Feature Selection Based on the SVM Weight Vector for Classification of Dementia

Esther E. Bron, Marion Smits, Wiro J. Niessen, Stefan Klein,
for the Alzheimer's Disease Neuroimaging Initiative

***Abstract*—Computer-aided diagnosis of dementia using a support vector machine (SVM) can be improved with feature selection. The relevance of individual features can be quantified from the SVM weights as a significance map (p-map). Although these p-maps previously showed clusters of relevant voxels in dementia-related brain regions, they have not yet been used for feature selection. Therefore, we introduce two novel feature selection methods based on p-maps using a direct approach (filter) and an iterative approach (wrapper).**

To evaluate these p-map feature selection methods, we compared them with methods based on the SVM weight vector directly, t-statistics and expert knowledge. We used MRI data from the Alzheimer's Disease Neuroimaging Initiative classifying Alzheimer's disease (AD) patients, mild cognitive impairment (MCI) patients who converted to AD (MCIc), MCI patients who did not convert to AD (MCInc), and cognitively normal controls (CN). Features for each voxel were derived from gray matter morphometry.

Feature selection based on the SVM weights gave better results than t-statistics and expert knowledge. The p-map methods performed slightly better than those using the weight vector. The wrapper method scored better than the filter method. Recursive feature elimination based on the p-map improved most for AD-CN: the area under the receiver-operating-characteristic curve (AUC) significantly increased from 90.3% without feature selection to 92.0% when selecting 1.5%-3% of the features. This feature selection method also improved the other classifications: AD-MCI 0.1% improvement in AUC (not significant), MCI-CN 0.7%, and MCIc-MCInc 0.1% (not significant).

Although the performance improvement due to feature selection was limited, the methods based on the p-map generally had the best performance and were therefore better in estimating the relevance of individual features.

***Index Terms*—Computer-aided diagnosis, Dementia, Feature selection, Recursive feature elimination, Significance maps, Support vector machine.**

I. Introduction

DEMENTIA affects 35.6 million individuals over 60 years of age worldwide as was estimated in 2010 [1]. Many of these individuals are never diagnosed [2], while an early and accurate diagnosis is important for providing optimal care. Accurate diagnostic methods are also important for research into understanding the disease process and developing new treatments [3], [4].

Computer-aided diagnosis methods can aid the diagnosis of neurodegenerative disease as they are trained on reference data and therefore potentially make use of subtle group differences that are not noted during qualitative visual inspection of brain imaging data [5]. These methods apply machine learning approaches to classify two or more classes, e.g. to distinguish Alzheimer's disease (AD) patients from normal (CN) controls. For this classification, the machine-learning methods are trained on features derived from imaging or related data.

For dementia diagnosis based on structural MRI, a survey of all recent work showed that the classification accuracy for AD-CN generally is 80-90% [6]. Many of the dementia classification methods used voxel-wise approaches based on brain morphometric analyses [6]–[8]. These voxel-wise approaches provide high-dimensional feature vectors of sizes up to ∼1 million features, while typically the sample size of such studies is much lower, in the order of hundreds, which can result in suboptimal performances. Therefore, researchers have explored feature selection methods for reducing dimensionality and improving performance [8], [9].

Although there exist many data-driven methods for feature selection, it can be difficult to choose the best method as the effectiveness depends on the specific application and data set [10]. Most feature selection methods rank the features based on a specific criterion that reflects their degree of relevance [11]. These feature selection methods can be divided into three main types of methods [6], [12]: 1) filter methods, 2) wrapper methods, and 3) embedded methods. Filter methods perform feature selection as a preprocessing step prior to the classification and compute some relevance measure on the training set to remove the least relevant features from the data set. A commonly used filter method is to perform a t-test for every feature [6], [9], [13]–[15]. Wrapper methods are iterative methods in which the classifier is trained several times using the feedback from every iteration to select a subset of features for the next iteration. A well-known wrapper method is recursive feature elimination (RFE) [16], in which the features that are ranked the lowest are iteratively removed. For embedded methods, the feature selection is incorporated in the classifier and selection is performed during training. In this work, we focus on filter and wrapper methods.

The support vector machine (SVM) classifier is frequently used for classification in medical imaging including computer-aided diagnosis in MR brain imaging [6], [8], [17], [18]. In

E.E. Bron, W.J. Niessen, and S. Klein are with the Biomedical Imaging Group Rotterdam, Departments of Medical Informatics and Radiology, Erasmus MC, Rotterdam, the Netherlands, e-mail: e.bron@erasmusmc.nl.

M. Smits is with the department of Radiology, Erasmus MC, Rotterdam, the Netherlands

W.J. Niessen is also with Imaging Physics, Applied Sciences, Delft University of Technology, the Netherlands

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.

training an SVM classifier, a weight vector is computed on the training data. This weight vector can be used as a importance measure of the features to the classifier. Therefore, it can serve as ranking measure for feature selection that can be used in a filter method or in a wrapper method. Feature selection using the SVM weight vector has been studied extensively in machine learning research [10], [16], [19]–[21] and has also been applied in neuroimaging [9], [22], [23].

The ranking of features based on the SVM weight vector may be suboptimal since the weights are not the result of a statistical test and therefore do not necessarily reflect the significance of a specific feature [24]. Using permutation testing, the SVM weight vector can be calibrated by taking into account the null distribution of the weights [17], [18]. The permutation test computes a p-value for every feature indicating the significance of its contribution to the classifier. As every feature represents a voxel, these p-values can be combined into a significance map (p-map) which reflects the regions consistently influencing the classifier. In previous work, we showed that these p-maps find clusters of significantly different voxels in regions known to be involved in neurodegenerative diseases underlying dementia [25]. Based on these results, it seems attractive to use the p-map for feature selection.

The SVM p-map has not been used for feature selection before, probably because SVM p-map computation with permutation testing is time-consuming. However, a recently published method for analytic estimation of significance maps [24] makes it computationally feasible to use p-maps for feature selection in both a filter and a wrapper approach. Like feature selection on the SVM weight vector, the p-map methods are purely data-driven and are from a methodological point of view closely linked to the SVM classifier, rendering interpretation clear.

In this paper, we validated several feature selection methods that are based on the weight vector of the SVM classifier. We evaluated feature selection using two relevance measures: 1) the SVM weight vector and 2) the SVM p-maps estimated with the analytic implementation as described in [24]. For both relevance measures, we evaluated filter and wrapper feature selection. We compared these methods to methods based on t-statistics and a method based on prior knowledge. For evaluation, we performed a classification experiment of AD, mild cognitive impairment (MCI) and CN based on T1-weighted MR scans using data from the Alzheimer's Disease Neuroimaging Initiative (ADNI).

This work is an extension of our conference paper [26], in which we presented an initial evaluation of the filter p-map feature selection method. That work was limited to comparison with the t-test and prior knowledge. We used a fixed threshold ($\alpha = 0.05$) on the p-map and t-test to select the significant features and compared the methods using different numbers of selected features. For the more thorough validation in this paper, we added other SVM-based methods and an additional method based on t-statistics to the comparison. We also analyzed the features that the methods selected. Finally, we now keep the number of features constant across methods.

II. METHODS

A. Support vector machine

The SVM classifier is based on maximization of the margin around the hyperplane ($\boldsymbol{w}^T\boldsymbol{x} + b$) separating samples of the different classes [27]. Each sample $i = 1, ..., m$ consists of an N-dimensional feature vector $\boldsymbol{x}_i$ and a class label $y_i \in \{+1, -1\}$. The maximization of the margin corresponds to the following minimization:

$$\boldsymbol{w}^*, b^*, \boldsymbol{\xi}^* = \arg\min_{\boldsymbol{w},b,\boldsymbol{\xi}} \frac{1}{2}||\boldsymbol{w}||^2 + C\sum_{i=1}^{m}\xi_i \quad (1)$$
$$s.t. \quad y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) \geq 1 - \xi_i; \quad \xi_i \geq 0; \quad i = 1, ..., m$$

In this soft-margin SVM equation, ξ_i is a penalty for misclassification or classification within the margin. Parameter C sets the weight of this penalty. The resulting weight vector $\boldsymbol{w}^*$ encodes the contributions of all features to the classifier.

B. Significance of the SVM weight vector

The p-value quantifies the significance of each feature's contribution to the SVM classifier. As every feature is a voxel, the p-values can be combined into a p-map image. To obtain p-values, permutation testing can be used to estimate a null distribution on the weight vector ($\boldsymbol{w}$) [17], [18]. Permutation testing, however, requires the training of a large number of SVM classifiers, which renders it very time-consuming for high-dimensional feature vectors.

A faster solution for estimation of the SVM p-map was presented by *Gaonkar et al.* [24], who derived an analytic approximation of the null distribution of $\boldsymbol{w}$. For this approximation, the SVM classifier is simplified by making two assumptions. First, under the assumption that the classes are separable, which is true if many features and a relatively small number of samples are used, the soft-margin SVM can be simplified to a hard-margin SVM, which does not use the misclassification penalty ξ_i. Second, under the assumption that for most permutations most samples will be support vectors, the hard-margin SVM can be simplified further to a least-squares SVM, which has a closed-form solution $\boldsymbol{w} = \boldsymbol{K}\boldsymbol{y}$, with:

$$\boldsymbol{K} = \boldsymbol{X}^{\boldsymbol{T}}\Big[(\boldsymbol{X}\boldsymbol{X}^{\boldsymbol{T}})^{-1} + (\boldsymbol{X}\boldsymbol{X}^{\boldsymbol{T}})^{-1}\boldsymbol{J}(-\boldsymbol{J}^{\boldsymbol{T}}(\boldsymbol{X}\boldsymbol{X}^{\boldsymbol{T}})^{-1}\boldsymbol{J})^{-1}\boldsymbol{J}^{\boldsymbol{T}}(\boldsymbol{X}\boldsymbol{X}^{\boldsymbol{T}})^{-1}\Big] \quad (2)$$

where $\boldsymbol{J}$ is a column matrix of ones and the matrix $\boldsymbol{X}$ contains one feature vector in each row. Given a sufficiently high number of subjects, the probability density function of every feature (j) can be approximated with a Gaussian distribution:

$$w_j \xrightarrow{d} \mathcal{N}\Big((2q-1)\sum_{i=1}^{m}K_{ij}, (4q-4q^2)\sum_{i=1}^{m}K_{ij}^2\Big) \quad (3)$$

where q is the fraction of the data with class label $y_i = +1$. A p-value for each feature is obtained by testing $\boldsymbol{w}^*$ against the analytic null distribution in (3). The experiments by Gaonkar et al. [24] showed that this approximation results in p-maps that are very similar to those obtained with permutation testing.

C. Feature selection using the SVM weight vector

In this work, we evaluated feature selection methods that are based on the SVM weight vector $\boldsymbol{w}^*$. Since these feature selection methods use information on which features contribute most to the classifier, they are expected to reduce features in a meaningful way. Intuitively, using such an SVM-based feature selection method prior to SVM classification is an attractive approach, as in this way the feature selection and the classification use the same decision model.

We defined four methods for feature selection on the SVM weights: 1) a filter method on the weight vector (*W-map*), 2) a wrapper method on the weight vector (*RFE W-map*), 3) a filter method on the significance of the weight vector (*P-map*), and 4) a wrapper method on the significance of the weight vector (*RFE P-map*). These methods are detailed below.

1) SVM weight map (W-map): The SVM weight vector $\boldsymbol{w}^*$ encodes the contributions of all features to the classifier. The highest absolute weights $|w_j^*|$ are assigned to the features j that have the largest contribution in the classification. The *W-map* image is used in a filter-based feature selection method by simply selecting the features with the highest absolute weights.

2) Recursive feature elimination using the SVM weight map (RFE W-map): Recursive feature elimination (RFE) [16] is a feature selection method originally developed in genetics, but it has been used in many applications including computer-aided diagnosis based on MRI [9]. RFE is not specifically developed for the SVM classifier, but it can use the SVM weight vector as its elimination criterion. Instead of 'naively' ranking the weights like in the *W-map* method, RFE uses a wrapper approach that removes a subset of features with the lowest classifier weights in every iteration. The approach is a form of backward feature elimination [28], but it removes multiple features at the same time to make the approach computationally feasible for high-dimensional feature spaces.

Similar to *W-map*, *RFE W-map* uses the SVM weight vector as its relevance measure. For genetic data, Guyon et al. [16] showed that *RFE W-map* outperformed the *W-map* approach. Unlike *W-map*, which orders the features on their individual relevance, RFE takes usefulness of the features into account by looking at feature sets instead of individual features. This is most important when the features are highly correlated. In that case, the feature selection methods should not select highly correlated features that have no additional information, which a filter method such as *W-map* might do. However, because of the iterative approach, *RFE W-map* is more likely to select features that are complementary to other features, but that might not individually have the highest relevance [16].

In our application, we use features based on voxel-wise morphometry of the gray matter (GM). These features are expected to be highly correlated, especially between neighboring voxels. Therefore, *RFE W-map* is expected to have some advantage over *W-map* in our application.

3) SVM significance map (P-map): The *W-map* and *RFE* methods are both based on $\boldsymbol{w}^*$, but do not perform any statistical testing. The analytic method to estimate the SVM p-map, which we explained in Section II-B, performs a significance test for each feature in the SVM classifier. In a previous conference paper, we introduced this p-map as a novel method for feature selection [26]. This method uses the p-map to select features that are most significant for the final classification. The advantage of this method over *W-map* is that it takes into account the null distribution of $\boldsymbol{w}^*$. This calibrates the weights and can make the ordering of the features more robust.

4) Recursive feature elimination using the SVM significance map (RFE P-map): This method combines the advantages of the previously described methods, performing both a wrapper approach and statistical testing. *RFE P-map* applies recursive feature elimination to the SVM p-map. To the best of our knowledge, this method has not been proposed before.

D. Feature selection using t-statistics

We compared the SVM weight vector feature selection methods with methods that use a more commonly applied relevance measure: t-statistics. These methods perform a t-test on the training set for every voxel. The resulting t-statistic can then be used in a filter-based approach (*T-test*). In addition, we can compute the t-statistic in a permutation test, similar to *P-map*. While the standard t-test makes the assumption that the data has a Gaussian distribution and is independently drawn, the permutation t-test does not make these assumptions. Therefore, we apply this randomized t-statistic in addition as a filter (*T-map*). For the permutation testing on the t-statistic, no analytic derivation is available, hence this method is more time-consuming than the other described methods. A wrapper-based approach, such as RFE, would have no added value for the t-statistics criteria, since these measures are univariate: the t-statistic is computed for each feature individually and does not give different results over several iterations.

E. Feature selection using prior knowledge (ROI)

The last feature selection method is region-of-interest (ROI) selection based on prior knowledge. In this method, we use the voxel-wise features only from certain ROIs that have been associated with dementia. We use the following ROIs (see Fig. 1): 1) Cingulate gyrus (CG), 2) Hippocampus including amygdala (HC), 3) Parahippocampal gyrus (PHG), 4) Fusiform gyrus (FG), 5) Superior parietal gyrus (SPG), 6) Middle/inferior temporal gyrus (MITG), 7) Temporal lobe (TL) including FG and MITG, 8) HC + PHG, and 9) TL + HC + PHG. The choice of these ROIs was based on those previously used for a similar study [9].

III. Experiments

A. Data

For the classification experiments, we used data from the ADNI[1]. The inclusion criteria for participants were defined in the ADNI GO protocol[2]. The ADNI was launched in 2003 by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, the Food and Drug Administration, private pharmaceutical companies and non-profit organizations, as a $60 million, 5-year public-private partnership. The primary goal of ADNI has been to

[1] http://adni.loni.usc.edu

[2] http://www.adni-info.org/Scientists/Pdfs/ADNI_Go_Protocol.pdf

test whether serial MRI, positron emission tomography, other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of MCI and early AD. Determination of sensitive and specific markers of very early AD progression is intended to aid researchers and clinicians to develop new treatments and monitor their effectiveness, as well as lessen clinical trial time and cost.

The used cohort is selected based on the paper by *Cuingnet et al* [8], who published a list of subjects included in their study. This cohort consists of AD patients, MCI patients that converted to AD within 18 months (MCIc), MCI patients that did not convert to AD within 18 months (MCInc), and CN. The participants were 137 AD patients (67 male, age: 76.0 ± 7.3 yrs, mini mental-state examination (MMSE) score: 23.2 ± 2.0), 76 MCIc (43 male, 74.8 ± 7.4 yrs, MMSE: 26.5 ± 1.9), 134 MCInc (84 male, 74.5 ± 7.2 yrs, MMSE: 27.2 ± 1.7), and 162 CN (76 male, 76.3 ± 5.4 yrs, MMSE: 29.2 ± 1.0). Acquisition of the data was performed according to the ADNI protocol [29]. T1w imaging was acquired at 1.5T with a voxel size of $\sim 1\text{mm}^3$.

B. Image processing

Probabilistic tissue segmentations were obtained for white matter, GM and cerebrospinal fluid using SPM8 (Statistical Parametric Mapping, UK) [30].

We constructed a template space specifically for the used data set based on a subset of 150 T1w images (81 CN, 69 AD [8]). To construct this template space, we derived the coordinate transformations from the template space to the subject's space from pairwise registration of the images in the subset [31]. We performed pairwise registrations with consecutively a rigid (including isotropic scaling), affine, and non-rigid B-spline transformation model. The non-rigid B-spline registration used a three-level multi-resolution framework with isotropic control-point spacing of 24, 12, and 6 mm at the three resolution levels respectively. Registrations were performed with Elastix registration software [32] by maximizing mutual information [33] within a brain mask [34]. A template image was created by averaging the deformed individual images. To transform the other subjects' images to template space, coordinate transformations were derived from pairwise registrations to the subset. The registrations to the template space were visually inspected to check if they were correct. This template space construction is detailed in [25].

We used multi-atlas segmentation to segment brain masks and the ROIs for the feature selection method based on prior knowledge. The segmentations were performed for every subject individually and subsequently transformed to template space. For the individual multi-atlas segmentations, we used 30 labeled T1w images, each containing 83 manually-segmented regions [35], [36]. The brain masks of the 30 atlas images were obtained with the Brain Extraction Tool (BET) [34]. These brain masks which were visually inspected and BET parameters were adjusted if necessary. The atlas images were registered to the subjects' image using a rigid, affine, and non-rigid B-spline transformation model consecutively. The labels of the regions and brain masks were fused using majority voting [37]. Using the definition of [35], [36], the listed regions were combined to obtain the nine ROIs defined in Section II-E. The numbers in brackets indicate the number of GM-containing voxels, i.e. the number of features, within an ROI:

1) CG: Cingulate gyrus anterior (supragenual) part right/left (r/l), Cingulate gyrus posterior part r/l, Subgenual anterior cingulate gyrus r/l, Pre-subgenual anterior cingulate gyrus r/l (45870 voxels)
2) HC: Hippocampus r/l, Amygdala r/l (9325)
3) PHG: Gyri parahippocampalis et ambiens r/l (11736)
4) FG: Lateral occipitotemporal gyrus (gyrus fusiformis) r/l (11115)
5) SPG: Superior parietal gyrus r/l (110875)
6) MITG: Medial and inferior temporal gyri r/l (43156)
7) TL: Anterior temporal lobe medial/lateral part r/l, Superior temporal gyrus central part r/l, Medial and inferior temporal gyri r/l, Lateral occipitotemporal gyrus (gyrus fusiformis) r/l, Posterior temporal lobe r/l, Posterior temporal lobe r/l (226908)
8) HC + PHG (21061)
9) TL + HC + PHG (245847)

C. Classification

For classification, we used features based on voxel-based morphometry. The features were the GM probabilistic segmentations in the template space that were modulated by the Jacobian determinant of the deformation field. This modulation is performed to take account of compression and expansion [38]. To correct for head size, features were divided by intracranial volume. The features were normalized to zero mean and unit variance.

Classification was performed with a linear SVM classifier using the LibSVM implementation [39]. A high value was assigned to the SVM slack parameter ($C = 10^5$) resulting in a hard-margin SVM classifier.

D. Experimental set-up

We compared seven feature selection methods: 1) Feature selection on the SVM feature weights (*W-map*), 2) Recursive feature elimination on the SVM feature weights (*RFE W-map*), 3) P-map feature selection (*P-map*), 4) Recursive feature elimination on the P-map (*RFE P-map*), 5) Univariate t-test for each voxel (*T-test*), 6) Randomized t-test for each voxel (*T-map*), and 7) ROI selection based on expert knowledge (*ROI*). In each cross-validation run, features were selected based on the training set. Using the selected features, an SVM was trained on the training set and applied to the test set.

The feature selection methods were evaluated at a set of fixed numbers of features to be selected. This set started from the total number of features within the GM mask, which was then iteratively divided by two, resulting in the following set: $N \in \{1406418, 803209, 351605, 87902, 43951, 21976, 10988, 5494, 2747, 1374, 687, 344\}$. To allow the hard-margin classifier to find a solution, the number of selected features was not decreased below $N = 344$ keeping the number features higher than or roughly equal to the number of samples. For *RFE W-map* and *RFE P-map*, which are iterative approaches,

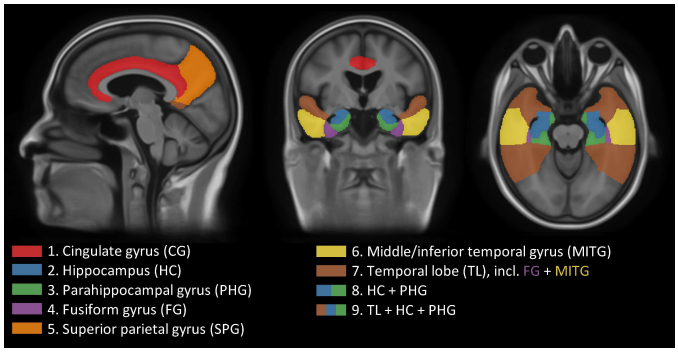


Fig. 1: ROIs for feature selection based on previous knowledge, adapted from [9].

the number of features to be eliminated in every iteration also decreased logarithmically in 16 steps between the points of N.

Classification experiments were performed in four settings: 1) AD-CN, 2) AD-MCI, 3) MCI-CN, and 4) MCIc-MCInc. For each setting, classification performance was quantified by the area under the receiver-operating-characteristic (ROC) curve (AUC) and accuracy with two-fold cross-validation. The cross-validation was iterated 100 times with random splits of the participants into a training and test set of the same size while preserving class priors.

We tested differences in AUC between classifiers with a paired t-test using the 100 iterations as samples. The consistency of the selected features was analyzed using heat maps showing the frequency of the selected features over the cross-validations. We visually inspected the heat maps for $N=43951$ on the axial slices for all methods simultaneously, paying specific attention to clusters of voxels that were selected more than 100 times. Computation times for the feature selection methods were measured in ten iterations of the AD-CN classification with $N=43951$.

IV. Results

A. Classification performance

Fig. 2 shows the AUC for each feature selection method for different numbers of selected features (N). Classification performance was improved by feature selection in all classification settings. For AD-CN classification, the AUC using all features was 90.3% on average over the 100 iterations. This AUC was significantly improved by *W-map* (up to 91.0% selecting 87902 features, $p<0.01$), *RFE W-map* (up to 91.6% selecting 43951 features, $p < 0.01$), *P-map* (up to 91.1% selecting 87902 features, $p<0.01$), *RFE P-map* (up to 92.0% selecting 21976 or 43951 features, $p < 0.01$), and *T-test* (up to 90.4% selecting 351605 features, $p < 0.01$). For AD-MCI classification, the AUC using all features was 68.5% on average. This was only slightly but not significantly improved by *RFE P-map* (up to 68.6% selecting 175803 ($p = 0.84$) or 351605 ($p = 0.88$) features). For MCI-CN classification, the AUC using all features was 72.8%. This was improved only significantly by *RFE P-map* (up to 73.5% selecting 87902 features, $p=0.02$), and slightly but not significantly improved by *RFE W-map* (up to 72.9% selecting 175803 ($p=0.58$) or 351605 ($p=0.69$) features) and *P-map* (up to 73.1% selecting 175803 features, $p = 0.41$). For MCIc-MCInc classification, the AUC using all features was 61.3%. This was slightly improved by *W-map* (up to 61.5% selecting 43951 features, $p=0.37$), *RFE W-map* (up to 61.4% selecting 43951 features, $p=0.49$), and *P-map* (up to 61.4% selecting 175803 features, $p=0.85$). Overall, the largest significant improvement, 1.7% increase in AUC, was achieved for AD-CN selecting 21976 or 43951 features (~1.5% or 3% of the total) with *RFE P-map*.

Feature selection based on the significance map (*P-map*, *RFE P-map*) methods performed slightly better than using methods directly based on the SVM weight vector (*W-map*, *RFE W-map*). This was significant in some cases ($p \leq 0.05$): AD-CN $N=\{21976, 43951\}$, AD-MCI $N \leq 21976$, MCI-CN $N \leq 87902$. In few cases the p-map methods performed significantly worse than the w-map methods: AD-CN $N \leq 5494$ ($p \leq 0.05$) and MCIc-MCI $N=2747$ ($p=0.03$).

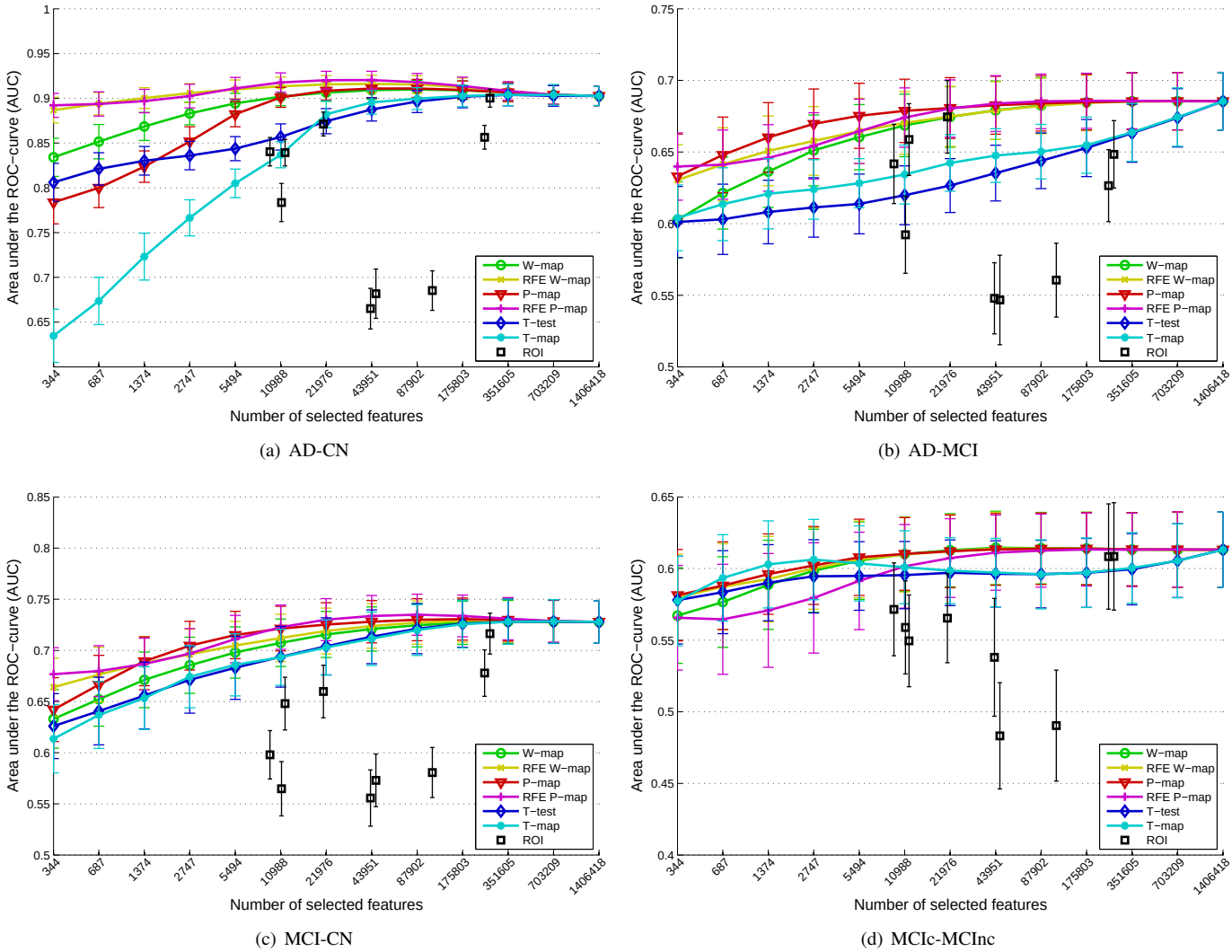
The wrapper methods (*RFE W-map*, *RFE P-map*) yielded generally a higher AUC than the filter methods (*W-map*, *P-map*). Especially when a smaller number of features was selected, the differences between the two approaches became larger. The differences were significant ($p \leq 0.05$) for: AD-CN $N \leq 175803$, AD-MCI $N = \{687, 344\}$, MCI-CN $N \leq 1374$. For MCIc-MCInc $N = \{1374, 2747, 5494\}$, the wrapper methods performed significantly worse than the filter methods ($p \leq 0.05$).

In all settings, the methods based on the SVM weights had a higher performance than those based on t-statistics. The AUC for the SVM weight-based methods was significantly higher in most experiments ($p < 0.01$): AD-CN for $N \leq 351605$, AD-MCI for all N, MCI-CN $N \leq 87902$, and MCIc-MCInc $N \geq 10988$. For MCIc-MCI $N = \{687, 1374\}$, the SVM weight-based methods were significantly worse than the t-statistics methods. The best performing ROI, consisting of the hippocampus, parahippocampal gyrus and the temporal lobe (ROI 9, 266908 features), did not improve AUC in any of the settings. Its AUC was: 90.0% for AD-CN, 64.8% for AD-MCI, 71.6% for MCI-CN, and 60.9% for MCIc-MCInc classification. For all classifications except for MCIc-MCInc, this ROI yielded a significantly lower performance ($p<0.01$) than all SVM-based methods selecting 351605 features.

In addition to the AUC, we analyzed classification accuracy which yielded slightly lower percentages than AUC (Appendix A, Fig. 4). The observed relations within and between the accuracies of the methods were the same as those for AUC.

B. Evaluation of selected features

We evaluated which features were selected by analyzing the heat maps showing the selection frequency of every feature. In cross-validation, a total of 200 feature sets were selected for a given N by every method. Fig. 3 shows the heat maps for the AD-CN classification when 43951 features were selected. Although all methods selected large clusters of voxels in the temporal lobe, the medial temporal lobe in particular, visual inspection of the heat maps for AD-CN showed some differences between the features selected by different methods. The t-statistics methods (*T-test*, *T-map*) selected voxels that



(a) AD-CN (b) AD-MCI

(c) MCI-CN (d) MCIc-MCInc

Fig. 2: Classification area-under-the-ROC-curve (AUC) as function of number of selected features for 7 feature selection methods. The mean and standard deviation of AUC are shown over 100 cross-validations for (a) AD-CN, (b) AD-MCI, (c) MCI-CN, and (d) MCIc-MCInc classification.

were mainly concentrated in the temporal lobe, while the SVM-weight based methods (*W-map*, *P-map*, *RFE W-map*, *RFE P-map*) selected voxels more dispersed over the brain. As mentioned, all methods frequently selected clusters of voxels in the temporal lobe (i.e. hippocampus including amygdala, PHG, FG, MITG, posterior temporal lobe), the insula and the thalamus, but the t-statistics methods did this more frequently and selected larger clusters in these brain regions than the SVM-weight based methods. The heat maps for SVM weight-based methods showed more clusters of frequently selected voxels in the frontal lobe (superior frontal gyrus, precentral gyrus, middle frontal gyrus), postcentral gyrus, and cingulate gyrus than those for the t-statistics methods. We also observed several small differences between the SVM-weight-based methods, of which the most important was that the p-map heat maps showed a more dispersed pattern over the brain than the w-map heat maps. Other differences were that the wrapper methods (*RFE W-map*, *RFE P-map*) selected more clusters of voxels in the superior frontal gyrus than the filter methods (*W-map*, *P-map*), and that the p-map selected more clusters of voxels in the insula than the w-map methods.

Appendix B shows the heat maps for the other classification settings. The patterns in these heat maps were similar to the AD-CN classification, but more dispersed over the brain and less pronounced in certain areas such as the temporal lobe. For most settings, like AD-CN, the voxels selected by the t-statistic methods were mostly concentrated in the temporal lobe, and the voxels selected by the p-map method were more dispersed over the brain. The AD-MCI (Fig. 5) classification was an exception to this, since in this setting the selected voxels were not only for the SVM-weight methods but also for the t-statistics methods more dispersed over the brain. For MCIc-MCInc (Fig. 7), the heat maps for all methods were quite flat with only few voxels that were consistently selected.

As observed in Fig. 3, both the hippocampus and the amygdala were frequently selected for AD-CN classification

by all methods and the t-statistics methods in particular. For AD-MCI and MCI-CN classification (Fig. 6), more amygdala voxels than hippocampus voxels were selected by all methods, while for MCIc-MCInc this was opposite. For MCI-CN, we further noted that the t-statistics methods selected fewer voxels in the insula than in the other settings, but more voxels in the cingulate gyrus and in the rim around the ventricles.

C. Computation times

We measured computation times for the AD-CN classification selecting 43951 features. On a training set of n={149,150}, the average time required for feature selection was; *W-map*: 11.4 (range 10.5-13.9) seconds, *RFE W-map*: 5.5 (5.5-5.6) minutes, *P-map*: 6.7 (6.2-7.6) minutes, *RFE P-map*: 2.0 (1.8-2.4) hours, *T-test*: 18.9 (17.9-20.1) seconds, and *T-map*: 5.6 (5.5-5.6) hours

V. Discussion

In classification experiments of AD, CN and MCI subjects based on structural MRI, we evaluated four feature selection methods that used the SVM weight vector. Two of these methods were novel because they used SVM significance maps as relevance measure for feature selection in a filter and in a wrapper approach. We compared these methods with more commonly used feature selection methods using t-statistics and expert knowledge ROIs.

A. Performance and selected features

In all classification settings (AD-CN, AD-MCI, CN-MCI, and MCIc-MCInc), the evaluated data-driven feature selection methods improved classification performance while the methods based on expert knowledge did not. The performance improvement was the largest using RFE based on the SVM p-map selecting 21976 or 43951 features for AD-CN, which significantly improved the AUC from 90.3% to 92.0%. This selection method also improved the other classifications: AD-MCI 0.1% improvement in AUC (not significant), MCI-CN 0.7%, and MCIc-MCInc 0.1% (not significant). In general, the SVM-weights-based methods performed better than those using t-statistics. Of the SVM-weight-based methods, the ones using the p-map instead of the w-map performed slightly better, while RFE also slightly improved performance.

In this study, we used the same ADNI cohort as used in the comparison study of *Cuingnet et al.* [8]. Their study found an AUC of 95% for AD-CN and 70% for MCIc-MCInc using a voxel-based approach without feature selection (method: *Voxel-Direct-D-gm*), which is somewhat higher than our results using all features. These differences might be attributed to differences in the methodology for template space construction [25]. *Cuingnet et al.* [8] also evaluated two methods that included feature selection and concluded that feature selection only improved performance for the MCIc-MCInc classification.

The evaluated feature selection methods frequently selected clusters of voxels in the hippocampus, amygdala and parahippocampal gyrus. This is in correspondence with the literature, as atrophy of these brain regions is well known to play an important role in AD [40]–[42]. Additionally, atrophy in the cingulate gyri [41]–[43], caudate nucleus [40], [41], insula [40], [41], thalamus [40], [43], superior parietal gyrus (precuneus) [41], [43], temporal gyri [41], [43] and frontal cortex [41] were reported in AD and MCI. The regions in which the data-driven methods frequently selected clusters of features roughly corresponded to these regions, which confirms the validity of these methods. The SVM-weight-based methods found most of these regions, except for the caudate nucleus and the superior parietal gyrus. In addition, the SVM-weight-based methods found a more global effect than the t-statistics methods by selecting regions dispersed over the entire brain.

The finding that classification performances were higher for the SVM-weight-based feature selection methods than for the t-statistics methods could be an indication that the classifier benefits from selecting some voxels that seem to be randomly distributed over the brain. If enough voxels in for example the hippocampus have been selected already, voxels from other brain regions may have complementary information for the classifier and may therefore be more beneficial than other hippocampal voxels that are possibly highly correlated with the hippocampal voxels that were already selected. RFE should be better at selecting complementary features [16], which might explain why the SVM-based RFE methods yielded somewhat higher performances than the filter methods.

Guyon et al. [16] showed that a small change in the feature set could result in a completely different feature ranking by RFE. This possibly causes the selected features for RFE to be even dispersed more over the brain than those for the filter methods. Since the heat map for *RFE P-map* showed that there was a lot of variation in the specific set of selected features, the performance may be improved even more by making the method more robust and less sensitive to small changes in the training set.

A paper by *Chu et al.* [9] found that feature selection only improved classification performance when expert knowledge was used. They compared an *ROI* method with three data-driven methods: *T-test*, *RFE W-map* which removed 3000 voxels in every iteration, and a method using the average absolute t-value in ROIs. In contrast to our work, *Chu et al.* found for AD-CN and MCI-CN classification improvement using some ROIs based on prior knowledge, but no improvement using any of the data-driven methods. The frequency maps shown in [9] for *T-test* and *RFE W-map* show the same pattern as we found in our work. For the *T-test* method the selected voxels were concentrated in the hippocampus and medial temporal lobe, while the *RFE W-map* method showed a more dispersed pattern of selected voxels. Our results suggest that data-driven feature selection methods do have potential to improve classification performance and are worth to be investigated further.

The performance improvements due to feature selection shown in this work could possibly be improved, e.g. by further optimizing the proposed methods to make them more robust or by exploring new methods. Such new methods could include feature reduction or regularization methods, for example one could incorporate principal component analysis [44], [45],

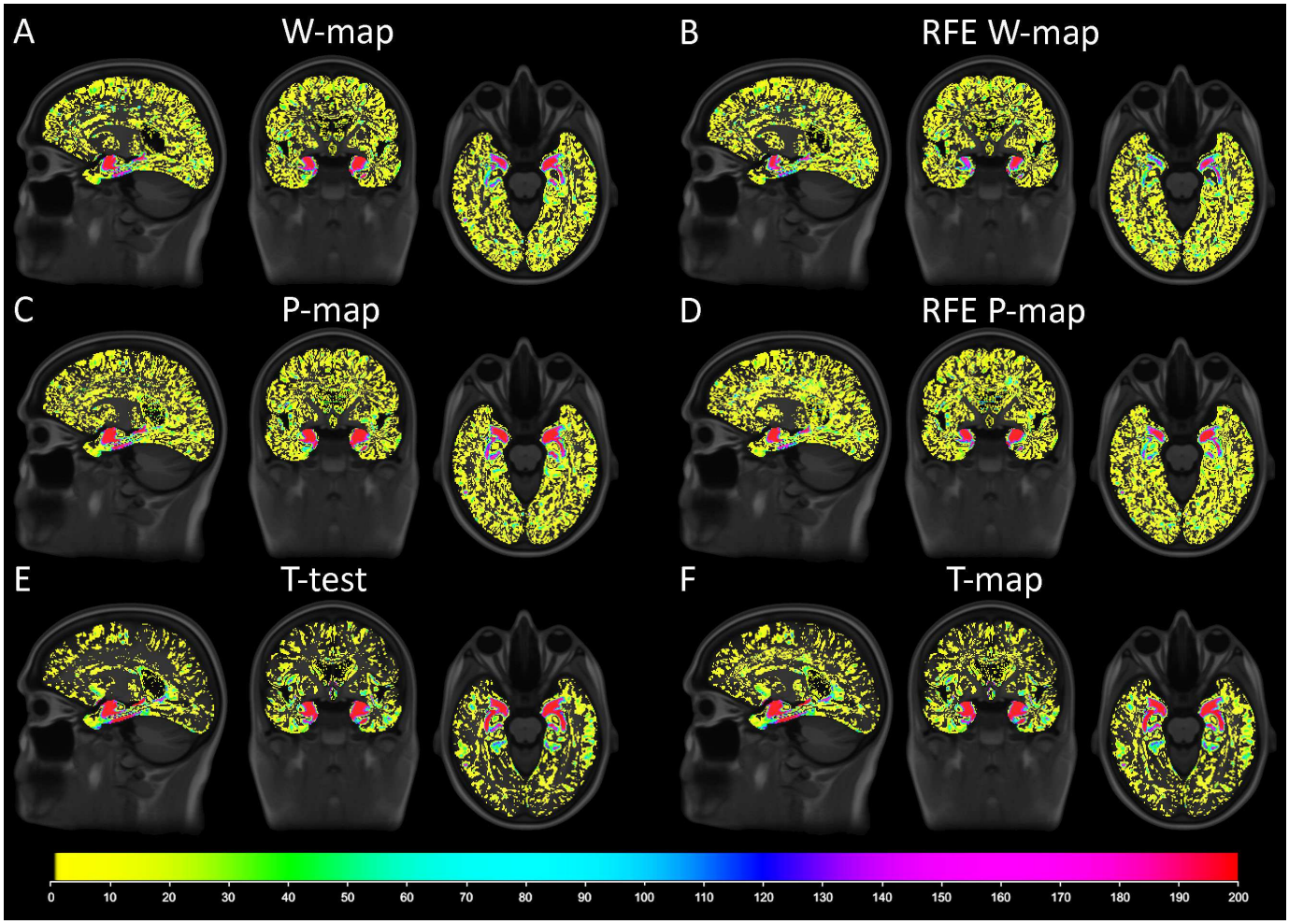


Fig. 3: Heat maps of the selected features for the AD-CN classification by the following methods: A) *W-map*, B) *RFE W-map*, C) *P-map*, D) *RFE P-map*, E) *T-test*, and F) *T-map*. In the 100 iterations of 2-fold cross-validation, a total of 200 sets of features are selected which are shown in the heat maps. The sample point of 43951 selected features is shown.

sparse regression [46], [47] or spatial regularization [48], [49].

B. Computation time

Feature selection increases the time needed for training of the classifier, but saves time in the application of the classifier since it uses fewer features. The *W-map* and the *T-test* methods were the fastest and only took 10-20 seconds. Significance map feature selection is more time-consuming than w-map feature selection and took a couple of minutes instead of seconds. The wrapper approaches are more time-consuming than the filter approaches as they iteratively train a classifier. Of the evaluated methods, the *T-map* method required the most time, up to 6 hours, as it uses permutations.

C. Challenges and limitations

Although four classes (AD, MCIc, MCInc, and CN) are considered in the analysis, we performed all classifications between pairs of classes because of better interpretability of the results.

For the experiments, we used a hard-margin classifier and kept the number of selected features higher than the number of samples. When the number of features is much higher than the number of samples, both soft-margin and hard-margin SVM yield the exact same solution. In that case, the largest Lagrange multiplier of the dual SVM equation is smaller than or equal to the slack parameter C and the misclassification penalty ξ_i does not have an effect. However, when the number of features is smaller, the solutions of hard-margin and soft-margin SVM differ depending on the used value for the C-parameter. For $N = 344$, a $C \approx 1$ or smaller would result in a soft-margin classification. Since *Chu et al.* [9] concluded that the effect of feature selection did not depend on value for the C-parameter, we only evaluated feature selection using hard-margin classification. Since the optimization of the C-parameter is generally performed in a grid-search loop and is therefore computationally expensive, using hard-margin SVM was also a pragmatic approach.

Like most current studies into computer-aided diagnosis of dementia, the reference standard for this study was based on clinical diagnosis. For the ADNI data used in this study, this clinical diagnosis is confirmed by a follow-up period of 18+ months. This may be a limitation, since the clinical diagnosis

[50] might not be always correct. The accuracy of the clinical diagnosis has been reported to be 70-90% compared to the ground truth which was assessed postmortem based on neuropathology [51]–[54]. However, due to the limited availability of data with ground truth diagnosis, we believe that the clinical diagnosis is the best reference standard for current research.

In this work we compared the performance of several feature selection methods for a range of numbers of selected features. For extension of this work, the number of features could be optimized using grid search in cross-validation on the training data.

D. Implications

Although performance improvements were small, some of the evaluated data-driven feature selected methods clearly were better at ranking the features than others. The RFE methods resulted in a better ranking than the filter methods, and the SVM-weight based methods gave a better ranking than the t-statistics methods. From these differences in results between feature selection methods, we learned that data-driven feature selection methods have potential, although we might not have found the ideal method yet. For the choice of the best feature selection methods, one should take into account the trade-off between AUC and complexity. For some applications, a method that requires a much smaller number of features to achieve similar performance might be preferred. Finally, we note that it is important to carefully choose the right method for feature selection as this can significantly reduce or improve the classification performance.

VI. Conclusion

In this work, we showed that data-driven feature selection methods can significantly improve computer-aided diagnosis of dementia. Especially recursive feature elimination on the SVM significance map works well but the performance improvement is still limited. More research and more data with a ground truth diagnosis is needed to further improve these methods for application in clinical diagnosis systems.

Acknowledgment

This work was funded by an Erasmus MC grant on Advanced MR neuroimaging in presenile dementia.

W.J. Niessen and S. Klein acknowledge funding from the European Union Seventh Framework Programme (FP7/2007 2013) under grant agreement no. 601055, VPH-DARE@IT.

Data collection and sharing of the data used in this work 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: Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen Idec Inc.; Bristol-Myers Squibb Company; 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.; Medpace, Inc.; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Synarc Inc.; and Takeda Pharmaceutical Company. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

References

[1] M. Prince, R. Bryce, E. Albanese, A. Wimo, W. Ribeiro, and C. P. Ferri, "The global prevalence of dementia: a systematic review and metaanalysis." *Alzheimers Dement*, vol. 9, no. 1, pp. 63–75.e2, 2013.

[2] Alzheimer's Association, "2011 Alzheimer's disease facts and figures," *Alzheimers Dement*, vol. 7, no. 2, pp. 208–244, 2011.

[3] S. Paquerault, "Battle against Alzheimer's disease: the scope and potential value of magnetic resonance imaging biomarkers," *Acad Radiol*, vol. 19, pp. 509–511, 2012.

[4] M. Prince, R. Bryce, and C. Ferri, *World Alzheimer Report 2011, The benefits of early diagnosis and intervention*. Alzheimer's Disease International, 2011.

[5] S. Klöppel, A. Abdulkadir, C. R. Jack, N. Koutsouleris, J. Mourão Miranda, and P. Vemuri, "Diagnostic neuroimaging across diseases." *Neuroimage*, vol. 61, no. 2, pp. 457–463, 2012.

[6] F. Falahati, E. Westman, and A. Simmons, "Multivariate Data Analysis and Machine Learning in Alzheimer's Disease with a Focus on Structural Magnetic Resonance Imaging." *J Alzheimer Disease*, vol. 41, no. 3, pp. 685–708, 2014.

[7] S. Klöppel, C. M. Stonnington, C. Chu, B. Draganski, I. Scahill, J. D. Rohrer, N. C. Fox, C. R. Jack Jr, J. Ashburner, and R. S. J. Frackowiak, "Automatic classification of MR scans in Alzheimer's disease." *Brain*, vol. 131, no. 3, pp. 681–689, 2008.

[8] R. Cuingnet, E. Gerardin, J. Tessieras, G. Auzias, S. Lehéricy, M.-O. O. Habert, M. Chupin, H. Benali, and O. Colliot, "Automatic classification of patients with Alzheimer's disease from structural MRI: A comparison of ten methods using the ADNI database," *Neuroimage*, vol. 56, no. 2, pp. 766–781, 2011.

[9] C. Chu, A.-L. Hsu, K.-H. Chou, P. Bandettini, and C. Lin, "Does feature selection improve classification accuracy? Impact of sample size and feature selection on classification using anatomical magnetic resonance images." *Neuroimage*, vol. 60, no. 1, pp. 59–70, 2012.

[10] V. Bolón-Canedo, N. Sánchez-Maroño, and A. Alonso-Betanzos, "A review of feature selection methods on synthetic data," *Knowl Inf Syst*, vol. 34, no. 3, pp. 483–519, 2012.

[11] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J Mach Learn Res*, vol. 3, pp. 1157–1182, 2003.

[12] W. Duch, "Filter Methods," in *Feature Extraction - Foundations and Applications*, I. Guyon, M. Nikravesh, S. Gunn, and L. A. Zadeh, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 89–117.

[13] D. Zhang, Y. Wang, L. Zhou, H. Yuan, and D. Shen, "Multimodal classification of Alzheimer's disease and mild cognitive impairment," *Neuroimage*, vol. 5, pp. 856–867, 2011.

[14] D. Salas-Gonzalez, J. M. Gorriz, J. Ramirez, I. A. Illan, M. Lopez, F. Segovia, R. Chaves, P. Padilla, and C. G. Puntonet, "Feature selection using factor analysis for Alzheimers diagnosis using 18F-FDG PET images," *Med Phys*, vol. 37, no. 11, pp. 6084–6084, 2010.

[15] E. Varol, B. Gaonkar, G. Erus, R. Schultz, and C. Davatzikos, "Feature ranking based nested support vector machine ensemble for medical image classification," *Proc IEEE Intl Symp Biomed Imag*, pp. 146–149, 2012.

[16] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene Selection for Cancer Classification using Support Vector Machines," *Mach Learn*, vol. 46, no. 1-3, pp. 389–422, 2002.

[17] J. Mourão Miranda, A. L. W. Bokde, C. Born, H. Hampel, and M. Stetter, "Classifying brain states and determining the discriminating activation patterns: Support Vector Machine on functional MRI data." *Neuroimage*, vol. 28, no. 4, pp. 980–995, 2005.

[18] Z. Wang, A. R. Childress, J. Wang, and J. A. Detre, "Support vector machine learning-based fMRI data group analysis," *Neuroimage*, vol. 36, no. 4, pp. 1139–1151, 2007.

[19] A. Rakotomamonjy, "Variable selection using SVM based criteria," *J Mach Learn Res*, vol. 3, pp. 1357–1370, 2003.

[20] D. Mladenić, J. Brank, M. Grobelnik, and N. Milic-Frayling, "Feature selection using linear classifier weights: interaction with classification models," in *Proc Ann Int ACM SIGIR Conf Research Developm Inform Retriev*, vol. 1, 2004, pp. 234–241.

[21] Y. Liu and Y. F. Zheng, "FS_SFS: A novel feature selection method for support vector machines," *Pattern Recognition*, vol. 39, no. 7, pp. 1333–1345, 2006.

[22] J. Rondina, T. Hahn, L. de Oliveira, A. Marquand, T. Dresler, T. Leitner, A. Fallgatter, J. Shawe-Taylor, and J. Mourao-Miranda, "SCoRS - a method based on stability for feature selection and mapping in neuroimaging." *IEEE Trans Med Imaging*, vol. 33, no. 1, pp. 85–98, 2013.

[23] Y. Fan, D. Shen, R. C. Gur, R. E. Gur, and C. Davatzikos, "COMPARE: classification of morphological patterns using adaptive regional elements," *IEEE Trans Med Imag*, vol. 26, no. 1, pp. 93–105, 2007.

[24] B. Gaonkar and C. Davatzikos, "Analytic estimation of statistical significance maps for support vector machine based multi-variate image analysis and classification." *Neuroimage*, vol. 78, pp. 270–283, 2013.

[25] E. E. Bron, R. M. E. Steketee, G. C. Houston, R. A. Oliver, H. C. Achterberg, M. Loog, J. C. van Swieten, A. Hammers, W. J. Niessen, M. Smits, and S. Klein, "Diagnostic classification of arterial spin labeling and structural MRI in presenile early stage dementia." *Hum Brain Mapp*, vol. 35, no. 9, pp. 4916–4931, 2014.

[26] E. E. Bron, M. Smits, J. C. van Swieten, W. J. Niessen, and S. Klein, "Feature Selection Based on SVM Significance Maps for Classification of Dementia," in *Mach Learn Med Imag*, vol. 8679. Lecture Notes in Computer Science, 2014, pp. 271–278.

[27] V. N. Vapnik, *The nature of statistical learning theory*. Springer-Verlag New York, Inc., 1995.

[28] R. Kohavi and G. John, "Wrappers for feature subset selection," *Artific intellig*, vol. 97, no. 1, pp. 273–324, 1997.

[29] C. R. Jack, M. Bernstein, N. C. Fox, P. Thompson, G. Alexander, D. Harvey, B. Borowski, P. J. Britson, J. L Whitwell, C. Ward, A. M. Dale, J. P. Felmlee, J. L. Gunter, D. L. G. Hill, R. Killiany, N. Schuff, S. Fox-Bosetti, C. Lin, C. Studholme, C. S. DeCarli, G. Krueger, H. Ward, G. J. Metzger, K. T. Scott, R. Mallozzi, D. Blezek, J. Levy, J. P. Debbins, A. S. Fleisher, M. Albert, R. Green, G. Bartzokis, G. Glover, J. Mugler, and M. W. Weiner, "The Alzheimer's Disease Neuroimaging Initiative (ADNI): MRI methods." *J Magn Reson Imaging*, vol. 27, no. 4, pp. 685–691, 2008.

[30] J. Ashburner and K. J. Friston, "Unified segmentation," *Neuroimage*, vol. 26, no. 3, pp. 839–851, Jul. 2005. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/15955494

[31] D. Seghers, E. D'Agostino, F. Maes, D. Vandermeulen, and P. Suetens, "Construction of a brain template from MR images using state-of-the-art registration and segmentation techniques," in *Proc Intl Conf Med Image Comput Comp Ass Intervent*. Springer, 2004, pp. 696–703.

[32] S. Klein, M. Staring, K. Murphy, M. A. Viergever, and J. P. W. Pluim, "Elastix: a toolbox for intensity-based medical image registration," *IEEE Trans Med Imaging*, vol. 29, no. 1, pp. 196–205, 2010.

[33] P. Thévenaz and M. Unser, "Optimization of mutual information for multiresolution image registration," *IEEE Trans Image Proc*, vol. 9, no. 12, pp. 2083–2099, 2000.

[34] S. M. Smith, "Fast robust automated brain extraction," *Hum Brain Map*, vol. 17, no. 3, pp. 143–155, 2002.

[35] A. Hammers, R. Allom, M. J. Koepp, S. L. Free, R. Myers, L. Lemieux, T. N. Mitchell, D. J. Brooks, and J. S. Duncan, "Three-dimensional maximum probability atlas of the human brain, with particular reference to the temporal lobe," *Hum Brain Mapp*, vol. 19, pp. 224–247, 2003.

[36] I. S. Gousias, D. Rueckert, R. A. Heckemann, L. E. Dyet, J. P. Boardman, A. D. Edwards, and A. Hammers, "Automatic segmentation of brain MRIs of 2-year-olds into 83 regions of interest," *Neuroimage*, vol. 40, pp. 672–684, 2008.

[37] R. A. Heckemann, J. V. Hajnal, P. Aljabar, D. Rueckert, and A. Hammers, "Automatic anatomical brain MRI segmentation combining label propagation and decision fusion," *Neuroimage*, vol. 33, pp. 115–126, 2006.

[38] J. Ashburner and K. J. Friston, "Voxel-based morphometry - the methods," *Neuroimage*, vol. 11, pp. 805–821, 2000.

[39] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM TIST*, vol. 2, no. 3, pp. 27–27, 2011.

[40] A. Bastos Leite, P. Scheltens, and F. Barkhof, "Pathological aging of the brain: an overview," *Top Magn Reson Imaging*, vol. 15, no. 6, pp. 369–389, 2004.

[41] G. Frisoni, C. Testa, A. Zorzan, F. Sabattoli, A. Beltramello, H. Soininen, and M. P. Laakso, "Detection of grey matter loss in mild Alzheimer's disease with voxel based morphometry," *J Neurol Neurosurg Psychiatry*, vol. 73, pp. 657–664, 2002.

[42] G. Chételat, B. Desgranges, V. De La Sayette, F. Viader, F. Eustache, and J.-C. Baron, "Mapping gray matter loss with voxel-based morphometry in mild cognitive impairment." *Neuroreport*, vol. 13, no. 15, pp. 1939–1943, 2002.

[43] C. Pennanen, C. Testa, M. P. Laakso, M. Hallikainen, E.-L. Helkala, T. Hänninen, M. Kivipelto, M. Könönen, A. Nissinen, S. Tervo, M. Vanhanen, R. Vanninen, G. B. Frisoni, and H. Soininen, "A voxel based morphometry study on mild cognitive impairment." *J Neurol Neurosurg Psychiatry*, vol. 76, no. 1, pp. 11–14, 2005.

[44] I. Jolliffe, *Encyclopedia of Statistics in Behavioral Science*, B. S. Everitt and D. C. Howell, Eds. Chichester, UK: John Wiley & Sons, Ltd, Oct. 2005.

[45] S. Duchesne, A. Caroli, C. Geroldi, C. Barillot, G. B. Frisoni, and D. L. Collins, "MRI-based automated computer classification of probable AD versus normal controls," *IEEE Trans Med Imag*, vol. 27, no. 4, pp. 509–520, 2008.

[46] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J Royal Statist Soc*, vol. 58, no. 1, pp. 267–288, 1996.

[47] J. Ye, M. Farnum, E. Yang, R. Verbeeck, V. Lobanov, N. Raghavan, G. Novak, A. DiBernardo, and V. a. Narayan, "Sparse learning and stability selection for predicting MCI to AD conversion using baseline ADNI data." *BMC Neurol*, vol. 12, no. 1, pp. 46–46, 2012.

[48] M. R. Sabuncu and K. Van Leemput, "The relevance voxel machine (RVoxM): a self-tuning Bayesian model for informative image-based prediction." *IEEE Trans Med Imaging*, vol. 31, no. 12, pp. 2290–2306, 2012.

[49] R. Cuingnet and M. Chupin, "Spatial and anatomical regularization of SVM for brain image analysis," in *Adv Neur Inf Proc Syst*, vol. 23, 2010, pp. 1–9.

[50] G. McKhann, D. Drachman, M. Folstein, R. Katzman, D. Price, and E. M. Stadlan, "Clinical diagnosis of Alzheimer's disease: report of the NINCDS-ADRDA Work Group under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease." *Neurology*, vol. 34, no. 7, pp. 939–944, 1984.

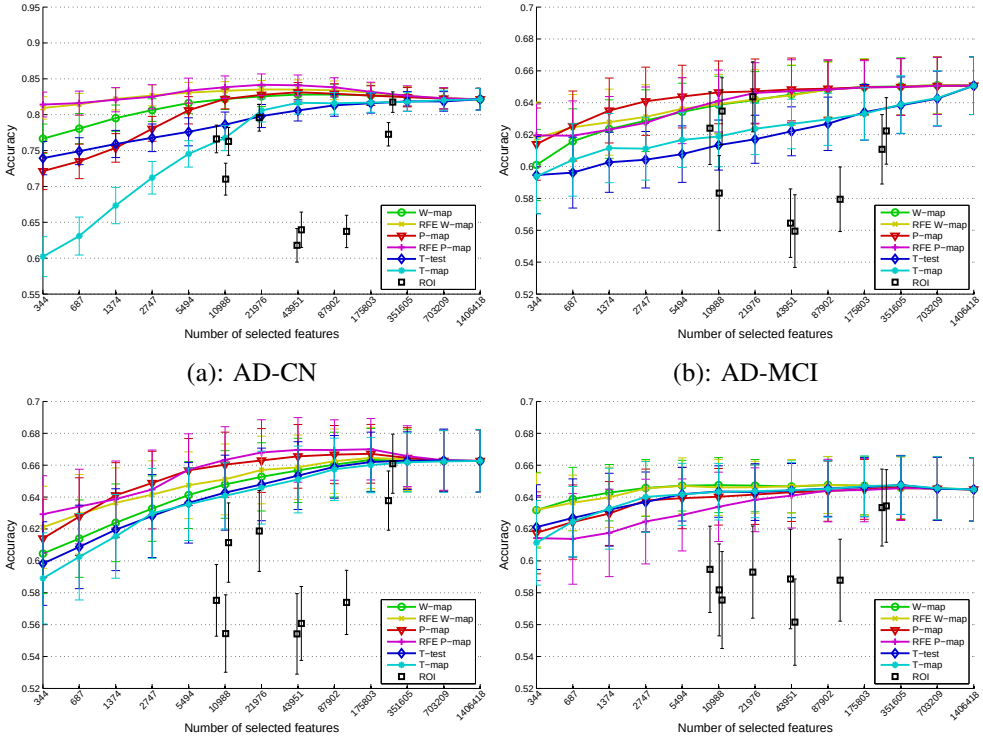
[51] J. Mattila, H. Soininen, J. Koikkalainen, D. Rueckert, R. Wolz, G. Waldemar, and J. Lötjönen, "Optimizing the diagnosis of early Alzheimer's disease in mild cognitive impairment subjects." *J Alzheimer Disease*, vol. 32, no. 4, pp. 969–979, 2012.

[52] A. Lim, D. Tsuang, W. Kukull, D. Nochlin, J. Leverenz, W. McCormick, J. Bowen, L. Teri, J. Thompson, E. R. Peskind, M. Raskind, and E. B. Larson, "Clinico-neuropathological correlation of Alzheimer's disease in a community-based case series." *J Am Geriat Soc*, vol. 47, no. 5, pp. 564–569, 1999.

[53] H. Petrovitch, L. R. White, G. W. Ross, S. C. Steinhorn, C. Y. Li, K. H. Masaki, D. G. Davis, J. Nelson, J. Hardman, J. D. Curb, P. L. Blanchette, L. J. Launer, K. Yano, and W. R. Markesbery, "Accuracy of clinical criteria for AD in the Honolulu-Asia Aging Study, a population-based study." *Neurology*, vol. 57, no. 2, pp. 226–234, 2001.

[54] A. M. Kazee, T. A. Eskin, L. W. Lapham, K. R. Gabriel, K. D. McDaniel, and R. W. Hamill, "Clinicopathologic correlates in Alzheimer disease: assessment of clinical and pathologic diagnostic criteria." *Alzheimer Dis Assoc Disord*, vol. 7, no. 3, pp. 152–164, 1993.

APPENDIX A
CLASSIFICATION ACCURACY



(a): AD-CN (b): AD-MCI

(c): MCI-CN (d): MCIc-MCInc

Fig. 4: Classification accuracy as function of number of selected features for 7 feature selection methods. The mean and standard deviation of accuracy are shown over 100 cross-validations for (a) AD-CN, (b) AD-MCI, (c) MCI-CN, and (d) MCIc-MCInc classification.

APPENDIX B
HEAT MAPS FOR AD-MCI, MCI-CN, AND MCIC-MCINC

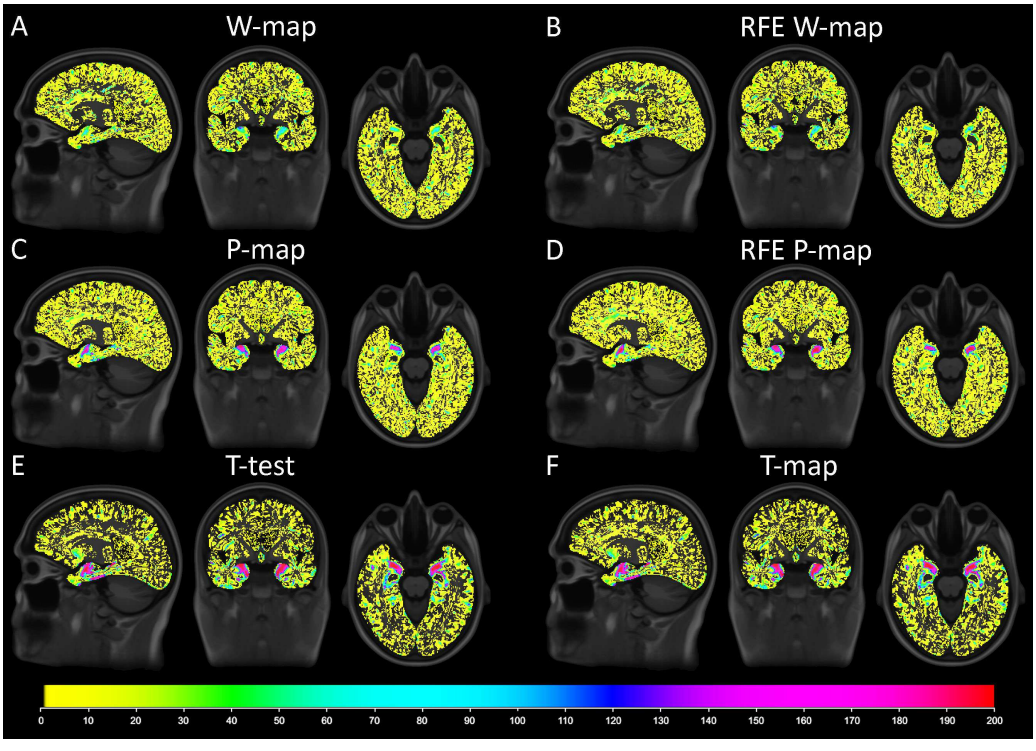


Fig. 5: Heat maps of the selected features for the AD-MCI classification by the following methods: A) *W-map*, B) *RFE W-map*, C) *P-map*, D) *RFE P-map*, E) *T-test*, and F) *T-map*. The sample point of 43951 selected features is shown.

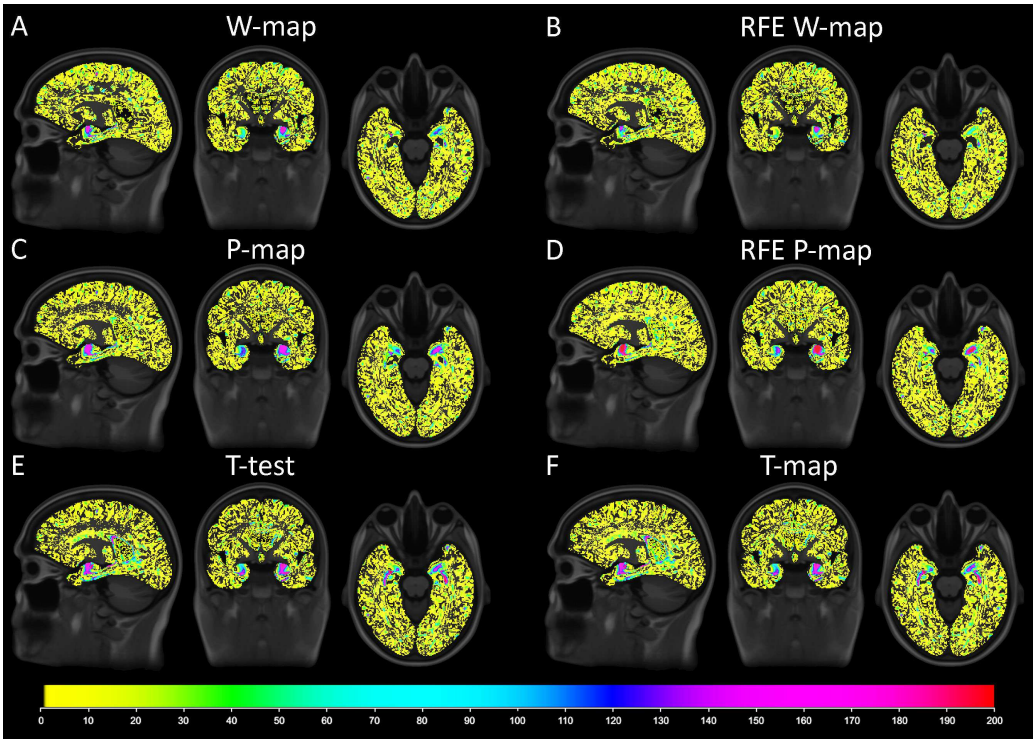


Fig. 6: Heat maps of the selected features for the MCI-CN classification by the following methods: A) *W-map*, B) *RFE W-map*, C) *P-map*, D) *RFE P-map*, E) *T-test*, and F) *T-map*. The sample point of 43951 selected features is shown.

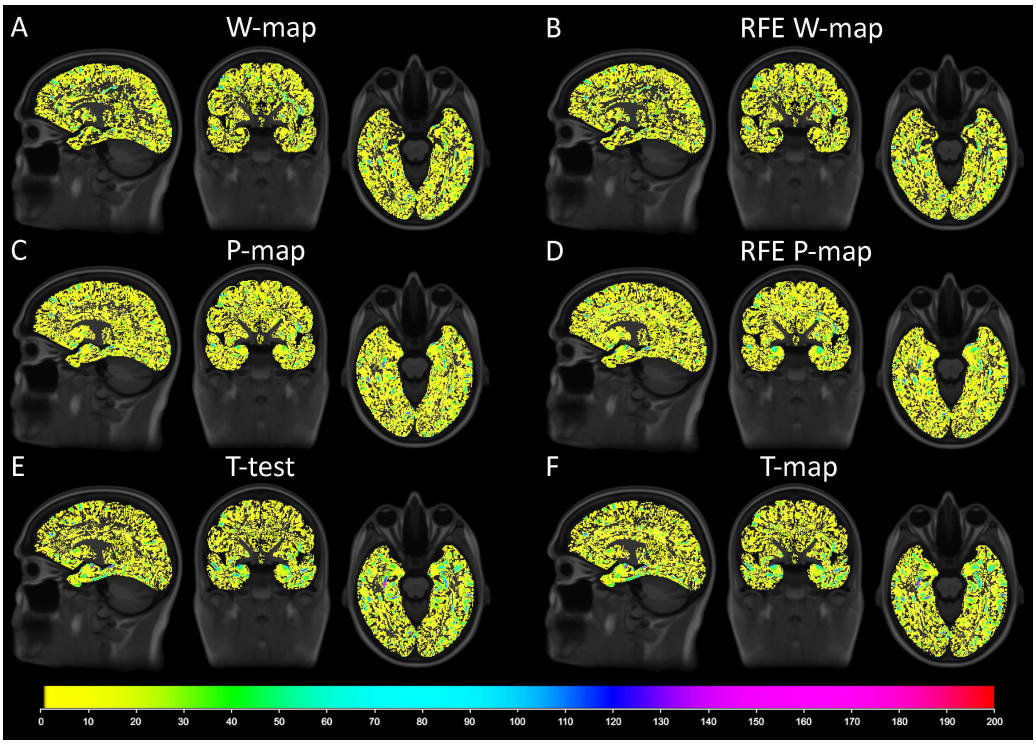


Fig. 7: Heat maps of the selected features for the MCIc-MCInc classification by the following methods: A) *W-map*, B) *RFE W-map*, C) *P-map*, D) *RFE P-map*, E) *T-test*, and F) *T-map*. The sample point of 43951 selected features is shown.