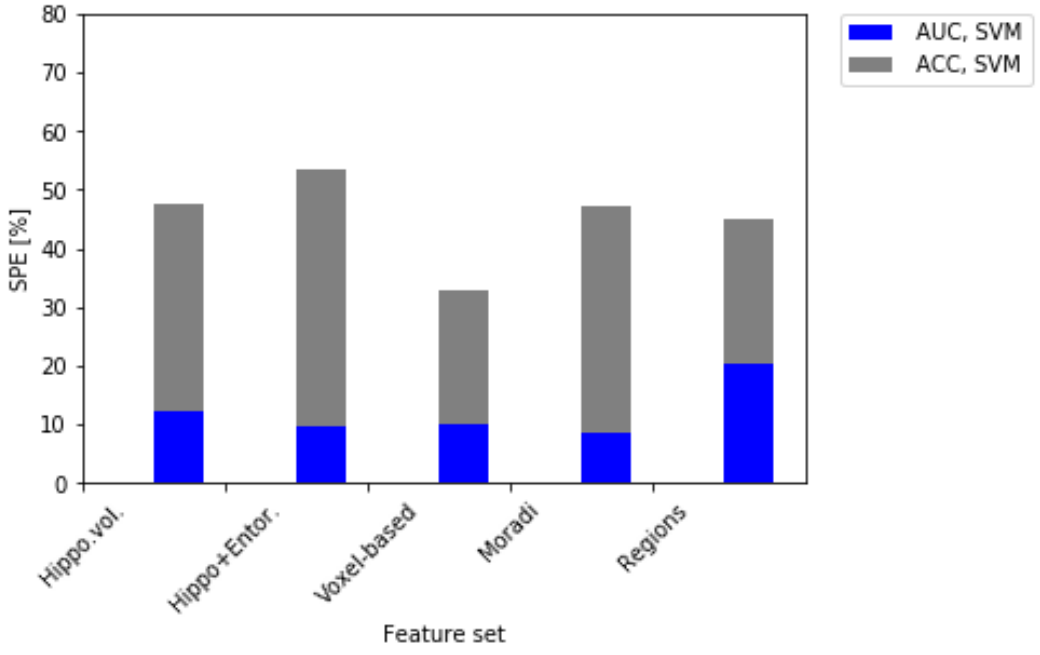


FIGURE 2 Specificity values of SVM classifiers when AUC and ACC were used for model selection. The models selected with ACC resulted in specificity values close 50 % whereas the models selected with AUC resulted in very low specificity values.

However, SVM yielded low specificity values and the relation between SPE and SEN was more balanced with the LR 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 avoids 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 Tables4 and 5 reveals, the final AUC values did not markedly differ between the two model selectors.

We evaluated the effects of age removal on the feature sets. For this purpose, Figure 3 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 3c). However, as visible in Tables 3 and 4, significant improvement (p-value < 0.1) was observed only for hippocampus and hippocampus and entorhinal volume feature sets.

Finally, Figure 4 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.

5 | 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 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, i.e., that is that the datasets contained twice the number of pMCIs (subjects who converted to AD) compared to sMCIs (subjects who remained as MCIs). The evaluation process was carried out with a nested fold CV repeated 10 times ensuring the insensitivity of the conclusions to random train/test division of the holdout method used previously [6]. Selecting the parameters of the classifiers 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* (see `https://github.com/MartaGomez/Regions-list-/wiki/Regions-list` for a detailed description) 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 [27] 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 age related effects from MRI data was not considered; second, the set of pMCI patients was about half of ours; and, third, the combined volumetrics 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 [27] to 80 % in the present study.

Retico et al. found that the voxel based VBM features best discriminate between sMCI and pMCI after applying Recursive Feature Elimination (RFE) [25]. However, again, the maximum accuracy in [25] 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 [31], concluded that the combination of different feature representations resulted into a better classification accuracy. Again, the classification accuracies were lower than in the present work. Moreover, [31] selected classifier hyperparameters based on test data that may cause upward bias in the reported accuracies [10].

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. 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 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 that was made 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 Table 5 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,

particularly as it seems to be a problem of a specific class of learning algorithms, that invites further research. However, as this issue is not central to the goals of this work, we do not analyze it further.

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 4. 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 3 illustrates, age removal always led to better classification performances, although the improvements were not always statistically significant. This agrees with a recent work of [28] which demonstrated the same for NC vs. MCI classification.

| 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, Alzheimer's Association; Alzheimer's 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.

References

REFERENCES

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

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

[3] Casanova R, Whitlow CT, Wagner B, Williamson J, Shumaker SA, Maldjian JA, et al. High dimensional classification of structural MRI Alzheimer's Disease data based on large scale regularization. Frontiers in neuroinformatics 2011;5.

[4] Chang CC, Lin CJ. LIBSVM: A library for Support Vector Machines. ACM Trans Intell Systems Tech 2011;2:27.

[5] Chu C, Hsu AL, Chou KH, Bandettini P, Lin C, Initiative ADN, et al. Does feature selection improve classification accuracy?

Impact of sample size and feature selection on classification using anatomical magnetic resonance images. Neuroimage 2012;60(1):59–70.

[6] Cuingnet R, Gerardin E, Tessieras J, Auzias G, Lehéricy S, Habert MO, et al. Automatic classification of patients with Alzheimer's Disease from structural MRI: a comparison of ten methods using the ADNI database. Neuroimage 2011;56(2):766–781.

[7] Delacourte A, David J, Sergeant N, Buee L, Wattez A, Vermersch P, et al. The biochemical pathway of neurofibrillary degeneration in aging and Alzheimer's Disease. Neurology 1999;52(6):1158–1158.

[8] Dietterich TG. Approximate statistical tests for comparing supervised classification learning algorithms. Neural computation 1998;10(7):1895–1923.

[9] Dukart J, Schroeter ML, Mueller K, Initiative ADN, et al. Age correction in dementia–matching to a healthy brain. PloS one 2011;6(7):e22193.

[10] Eskildsen SF, Coupé P, García-Lorenzo D, Fonov V, Pruessner JC, Collins DL. Prediction of Alzheimer's disease in subjects with mild cognitive impairment from the ADNI cohort using patterns of cortical thinning. NeuroImage 2013;65:511–521.

[11] Franke K, Ziegler G, Klöppel S, Gaser C, Initiative ADN, et al. Estimating the age of healthy subjects from T1-weighted MRI scans using kernel methods: Exploring the influence of various parameters. Neuroimage 2010;50(3):883–892.

[12] Friedman JH, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. J Stat Software 2010;33(1):1–22.

[13] Frisoni GB, Fox NC, Jack CR, Scheltens P, Thompson PM. The clinical use of structural MRI in Alzheimer Disease. Nature Reviews Neurology 2010;6(2):67–77.

[14] Gaser C, Franke K, Klöppel S, Koutsouleris N, Sauer H, Initiative ADN, et al. BrainAGE in mild cognitive impaired patients: predicting the conversion to Alzheimer's Disease. PloS ONE 2013;8(6):e67346.

[15] Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982;143(1):29–36.

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

[17] Johnson KA, Fox NC, Sperling RA, Klunk WE. Brain imaging in Alzheimer Disease. Cold Spring Harbor perspectives in medicine 2012;2(4):a006213.

[18] Klöppel S, Stonnington CM, Chu C, Draganski B, Scahill RI, Rohrer JD, et al. Automatic classification of MR scans in Alzheimer's Disease. Brain 2008;131(3):681–689.

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

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

[21] Mosconi L, Brys M, Glodzik-Sobanska L, De Santi S, Rusinek H, De Leon MJ. Early detection of Alzheimer's Disease using neuroimaging. Experimental gerontology 2007;42(1):129–138.

[22] Mwangi B, Tian TS, Soares JC. A review of feature reduction techniques in neuroimaging. Neuroinformatics 2014;12(2):229–244.

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

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

[25] Retico A, Bosco P, Cerello P, Fiorina E, Chincarini A, Fantacci ME. Predictive Models Based on Support Vector Machines: Whole-Brain versus Regional Analysis of Structural MRI in the Alzheimer's Disease. Journal of Neuroimaging 2015;25(4):552–563.

[26] Serrano-Pozo A, Frosch MP, Masliah E, Hyman BT. Neuropathological alterations in Alzheimer Disease. Cold Spring Harbor perspectives in medicine 2011;1(1):a006189.

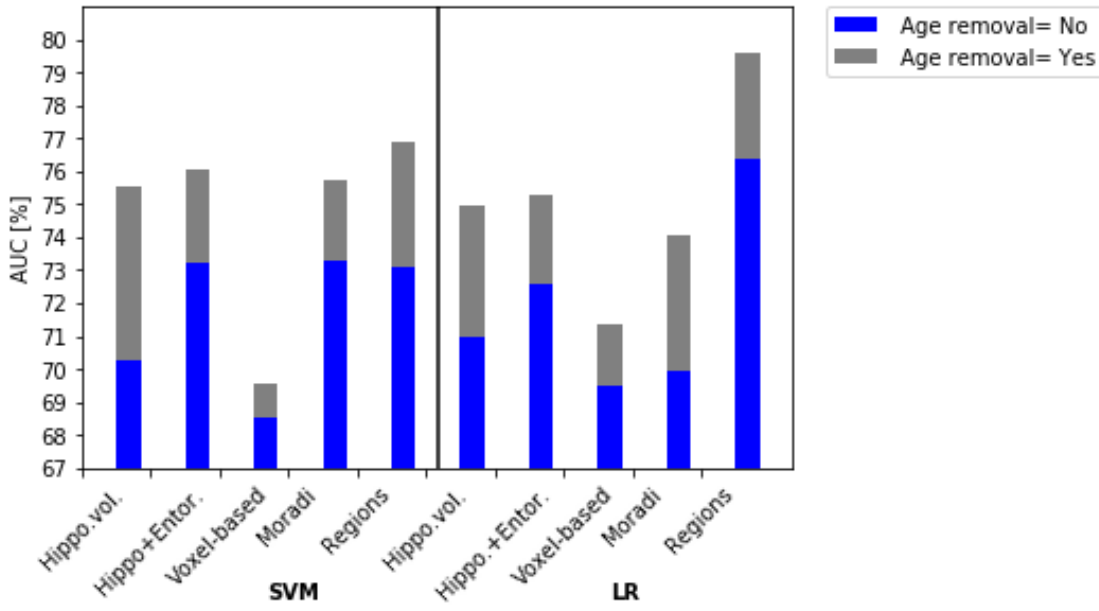
[27] Tanpitukpongse T, Mazurowski M, Ikhena J, Petrella J. Predictive Utility of Marketed Volumetric Software Tools in Subjects at Risk for Alzheimer Disease: Do Regions Outside the Hippocampus Matter? American Journal of Neuroradiology 2017;38(3):546–552.

[28] Tohka J, Moradi E, Huttunen H. Comparison of feature selection techniques in machine learning for anatomical brain MRI in dementia. Neuroinformatics 2016;p. in press.

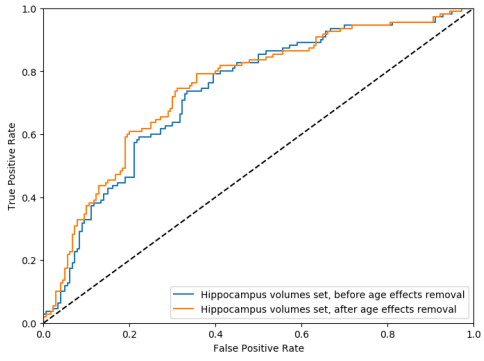
[29] Vos SJ, Verhey F, Frölich L, Kornhuber J, Wiltfang J, Maier W, et al. Prevalence and prognosis of Alzheimer's Disease at the mild cognitive impairment stage. Brain 2015;138(5):1327–1338.

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

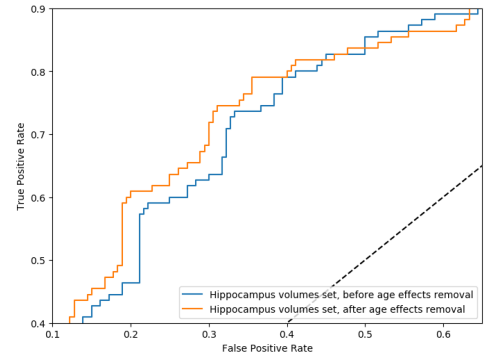
[31] Wolz R, Julkunen V, Koikkalainen J, Niskanen E, Zhang DP, Rueckert D, et al. Multi-method analysis of MRI images in early diagnostics of Alzheimer's Disease. PloS one 2011;6(10):e25446.



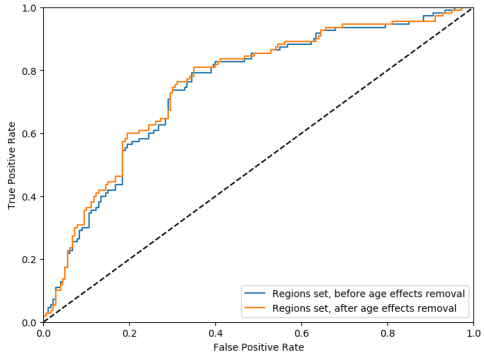
(a) AUC comparison



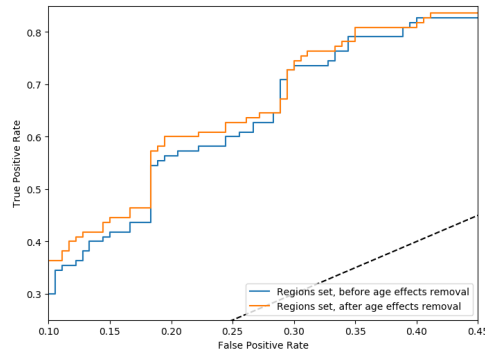
(b) ROC for RLR classifier using hippocampus volumes



(c) Zoom over the ROC for RLR using hippocampus volumes



(d) ROC for RLR classifier using region features



(e) Zoom over the ROC for RLR using region features

FIGURE 3 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 LR classifier using region features. Age removal improved predictions in all cases.

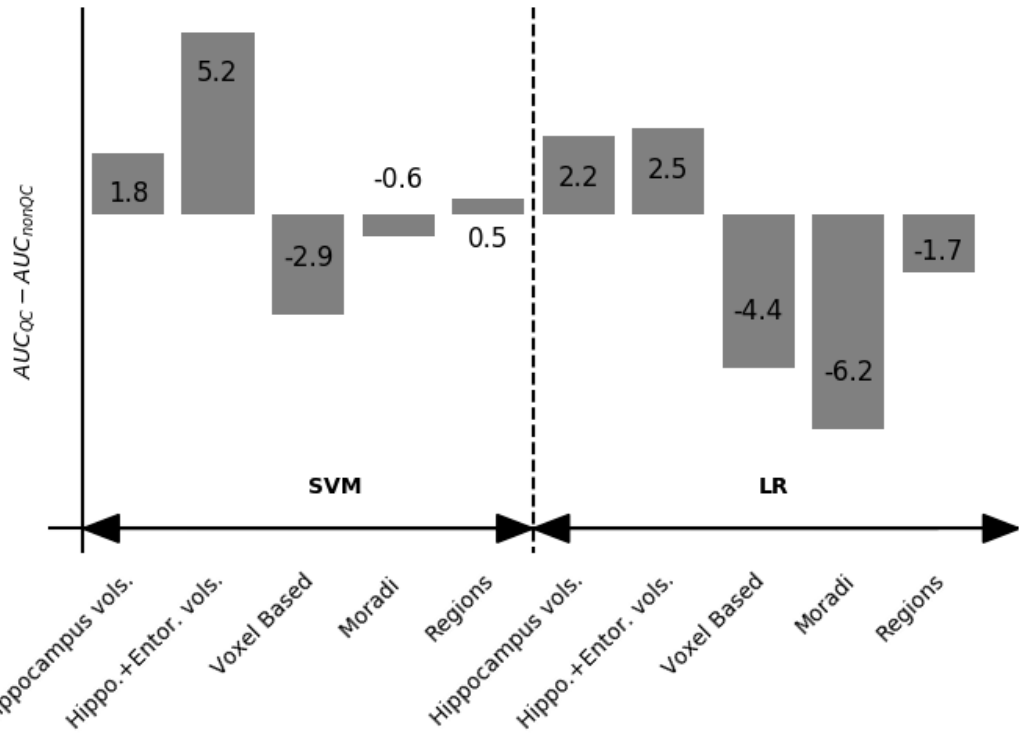


FIGURE 4 Differences between the AUC values with the QC dataset and the non-QC dataset for SVM (left) and RLR (right).