Neuroinform (2016) 14:279–296
DOI 10.1007/s12021-015-9292-3

ORIGINAL ARTICLE

Comparison of Feature Selection Techniques in Machine Learning for Anatomical Brain MRI in Dementia

Jussi Tohka[1,2] · Elaheh Moradi[3] · Heikki Huttunen[3] · Alzheimer's Disease Neuroimaging Initiative

Published online: 23 January 2016

Abstract We present a comparative split-half resampling analysis of various data driven feature selection and classification methods for the whole brain voxel-based classification analysis of anatomical magnetic resonance images. We compared support vector machines (SVMs), with or without filter based feature selection, several embedded feature selection methods and stability selection. While comparisons of the accuracy of various classification methods have been reported previously, the variability of the out-of-training sample classification accuracy and the set of selected features due to independent training and test sets have not been previously addressed in a brain imaging context. We studied two classification problems: 1) Alzheimer's disease (AD) vs. normal control (NC) and 2) mild cognitive impairment (MCI) vs. NC classification. In AD vs. NC classification, the variability in the test accuracy due to the subject sample did not vary between different methods and exceeded the variability due to different classifiers. In MCI vs. NC classification, particularly with a large training set, embedded feature selection methods outperformed SVM-based ones with the difference in the test accuracy exceeding the test accuracy variability due to the subject sample. The filter and embedded methods produced divergent feature patterns for MCI vs. NC classification that suggests the utility of the embedded feature selection for this problem when linked with the good generalization performance. The stability of the feature sets was strongly correlated with the number of features selected, weakly correlated with the stability of classification accuracy, and uncorrelated with the average classification accuracy.

Keywords Magnetic Resonance Imaging · Machine Learning · Feature selection · Alzheimer's Disease · Classification · Multivariate pattern analysis

Alzheimer's Disease Neuroimaging Initiative (ADNI) is a Group/Institutional Author

Data used in preparation of this article were obtained from the Alzheimers 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 adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

Electronic supplementary material The online version of this article (doi:10.1007/s12021-015-9292-3) contains supplementary material, which is available to authorized users.

✉ Jussi Tohka
jtohka@ing.uc3m.es

[1] Department of Bioengineering and Aerospace Engineering, Universidad Carlos III de Madrid, Avd. de la Universidad, 30, 28911, Leganes, Spain

[2] Instituto de Investigación Sanitaria Gregorio Marañon, Madrid, Spain

[3] Department of Signal Processing, Tampere University of Technology, P.O. Box 553, FI-33101 Tampere, Finland

Introduction

Given a training set of brain images and the associated output variables (for example, the diagnosis of the subject), machine learning algorithms try to solve the model that generated the output variables based on the input data (brain images). The idea is that the inferred model predicts accurately and automatically the outputs corresponding to inputs not belonging to the training set. This not only has

direct applications to the design of imaging biomarkers for various brain disorders, but the inferred models can be also analysed as multivariate, discriminative representations of the brain feature of interest. It has been demonstrated that these multivariate representations can provide complementary information to the ordinary massively univariate analysis, both in anatomical and functional imaging (Jimura and Poldrack 2012; Davis et al. 2014; Khundrakpam et al. 2015; Mohr et al. 2015). However, these two analysis techniques and their interpretation differ (Haufe et al. 2014) and they possess distinct advantages and disadvantages (Davis et al. 2014; Kerr et al. 2014).

A fundamental problem in using voxel-based supervised classification algorithms for brain imaging applications is that the dimensionality of data (the number of voxels in the images of a single subject) far exceeds the number of training samples available (subjects whose response variable is known). Rigorous solutions to this problem, termed feature or variable selection, include regularization and subset selection (Hastie et al. 2009). The reasons for using feature selection (FS) are two-fold: 1) using only a selected subset of features tends to improve the classification performance by eliminating the non-informative features, and 2), recognizing only the significant features contributing to the classification can be analysed as a multivariate representation of the brain disorder of interest (Kerr et al. 2014). While comparisons of the accuracy of various classification methods have been reported previously (Cuingnet et al. 2011; Chu et al. 2012; Bron et al. 2015; Sabuncu et al. 2015), the stability of the out-of-training sample classification accuracy and the set of selected features due to independent training and test sets have not been previously addressed in an anatomical brain imaging context. This paper addresses two questions: 1) How do the variability among the subject pool alter the classification accuracy and the selected feature set and 2) do different feature selection and classification techniques differ in their generalization performance?

Data driven FS selection methods are often divided into filter, wrapper and embedded methods (Huttunen et al. 2012; Mwangi et al. 2014). Especially, embedded FS methods have been increasingly applied and developed for brain imaging applications (Grosenick et al. 2008; Ryali et al. 2010; Huttunen et al. 2013a; Casanova et al. 2011b; Khundrakpam et al. 2015). Embedded FS algorithms solve the learning and variable selection problems jointly by optimizing a suitably regularized objective function consisting of a data term and a regularization term whose trade-off is controlled by regularization parameters. Importantly, a regularization term can be designed so that the feature selector possesses the grouping effect (Carroll et al. 2009; Zou and Hastie 2005), forcing simultaneous selection of features that contain correlated information, and takes into account the spatial structure in data inherent to brain imaging (Grosenick et al. 2013; Van Gerven et al. 2010; Michel et al. 2011; Baldassarre et al. 2012; Cuingnet et al. 2013). These brain imaging specific regularizers utilizing the spatial structure in the data often outperform standard regularizers, not taking the spatial structure in data into account, in terms of interpretability of the classifiers (Fiot et al. 2014; Mohr et al. 2015).

The typical logic of the embedded FS is to train a classification model for various values of regularization parameters and then select the best of these classification models, usually using the out-of-the-training-set predictive performance as the selection criterion. Thus, embedded FS can be seen as a two-stage problem, where, in the first stage, one trains a series of classifiers and, in the second stage, selects the best of these classifiers. The research effort in brain imaging community has been strongly focused on the first of these stages and very little effort has been placed on studying the second stage. A particular problem in the second stage is that many feature selection techniques in brain imaging rely on the cross-validation (CV) based estimation of the generalization error to select the regularization parameters. This is problematic because CV-based error estimates with small sample sizes have an extremely large variance. This fact was first demonstrated already by Glick (1978) but it still remains as little known caveat in small sample classification analysis (Dougherty et al. 2010). Stability selection is a relatively new feature selection approach that utilizes the above mentioned variability (Meinshausen and Bühlmann 2010). The key idea is that, using random subsamping of the data, one selects those features that are most frequently selected on the subsamples of data. Although this idea has been applied in neuroimaging applications (Ye et al. 2012; Rondina et al. 2014), its suitability for neuroimaging has received little direct attention.

A closely related question concerns the replicability of the selected voxel sets. More specifically, the question is how much do the error rates and selected features depend on the subject-set studied and to what extent the classifiers represent generalizable discrimination pattern across the classes. In a very interesting study, Rasmussen et al. (2012) demonstrated that within the context of fMRI choosing the regularization parameters relying only on the predictive accuracy has a negative impact on the replicability of the discrimination patterns between the two tasks.

In this paper, we study different linear whole-brain voxel-based classifiers (listed in Table 1) for the Alzheimer's disease (AD) and mild cognitive impairment (MCI) classification based on structural MRI. The studied classification methods include embedded FS methods based on penalized logistic regression, support vector machines with or without filter based FS, and stability selection followed by the SVM

Table 1 Learning algorithms studied in this work. CV and BEE after the abbreviation refer to the criterion used to select λ and possibly α_2

Abbreviation	Algorithm
EN-VA	Logistic regression with elastic-net penalty;variable α_2, $\alpha_1 = 1 - \alpha_2$, $\alpha_3 = 0$
EN-05	Logistic regression with elastic-net penalty with $\alpha_1 = \alpha_2 = 0.5$ fixed, $\alpha_3 = 0$
LASSO	Logistic regression with LASSO penalty $\alpha_1 = 1$, $\alpha_2 = \alpha_3 = 0$
LASSOSTAB	LASSO with stability selection (see Section "Stability Selection").
EN-05STAB	Elastic net with stability selection (see Section "Stability Selection").
GN	GraphNet with $\alpha_1 = 1$, $\alpha_2 = 1$; $\alpha_3 = 1$ for 4 mm data, $\alpha_3 = 10$ for 8 mm data
SVM-Fx	SVM with t-test filter selecting x (125 or 1000) best ranked voxels
SVM-FFDR	SVM with t-test filter selecting voxels surviving a given FDR threshold
SVM-ALL	SVM with all voxels

The regularization parameter ($\lambda\alpha_2$ in our notation) for all SVM algorithms was selected by cross-validation on the training set. The stability selection algorithms were followed by SVM classification

classification. We also contrast non-parametric CV based model selection to a recent parametric classification error estimation based model selection (Huttunen et al. 2013b; Huttunen and Tohka 2015). We proceed with an experimental setup based on split-half resampling similar to the one used in the NPAIRS framework (Strother et al. 2002). The subjects are randomly divided in two non-overlapping sets, test and train, and random divisions are repeated 1000 times. We study both the replicability of the selected variables (voxels) and the error rates of the classifiers. We vary the number of subjects used for training the classifiers and the number of variables.

We chose MRI-based AD/MCI classification applications for several reasons. 1) They are well studied problems that can be solved accurately using linear classifiers (Cuingnet et al. 2011; Bron et al. 2015; Chu et al. 2012). 2) A large enough (at least 200 subjects per class) high quality dataset is available (ADNI) (Weiner et al. 2012) that is a necessity for performing the analysis. We note that this requirement cannot be fulfilled for stable vs. progressive MCI classification with ADNI1 data (Moradi et al. 2015). 3) The uses of supervised machine learning are more varied in functional imaging because of the additional time dimension and more complex experimental designs. We use voxel based morphometry (VBM)-style feature extraction as it has proved effective for this and related applications (Gaser et al. 2013; Moradi et al. 2015; Cuingnet et al. 2011; Bron et al. 2015; Retico et al. 2015), and unlike region of interest (ROI) based methods, provides a feature set that retains the high-dimensional nature of the data and allows to draw conclusions perhaps extendable to other whole brain pattern classification approaches.

We note that computing the results presented in this study required approximately 6 years of CPU time.

Classification and Feature Selection

Linear Classifiers

The image of the subject i is denoted by $\mathbf{x}_i = [x_{i1}, \ldots, x_{iP}]$ where x_{ij} is the gray matter density at the voxel j. Only voxels within the brain mask are considered. The observation matrix is denoted by $\mathbf{X} \in \mathbb{R}^{N \times P}$, whose rows $\mathbf{x}_i$ are the images with corresponding class labels $\mathbf{y} = (y_1, \ldots, y_N)^T$ with $y_i = \{-1, 1\}$. -1 is interpreted as not healthy (AD or MCI) and 1 is interpreted as normal control. The observation matrix is normalized so that $(1/N)\sum_i x_{ij} = 0$ and $(1/N)\sum_i (x_{ij})^2 = 1$. We use N_c to denote the number of training examples from the class c.

The predicted class label $\hat{y}$ for the feature vector $\mathbf{x}$ is given by $\hat{y} = \text{sign}(\beta_0 + \boldsymbol{\beta}^T\mathbf{x}) \doteq g(\mathbf{x})$, where the classifier parameters $\beta_0 \in \mathbb{R}$ and $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_P)^T \in \mathbb{R}^P$ are learned from training data.

Filters for Feature Selection

Filters form the simplest approach to feature selection. Filters work as a pre-processing step for classifiers and are completely independent of the classification, which is often interpreted as their downside (Guyon and Elisseeff 2003). We here consider only a simple t-test based filter (Inza et al. 2004). For each feature j, a t-score is computed

$$t_j = \frac{|\mu_{-1}(j) - \mu_1(j)|}{\sqrt{0.5(\sigma^2_{-1}(j) + \sigma^2_1(j))}}, \quad (1)$$

where $\mu_c(j)$ and $\sigma^2_c(j)$ are mean and variance of the feature j for the class c, respectively, and we have assumed that the classes are balanced. Based on the t-scores t_j, the features

are ranked and the ones with the highest t-scores are selected to be used in classification. We used two different kinds of selection thresholds in this study. We either selected 125 or 1000 highest ranking features or selected these according to a false discovery rate (FDR) corrected threshold (Genovese et al. 2002). This filter method is particularly interesting to this work since it resembles the standard statistical analysis used in VBM.

Embedded Feature Selection

In the embedded FS, the idea is to jointly train the classifier and select the relevant features. This can be formulated as a cost function optimization, where the data term $D(\mathbf{X}, \mathbf{y}, \boldsymbol{\beta}, \beta_0)$ models the likelihood of training data given the classifier parameters and the regularization terms penalize *a* priori unlikely classification parameters. The general form of the cost function used in this paper is Grosenick et al. (2013)

$$C(\boldsymbol{\beta}, \beta_0) = D(\mathbf{X}, \mathbf{y}, \boldsymbol{\beta}, \beta_0) + \lambda \left(\alpha_1 ||\boldsymbol{\beta}||_1 + (\alpha_2/2)||\boldsymbol{\beta}||^2 + \alpha_3 \left(\sum_{i=1}^{P} \frac{1}{2|\mathcal{N}_i|} \sum_{j \in \mathcal{N}_i} (\beta_i - \beta_j)^2 \right) \right), \quad (2)$$

where λ and $\alpha_i, i = 1, 2, 3$ are the parameters that are selected by a model selection criteria and $\mathcal{N}_i$ is the 6-neighborhood of the voxel i. In above, if $\alpha_2 = \alpha_3 = 0$, the sparsity promoting LASSO penalty follows (Tibshirani 1996). If $\alpha_3 = 0$, then elastic-net penalty follows (Zou and Hastie 2005), and if all α_i are allowed to take non-zero values, we talk about GraphNet penalty (Grosenick et al. 2013). If $\alpha_1 = \alpha_3 = 0$, we have a regularizer that does not promote sparsity that is used in the SVM (Hastie et al. 2004). Note that it is possible to adopt a convention that $\sum_j \alpha_j = 1$.

For logistic regression models (Friedman et al. 2010)

$$D(\mathbf{X}, \mathbf{y}, \boldsymbol{\beta}, \beta_0) = (1/N) \sum_{i=1}^{N} Log \Pr(y_i | \mathbf{x}_i)$$

and

$$\Pr(c \mid \mathbf{x}) = \frac{1}{1 + \exp\left[c(\beta_0 + \mathbf{x}^T \boldsymbol{\beta})\right]}.$$

for $c \in \{-1, 1\}$ and for SVM models (Hastie et al. 2004)

$$D(\mathbf{X}, \mathbf{y}, \boldsymbol{\beta}, \beta_0) = \sum_{i=1}^{N} [1 - y_i(\beta_0 + \boldsymbol{\beta}^T \mathbf{x}_i)]_+$$

where $[x]_+ = \max(0, x)$.

Different parameter values (λ, α_j) produce different classifiers and the idea of the embedded FS methods is to train several classifiers with different parameter values and then select the best classifier according to some model selection criteria. Particularly, the product $\lambda\alpha_1$ controls the strength of the L1 regularization effectively deciding how many voxels to select.

Parameter Selection Based on Error Estimation

Cross-validation

K-fold cross-validation is the most widely used technique for the parameter selection in the embedded FS. The training set is divided into K equally sized sets (folds), $K - 1$ of which are used for the classifier training and the remaining one for testing the classifier. This is iterated over the K folds, having a different fold as the test fold during each iteration. Then, the K obtained test accuracies are averaged and the parameter combination giving the highest average accuracy is selected. In this work, we always set $K = 10$ according to Kohavi (1995).

Bayesian Error Estimation

The non-parametric error estimation techniques (such as CV or bootstrap) suffer from excess variability of the error estimates especially in small sample situations (Dougherty et al. 2010). The parametric Bayesian error estimator (BEE) was recently proposed as an alternative to non-parametric error estimation techniques (Dalton and Dougherty 2011) and we have demonstrated that it can be applied to model selection also when its parametric assumptions are only approximately satisfied (Huttunen et al. 2013b; Huttunen and Tohka 2015).

The BEE is defined as the minimum mean squared estimator (MMSE) minimizing the expectation between the error estimate and the true error (Dalton and Dougherty 2011). If we assume Gaussian model for the class-conditional density, a closed form expression can be derived for the posterior expectation of the classification error in the binary classification case under mild assumptions about the covariance structure. The method is attractive, because the errors are estimated directly from the training data, and no iterative resampling or splitting operations are required. This also means substantial savings in the computation time. The closed form equations for BEE are complex and we refer to Dalton and Dougherty (2011), Huttunen and Tohka (2015) for them. The model selector we use is the BEE with the full covariance and the proper prior with the hyper-parameters set exactly as in Huttunen and Tohka (2015). For the completeness, the hyper-parameter values along with a short explanation of their meaning are available in the supplement. The implementation of the BEE model selector is available at https://sites.google.com/site/bayesianerrorestimate/.

Stability Selection

Stability selection is a recently proposed approach by Meinshausen and Bühlmann (2010) for addressing the problem of selecting the proper amount of regularization in embedded FS algorithms. This approach is based on subsampling combined with the FS algorithm. The key idea of this method is that, instead of finding the best value of the regularization and using it, one applies a FS method many times to random subsamples of the data for different value of the regularization parameters and selects those variables that were most frequently selected on the resulting subsamples.

Given a set of regularization parameters Λ, fixed parameters α_i, the number of iterations M, and the threshold value π_{thr}, the stability selection performs following steps:

1) For each regularization parameter $\lambda \in \Lambda$,

- Draw a subsample of training data D_i of size $\lfloor \frac{N}{2} \rfloor$, where N is the number of training data, without replacement.
- Run the regularized logistic regression on D_i using parameter λ (see Eq. 2) and obtain β^i. Keep the selected features $S^\lambda(D_i) = \{j : \beta_j^\lambda \neq 0\}$.
- Repeat the above step M times and compute the selection probability for all features $j = \{1, ..., p\}$,

$$\Pi_j^\lambda = \frac{1}{M} \sum_{i=1}^{M} \mathbf{1}\{j \in S^\lambda(D_i)\}, \tag{3}$$

where the $\mathbf{1}\{.\}$ is the indicator function.

2) Calculate the stability score for each variable $j = \{1, ..., p\}$,

$$S_{stable}(j) = \max_{\lambda \in \Lambda}(\Pi_j^\lambda) \tag{4}$$

3) Finally, select the features with higher stability score than π_{thr}.

In this work, we used $R = 1000$ iterations and the studied regularization parameter values were $\Lambda = \{k \times 0.005; k = 1, 2, ..., 60\}$ for LASSO ($\alpha_1 = 1, \alpha_2 = \alpha_3 = 0$) and $\Lambda = \{k \times 0.01; k = 1, 2, ..., 60\}$ for elastic-net ($\alpha_1 = \alpha_2 = 0.5, \alpha_3 = 0$). The GraphNet penalty was not considered with the stability selection as the computation time would have been prohibitive. The experiments were done with two different threshold values $\pi_{thr} = \{0.1, 0.2\}$, meaning that a feature was selected if at least for one value of $\lambda \in \Lambda$, it was selected 100 ($\pi_{thr} = 0.1$) or 200 ($\pi_{thr} = 0.2$) times among 1000 subsampling experiments. We present the results only for the better threshold value, which was $\pi_{thr} = 0.2$ for 8mm data and $\pi_{thr} = 0.1$ for the 4 mm data. After the stability selection, we still have to select the classifier for classifying the data based on selected features. We decided to use SVM in accordance to Ye et al. (2012).

Materials

ADNI Data

Data used in this work is obtained from the Alzheimers Disease Neuroimaging Initiative (ADNI) database http://adni.loni.usc.edu/. The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimers disease (AD). For up-to-date information, see www.adni-info.org.

We used MRIs from 200 AD subjects, 400 MCI subjects, and 231 normal controls 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 1 mm x 1 mm x 1.2 mm) were available.

Pre-processing

As described by Gaser et al. (2013), Moradi et al. (2015) preprocessing of the T1-weighted images was performed using the SPM8 package (http://www.fil.ion.ucl.ac.uk/spm) and the VBM8 toolbox (http://dbm.neuro.uni-jena.de), running under MATLAB. All T1-weighted images were corrected for bias-field inhomogeneities, then spatially normalized and segmented into grey matter (GM), white matter, and cerebrospinal fluid (CSF) within the same generative model (Ashburner and Friston 2005). The segmentation procedure was further extended by accounting for partial volume effects (Tohka et al. 2004), by applying adaptive maximum a posteriori estimations (Rajapakse et al. 1997), and by using an hidden Markov random field model (Cuadra et al. 2005) as described previously (Gaser 2009). This procedure resulted in maps of tissue fractions of WM and GM. Only the GM images were used in this work. Following the pipeline proposed by Franke et al. (2010), the GM images were processed with affine registration and smoothed with 8-mm full-width-at-half-maximum smoothing kernels.

After smoothing, images were resampled to 4 mm and 8mm isotropic spatial resolution, producing two sets of the images with different resolutions. This procedure generated, for each subject, 29852 or 3747 aligned and smoothed GM density values that were used as MRI features. Image downsampling is often used in machine learning to reduce the number of redundant features in order to improve the classification performance. For example, Franke et al. (2010) concluded that the voxel size had negligible effect on age estimation accuracy. For this study, even more important reason for downsampling is the reduction in computational time and the memory requirements for classifier training.

Normal aging and AD have partially overlapping effects on the brain (Fjell et al. 2013; Dukart et al. 2011), and therefore age effect removal has been suggested to improve the classification performance in the AD related classification tasks (Dukart et al. 2011; Moradi et al. 2015). Briefly, given a set of pre-processed images of normal controls (representing the GM density values), we estimated the effects of normal aging to each voxel separately using linear regression. Then, the learned regression coefficients are used to remove aging effect in any image. The procedure applied is detailed by Moradi et al. (2015), where the rationale behind it is also more thoroughly described. We performed the AD vs. NC experiments for both the images with and without age-removal.

Methods

Experimental Procedure

We performed a split-half resampling type analysis that was introduced by Strother et al. (2002) for their NPAIRS framework and applied by Rasmussen et al. (2012) to study classification analysis of fMRI data. Specifically, we sampled without replacement $N_C = 100$ or $N_C = 50$ subjects from each of the two classes so that $N = 200$ or $N = 100$ and the classification problems were balanced. This procedure was repeated $R = 1000$ times. We denote the two subject samples (split halves; train and test) A_i and B_i for the iteration $i = 1, \ldots, R$ and drop the index where it is not necessary. The sampling was without replacement so that the split-half sets A_i and B_i were always non-overlapping and are considered as independent train and test sets. Each learning algorithm, listed in Table 1, was trained on the split A_i and tested on the split B_i and, vice versa, trained on B_i and tested on A_i. This was done with each image set (4 mm, 8 mm, Age removed 4 mm, Age removed 8 mm for the AD vs. NC problem and age removed 4 mm and age removed 8 mm for the MCI vs. NC problem). Thus, each algorithm was trained and tested 24000 times. All the training operations (estimation of regression coefficients for age removal, parameter selections) were done in the training half. The test half was used only for the evaluation of the algorithms.

We recorded the test accuracy (ACC) of each algorithm (the fraction of the correctly classified subjects in the test half) averaged across $R = 1000$ re-sampling iterations. Moreover, we computed the average absolute difference in ACC between the two split-halves, i.e.,

$$\Delta ACC = \frac{1}{R}\sum_{i=1}^{R} |ACC(A_i, B_i) - ACC(B_i, A_i)|, \qquad (5)$$

where $ACC(A_i, B_i)$ means accuracy when the training set is A_i and the test set is B_i. We additionally recorded the average area under the curve (AUC) for the test subjects. As expected for balanced problems, AUC correlated almost perfectly with ACC and to simplify the exposition of the results, we decided not to present AUCs in the paper.

Statistical testing on ACCs was done to confirm whether the generalization performance of the classifiers differed. Note that just performing the standard t-test or some non-parametric alternative (e.g., a permutation test) on test-accuracies is not correct if we are interested in the true generalization ability to new subjects (not part of the ADNI sample). This is because different replications of the train/test procedure are not independent (Bouckaert and Frank 2004; Nadeau and Bengio 2003). As we performed 1000 replications on different split-halves, we used 1000x2 CV approach known as the corrected repeated 2-fold CV t-test (Bouckaert and Frank 2004). This corrected t-test, which is an improvement of 5X2 CV test of Dietterich (1998) and McNemar's test (see Bouckaert and Frank 2004), relies on the covariance correction of Nadeau and Bengio (2003). The test can be assumed to be conservative in our setting as the correction factor of Nadeau and Bengio (2003) was derived using the assumption that the classifiers are stable with respect to a change in the training set. This is not the case here, and thus the correction overestimates the correlation between the accuracies of different replication rounds. However, we feel that this conservative test is better for the purposes of this work than a liberal uncorrected test, however, for this reason we report the significance at $p = 0.1$ level in addition to the standard $p = 0.05$ level. We used similar correction in the case where an unpaired t-test had to be used, that is, when comparing the ACCs of classifiers trained with a different number of subjects. Finally, where it was appropriate, we combined the test-statistics using a simple average t method (Lazar et al. 2002), which is nearly equivalent to Stouffer's statistic due to the high degrees of freedom.

Hypothesis tests on ΔACC were performed using a permutation test. This assumes the independence of ACC differences between different replications and therefore these tests might be more liberal than the nominal alpha level indicates.

Feature Agreement Measures

We used two measures to quantify the agreement of the selected voxels between two non-overlapping datasets: Dice index and modified Hausdorff distance. The Dice index measures the similarity of two sets (or binarized maps) of selected voxels and is widely used performance measure for evaluating image segmentation algorithms and has been also used to compare fMRI activation maps (Pajula et al. 2012).

The Dice index between the voxel sets V_A and V_B is defined as Dice (1945)

$$DICE(V_A, V_B) = \frac{2|V_A \cap V_B|}{|V_A| + |V_B|} \quad (6)$$

and it varies between 0 (when the two sets do not share any voxels/features) and 1 (when $V_A = V_B$). The Dice index has a close the relationship to Kappa coefficient (Zijdenbos et al. 1994) and we will interpret the Dice values according to well-known but subjective Kappa categorizations (Pajula et al. 2012).

The Dice index does not take into account the spatial closeness of the voxels and returns the value 0 if the data indicates close-by (but not exactly matching) voxels. Also, for this reason, the Dice index might favor dense voxel sets over sparse sets. Therefore, we introduced another similarity measure, modified Hausdorff distance (mHD), which takes into account spatial locations of the voxels (Dubuisson and Jain 1994). Let each of the voxels **a** be denoted by its 3-D coordinates (a_x, a_y, a_z). Then, the mHD is defined as

$$H(V_A, V_B) = \max(d(V_A, V_B), d(V_B, V_A)), \quad (7)$$

where

$$d(V_A, V_B) = \sum_{\mathbf{a} \in V_A} \min_{\mathbf{b} \in V_B} ||\mathbf{a} - \mathbf{b}||.$$

The rationale of using modified Hausdorff distance instead of the (original) Hausdorff distance is that the values of the original Hausdorff distance are large even in the presence of small differences between the voxel sets and typically remains constant when difference increases. The modified Hausdorff distance does not suffer from such a problem; we refer to Dubuisson and Jain (1994) for details. The permutation test was applied for comparison of the feature agreement measures between different algorithms.

Studied Classification Methods and their Implementation

We studied several learning algorithms that are summarized in Table 1. The elastic net and LASSO based methods were implemented with the GLMNET package ((Friedman et al. 2010); http://web.stanford.edu/hastie/glmnet_matlab/) with the default parameters and default grid to search for the optimal λ. The SVMs were implemented with LIbSVM ((Chang and Lin 2011); http://www.csie.ntu.edu.tw/cjlin/libsvm/) and the regularization parameter was always selected based on CV in the training set. The stability selection was based on an in-house Matlab implementation following the guidelines of Ye et al. (2012) and it was followed by SVM classification. For this reason, when referring to Elastic-Net or LASSO later on, we do not typically mean stability selection. The GraphNet was implemented based on an in-house C code implementing the cyclical coordinate descent of Friedman et al. (2010), which uses a quadratic approximation to the log-likelihood, and then coordinate descent on the resulting penalized weighted least-squares problem. However, the coordinate descent is modified to account for the spatial regularizer.[1] For stability selection and GraphNet, we had to fix certain parameter values for computational reasons. For stability selection, these were fixed following suggestions by Ye et al. (2012). For GraphNet, these were fixed using a small-scale pilot study on the AD vs. NC problem with $N_c = 100$. We selected the parameter values for the main experiment so that the numbers of selected features were appropriate , i.e., classification accuracy was not used as the parameter selection criterion but the same data as for the main experiment was used. Note that slightly different parameter values were appropriate for 4 mm and 8 mm data. The studied parameters for the grid search for all the algorithms are provided in the supplement, where full details about parameter tuning experiments can be found. We performed full-scale experiments for the GraphNet with $\alpha_1 = 1, \alpha_2 = 0, \alpha_3 = \{1, 10\}$ called Sparse Laplacian in Baldassarre et al. (2012). However, all the results (ACC, ΔACC, mHD, and Dice) were practically equal to those of GraphNet with parameters as in Table 1, and therefore, they are omitted from the paper.

With SVMs, the filter parameters, the number of features to select (we selected 1000 features for 4 mm voxel size and repeated the experiments selecting 125 as well as 1000 features for 8 mm voxel size) and the FDR thresholds, were selected based on our previous experience on the similar classification problems (Moradi et al. 2015; Moradi et al. 2014). We were unable to find a single FDR-threshold which would have worked well for all settings and chose the values: $q = 0.0005$ for $N_C = 100$ and $q = 0.005$ for $N_C = 50$ in the AD vs. NC classification (the same values were used for both 4 mm and 8 mm data); For the MCI vs. NC problem, when N_C was 100, we used $q = 0.005$ for 4 mm data and $q = 0.05$ for 8 mm data and when $N_C = 50$, we used $q = 0.5$ to prevent empty feature sets that often resulted with normal q thresholds. The rationale for these selections is explained in more detail in the supplement.

Results

Classification Accuracy and its Variability

AD vs. NC

The average ACC and ΔACC for the AD vs. NC problems are listed in Table 2. We discuss only the results with the

[1] This is akin to the implementation in the Donders Machine Learning Toolbox https://github.com/distrep/DMLT

Table 2 The average ACCs and ΔACC for the AD vs. NC experiments. The columns ACC refer to the averages over the $R = 1000$ resamplings

	$N_C = 50$, 4 mm		$N_C = 100$, 4 mm		$N_C = 50$, 8 mm		$N_C = 100$, 8 mm	
	ACC	ΔACC	ACC	ΔACC	ACC	ΔACC	ACC	ΔACC
EN-VACV	0.821	0.041	0.844	0.028	0.823	0.041	0.846	0.027
EN-VABEE	0.815	0.039	0.842	0.027	0.817	0.039	0.841	0.026
EN-05CV	0.820	0.040	0.844	0.027	0.824	0.041	0.846	0.027
EN-05BEE	0.811	0.039	0.837	0.027	0.814	0.039	0.837	0.026
LASSOCV	0.813	0.042	0.840	0.029	0.817	0.041	0.842	0.028
LASSOBEE	0.799	0.043	0.828	0.027	0.801	0.042	0.827	0.027
LASSOSTAB	0.809	0.044	0.829	0.034	0.805	0.047	0.822	0.034
EN-05STAB	0.814	0.041	0.827	0.030	0.813	0.041	0.827	0.032
GNCV	0.822	0.043	0.847	0.029	0.820	0.044	0.838	0.030
GNBEE	0.814	0.039	0.838	0.026	0.807	0.038	0.830	0.026
SVMF-FDR	0.819	0.044	0.841	0.029	0.817	0.049	0.840	0.030
SVMF-1000	0.829	0.043	0.847	0.028	0.809	0.044	0.839	0.031
SVMF-125	–	–	–	–	0.827	0.044	0.846	0.029
SVM-ALL	0.802	0.038	0.830	0.027	0.798	0.040	0.825	0.027
Mean	0.814	0.041	0.838	0.028	0.814	0.042	0.836	0.029

– means that a measure is not available. Slightly different parameter settings are used for GN, SVM-FDR and stability selection depending on the data dimensionality (4 mm voxels vs. 8 mm voxels.)

age removal because it improved the average ACC with all the classifiers. The improvement remained non-significant with respect to generalization performance at the $p = 0.05$ level (corrected t-test) with any of the classifiers, but the combined effect measured using the average t-statistic was highly significant ($p < 10^{-5}$). The improvement in ACC was from 0.004 (GNBEE with 8mm data and $N_c = 100$) to 0.021 (LASSOSTAB with 4mm data and $N_c = 50$) and the average improvement in ACC was 0.014. The classification accuracies without age removal are given in the supplementary Table S1.

The average ACC varied from 0.798 (SVM-ALL, 8 mm, $N_c = 50$) to 0.847 (GN1CV, 4 mm, $N_C = 100$) and showed little dependence on whether 4 mm or 8 mm data was used (mean ACC was 0.838 for 4 mm data and 0.835 for 8 mm data when $N_C = 100$, the difference was not significant in terms of generalization performance, neither with individual classifiers nor when studying average t-statistic). The accuracy was improved by 0.023 (on average) when doubling the number of training subjects. Adding more subjects improved the classification accuracy with all the classifiers, but the improvement remained non significant. However, the average t was again highly significant ($p < 10^{-5}$) suggesting that the addition of subjects was useful as expected.

The average variability of classification accuracies between independent samples ΔACC was greater than the difference between the average classification accuracy between any two classifiers: the smallest ΔACC among independent samples was 0.026 by GraphNet combined with BEE with $N_c = 100$ while the largest difference of the classification accuracy among two different classifiers was 0.025 (between EN-VACV and SVMALL with $N_c = 50$ and 8 mm data). The Fig. 1 illustrates this phenomenon. It shows the scatter plot between the ACC difference of EN-05CV classifier in the two independent splits of the data and the ACC difference between EN-05CV and SVMALL trained with the same data. Even in the case, where the difference between classifiers was maximal (8 mm and $N_c = 50$ red balls in the figure), the ACC differences between the classifiers were about at the same level as the ACC differences due to different train and test sets.

The average ΔACC was reduced by one third (from 0.043 to 0.029 with 4 mm data and 0.042 to 0.028 with 8 mm data) when going from $N_c = 50$ to $N_c = 100$. The reduction was significant with all the classifiers according to the permutation test ($p < 10^{-5}$). There were no striking differences between ΔACC values of different methods; however, ΔACC for the feature selection methods that do not try to estimate classification error (filters and stability selection) was higher on average than for the methods that select features based on the estimate of the classification accuracy (CV and BEE based methods). However, the differences were significant at $p = 0.05$ level only for certain setups, for example, elastic-net based methods with

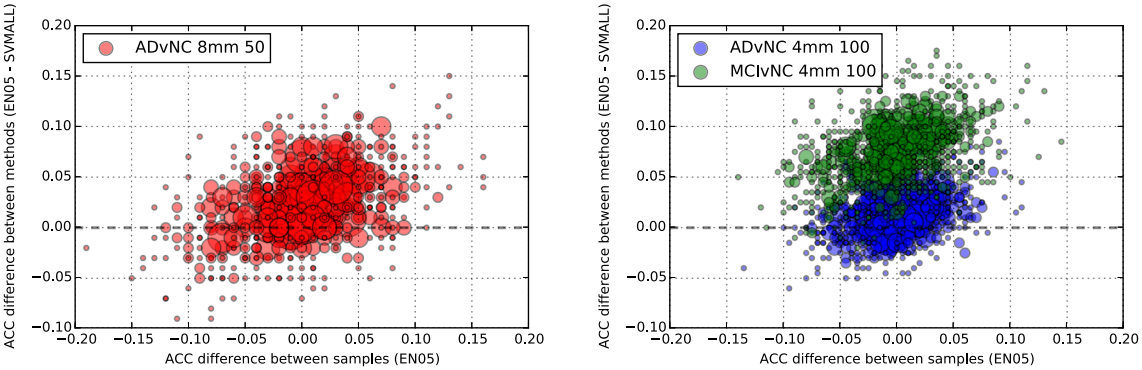

Fig. 1 The ACC difference of EN-05CV classifier in the two independent splits of the data ($ACC_{EN-05CV}(A_i, B_i)$ - $ACC_{EN-05CV}(B_i, A_i)$) plotted against the ACC difference between EN-05CV and SVMALL trained with the same data ($ACC_{EN05-CV}(A_i, B_i)$ - $ACC_{SVMALL}(A_i, B_i)$). The size of the balls correspond to the number of replications with a certain ACC difference. Left panel: For the AD vs. NC problem, the train and test sample had equal or larger influence on ACC than the classifier choice even with the classifiers with the largest difference in average ACC. Right: For the MCI vs. NC problem (green balls), the situation was different than for the AD vs. NC problem (blue balls): the choice of the classifier was important as the green balls are consistently in the positive half of y-axis

$N_c = 50$ and 4 mm data showed significantly smaller ΔACC than the filter based methods (SVMF-1000 and SVMF-FDR).

MCI vs. NC

The classification between MCI and NC subjects can be considered as a much harder problem than the AD vs. NC classification. We did not consider LASSO-based methods or the elastic net with variable α (EN-VA) to simplify the analysis of the results.[2] The results concerning the classification accuracy are presented in Table 3.

The average classification accuracy varied from 0.674 (SVMF-125, $N_C = 50$ 8 mm voxel size) to 0.847 (EN05-CV, $N_C = 100$, 4 mm voxel size). Unlike in the AD vs. NC problem, the choice of the method mattered in this case. GraphNet and Elastic Net were clearly the most accurate methods: With $N_C = 100$ the generalization performance improvement was always significant at $p = 0.05$ level when comparing Elastic-Net or Graphnet method to any SVM-based method with 4 mm data; with 8 mm data, the differences were significant at $p = 0.05$ level against SVMs with filters (SVMF-1000 and SVMF-FDR) and at $p = 0.1$ level against SVMALL and stability selection. This is visible in the scatter plot of the right panel of Fig. 1, where the green balls corresponding to the MCI vs. NC problem lie predominantly in the positive half of the y-coordinate. With the smaller number of subjects $N_C = 50$ and 4 mm data, the performance of Elastic-Net and GraphNet still remained superior, however, the improvement was typically significant only at $p = 0.1$ level. With 8 mm data and $N_C = 50$, the performance differences were not significant except for SVMF-125 which was less accurate than the embedded methods at $p = 0.1$ level. The Elastic Net based stability selection, which used the SVM classifier, performed similarly to the other SVM-based methods and featured poorer classification performance than the standard Elastic Net. The CV and BEE based models for the parameter selection performed similarly in the terms of the average classification performance. Again, and not surprisingly, the addition of subjects improved the performance of all classifiers. With GraphNet and Elastic Net, the average ACC was higher with 4 *mm* data than with 8 *mm* data, however, the improvement was not statistically significant due to high variability between independent samples.

The average variability of the classification accuracy ΔACC was higher (means 0.038 ($N_C = 100$) and 0.052 ($N_C = 50$) for 4 mm data and 0.037 ($N_C = 100$) and 0.051 ($N_C = 50$) for 8 mm data) than with the AD vs. NC problem with the same setups (means 0.041 ($N_C = 50$) and 0.028 ($N_C = 100$) for 4 mm data and 0.029 ($N_C = 50$) and 0.042 ($N_C = 100$) for 8 mm data). Typically, ΔACC did not vary much between the methods. However, with $N_C = 50$, the methods that select the parameters based on CV-error estimate (EN-05CV and GNCV) produced higher ΔACC than the other methods ($p < 0.001$ always). With EN-05CV, ΔACC decreased to the level of other methods when more subjects were added. In contrast, even with $N_C = 100$, ΔACC for GraphNet using the CV-based model selection was higher than ΔACC for other methods. Especially, the ACC difference was large in the iterations i where the differences between MCI classes of A_i and B_i were

[2] Briefly, as the LASSO does not enforce grouping, it is sometimes considered as inappropriate for neuroimaging applications (Carroll et al. 2009). The performance of EN-VA was very similar with EN-05 in the AD vs. NC problem. For these reasons, we decided not to perform the experiments for these methods for MCI vs. NC problem.

Table 3 The average ACCs and ΔACC for MCI vs. NC experiments, see Table 2 for notation

	$N_C = 50$, 4 mm		$N_C = 100$, 4 mm		$N_C = 50$, 8 mm		$N_C = 100$, 8 mm	
	ACC	ΔACC	ACC	ΔACC	ACC	ΔACC	ACC	ΔACC
EN-05CV	0.785	0.058	0.836	0.033	0.739	0.057	0.797	0.038
EN-05BEE	0.782	0.053	0.833	0.032	0.746	0.050	0.800	0.034
GNCV	0.767	0.070	0.810	0.057	0.732	0.059	0.789	0.037
GNBEE	0.775	0.050	0.828	0.031	0.739	0.045	0.794	0.031
EN-05STAB	0.695	0.051	0.753	0.036	0.689	0.049	0.747	0.034
SVMF-FDR	0.700	0.044	0.720	0.043	0.692	0.046	0.710	0.039
SVMF-1000	0.684	0.052	0.719	0.041	0.684	0.054	0.721	0.045
SVMF-125	–	–	–	–	0.674	0.051	0.706	0.040
SVMALL	0.704	0.042	0.758	0.030	0.700	0.045	0.753	0.031
Mean	0.736	0.052	0.782	0.038	0.711	0.051	0.757	0.037

large. For analyzing the differences in the MCI groups, we used the information from the three year follow-up of these patients, specifically the information whether or not they converted to AD within the 3 year time window (see Moradi et al. 2015). We could not find a clear answer to the question why Graphnet with the CV-based model selection was particularly sensitive to differences in MCI classes.

Selected Features

As listed in Table 4, the LASSO methods produced the most sparse voxel set, followed by Elastic-net, and then Graphnet. The filter-based SVMs were designed to give dense voxel sets and it is not particularly informative to analyze the numbers of features selected by the filter methods as the user has a direct control over the sparsity of the classifier.

The elastic net with variable α_2 tended to select more voxels than its fixed α_2 counterpart indicating that model selection strategies favored more dense models. The approach used for the parameter selection in the embedded FS methods had a marked influence on the number voxels selected. The stability selection and CV yielded similar numbers of features whereas the BEE favored more dense models than the other two model selection strategies. For both 4 mm and 8 mm data, the voxel sets were slightly more numerous for the MCI vs. NC problem than for the AD vs. NC problem with the embedded FS methods.

The selection probabilities of the voxels by different methods are illustrated in Fig. 2 through two axial planes passing through hippocampus. For the AD vs. NC problem, the embedded variable selection methods focused on hippocampus and superior temporal cortex and the filter-based

Table 4 Numbers of voxels selected with different classifiers

	$N_C = 50$, 4 mm		$N_C = 100$, 4 mm		$N_C = 50$, 8 mm		$N_C = 100$, 8 mm	
	AD	MCI	AD	MCI	AD	MCI	AD	MCI
EN-VACV	214	–	269	–	121	–	144	–
EN-VABEE	666	–	1002	–	402	–	543	–
EN-05CV	113	109	145	173	72	77	91	131
EN-05BEE	229	225	308	305	142	161	192	225
LASSOCV	32	–	50	–	29	–	44	–
LASSOBEE	57	–	98	–	53	–	91	–
LASSOSTAB	28	–	66	–	17	–	37	–
EN-05STAB	294	250	411	369	103	100	143	159
GNCV	212	225	255	369	358	655	476	1107
GNBEE	814	829	1080	1104	1544	1647	1742	1835
SVMF-FDR	4631	13247	7556	1058	577	1662	942	515

Columns AD refer to the AD vs.NC problem and columns MCI refer to the MCI vs. NC problem. Note that parameters for GN, SVM-FDR, and stability selection were different for 4 mm and 8 mm data and thus the numbers of selected voxels are not comparable between 4 mm and 8 mm data

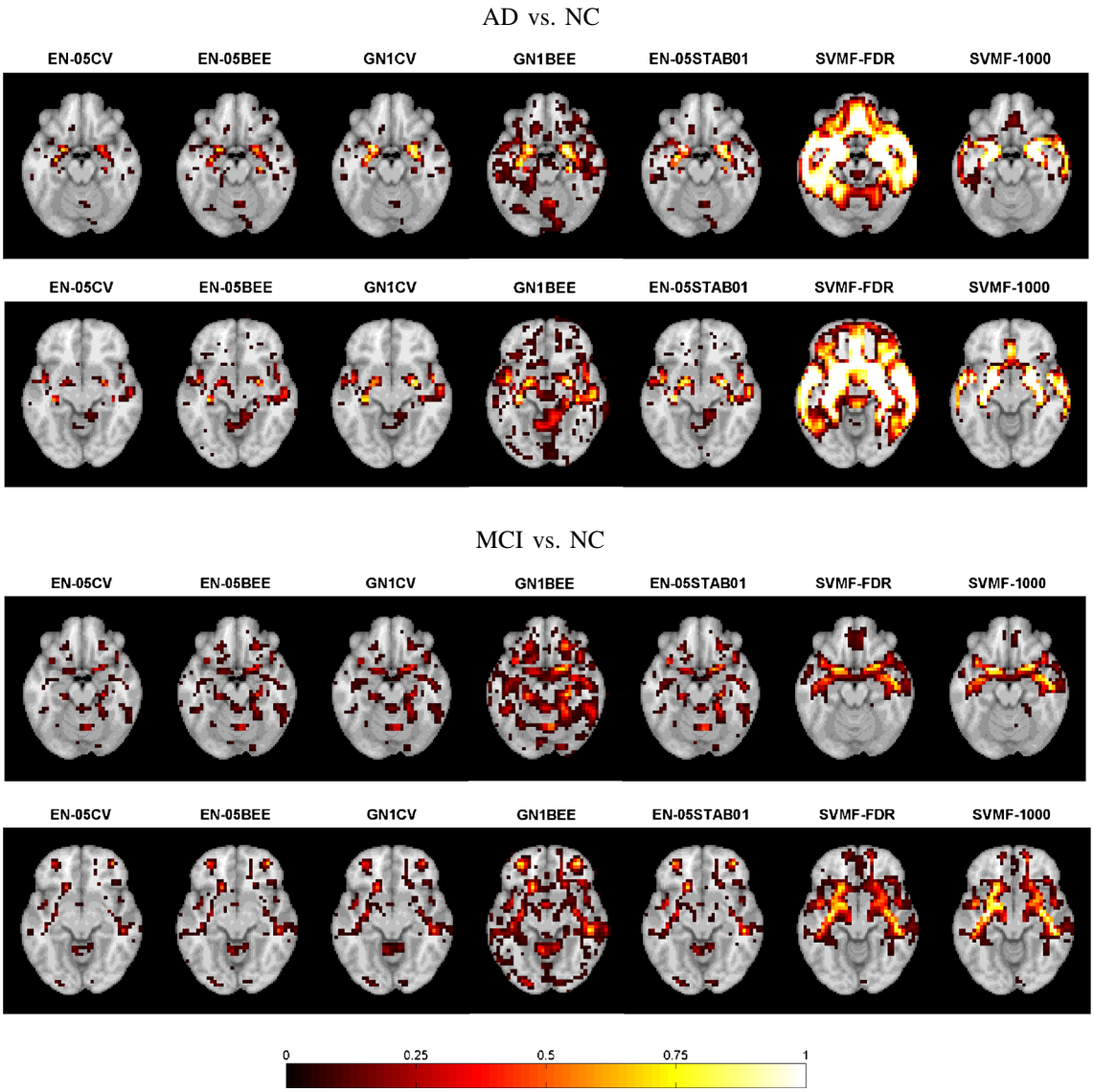

Fig. 2 The probability of voxels being selected for different classification methods over 2000 training replications ($N_c = 100$ and 4 mm data was used). Axial slices at MNI coordinates $z = -18\ mm$ (showing Hippocampus, upper row) and $z = -10\ mm$ (showing Hippocampus and mid-temporal cortices,bottom row) are shown

methods equally included voxels from the middle temporal and frontal cortices. In addition, it can be seen that GNBEE included voxels from cerebellum. All these locations have been implicated to be involved in AD pathology previously (Weiner et al. 2012) and have been found to be effective in classifying between AD patients and normal controls (Casanova et al. 2011b). For the MCI vs. NC problem, the voxel selection probability patterns were somewhat different: for all the methods, the selected voxels concentrated in the frontal regions more than in the AD vs. NC problem. Also, filter and embedded feature selection methods seemingly disagreed which frontal voxels to include - the filters favoring medial frontal gyrus and the embedded methods favoring the middle frontal gyrus.

Stability of Selected Feature Sets

The feature selection stability measured with Dice coefficient (Tables 5 and 6) varied from 0.009 (LASSOBEE, AD vs. NC, 4 mm, $N_c = 50$) to 0.710 (with SVM-F1000, AD vs. NC, 8 mm, $N_c = 100$). The Dice coefficients for the off-the-shelf embedded feature selection methods (LASSO and Elastic-net) were very low. The stability of feature sets was increased by taking the spatial context account (Graphnet) and the most stable feature sets were those based on the fixed number of features to be selected (SVMF-1000). The stability selection increased the Dice coefficients compared to the error estimation based parameter selection - however, typically GraphNet algorithms produced higher Dice

Table 5 The average mHD and Dice values for AD vs. NC experiments

	$N_C = 50$, 4 mm		$N_C = 100$, 4 mm		$N_C = 50$, 8 mm		$N_C = 100$, 8 mm	
	mHD	Dice	mHD	Dice	mHD	Dice	mHD	Dice
EN-VACV	4.431	0.050	3.931	0.063	2.374	0.092	2.113	0.109
EN-VABEE	3.235	0.060	2.499	0.087	1.430	0.146	1.146	0.189
EN-05CV	4.438	0.048	3.951	0.059	2.272	0.101	2.073	0.120
EN-05BEE	3.586	0.041	3.084	0.047	1.763	0.090	1.554	0.101
LASSOCV	6.040	0.014	5.192	0.023	3.060	0.064	2.670	0.072
LASSOBEE	4.908	0.009	4.008	0.015	2.431	0.043	2.003	0.049
LASSOSTAB	5.725	0.041	4.093	0.057	3.010	0.121	2.299	0.155
EN-05STAB	3.000	0.125	2.509	0.164	1.786	0.163	1.557	0.182
GNCV	5.235	0.113	4.318	0.183	2.530	0.201	2.310	0.243
GNBEE	2.643	0.079	2.254	0.093	0.626	0.435	0.557	0.486
SVMF-FDR	1.319	0.440	0.614	0.669	1.011	0.440	0.506	0.668
SVMF-1000	1.648	0.345	1.141	0.490	0.490	0.605	0.343	0.710
SVMF-125	–	–	–	–	1.296	0.332	0.934	0.477
mean	3.851	0.114	3.133	0.163	1.852	0.218	1.543	0.274

The values refer to the averages over the $R = 1000$ resamplings. mHDs are expressed in voxels; the values in millimeters can be obtained by multiplying the mHD in voxels by the voxel size. The standard deviations of mHD and Dice values across 1000 resamplings are presented in the supplement. Other notation is as in Table 2

coefficients than the EN-05STAB. Not surprisingly, the larger the voxel-size and N_c, the higher the Dice coefficient. All the quoted differences in the Dice coefficient value were significant ($p < 10^{-5}$).

While the Dice index values were very low for the off-the shelf embedded methods and also somewhat discouraging for the GraphNet and stability selection methods for 4mm data (indicating 'slight agreement' in the Landis-Koch categorization which is applicable for Dice indeces in addition to Kappa coefficients Pajula et al. 2012), the modified Hausdorff distances showed the feature-selection stability of several embedded methods in a more positive light (see Tables 5 and 6). For problems with $N_c = 100$ and 4 mm data, average mHDs for the embedded methods varied from 5.2 voxels (21 mm, LASSOCV, AD vs. NC) to 2.1 voxels (8.3 mm, GN1BEE) compared to the range between 0.614 and 3.355 voxels (2.5 mm and 14 mm) for the filter methods. mHD values are easy to interpret, a value of 2.25 voxels (GN1BEE, AD vs. NC, 4 mm $N_c = 100$) means that, on average, the maximal distance from voxel selected in one subject sample was 2.25 voxels (10 mm) to a voxel selected in an independent subject sample. The average mHD values for selected methods are visualized in Fig. 3 in millimeters. With the MCI vs. NC problem, the most stable embedded methods featured lower mHD values than SVMF-FDR, which, in the sense of the selection stability, is equivalent

Table 6 The average mHD and Dice values for MCI vs. NC experiments

	$N_C = 50$, 4 mm		$N_C = 100$, 4 mm		$N_C = 50$, 8 mm		$N_C = 100$, 8 mm	
	mHD	Dice	mHD	Dice	mHD	Dice	mHD	Dice
EN-05CV	4.484	0.050	3.423	0.070	2.346	0.072	1.693	0.127
EN-05BEE	3.404	0.046	2.887	0.062	1.641	0.091	1.389	0.135
GNCV	6.433	0.057	4.529	0.076	3.195	0.152	1.293	0.328
GNBEE	2.463	0.077	2.075	0.105	0.578	0.463	0.521	0.516
EN-05STAB	3.093	0.118	2.511	0.146	1.847	0.119	1.435	0.189
SVMF-FDR	0.879	0.501	3.355	0.181	0.708	0.499	1.327	0.300
SVMF-1000	2.345	0.154	1.906	0.255	0.715	0.420	0.612	0.502
SVMF-125	–	–	–	–	1.816	0.146	1.460	0.259
mean	3.300	0.143	2.955	0.128	1.606	0.245	1.216	0.295

The standard deviations of mHD and Dice values across 1000 resamplings are presented in the supplement. See Table 5 for notation

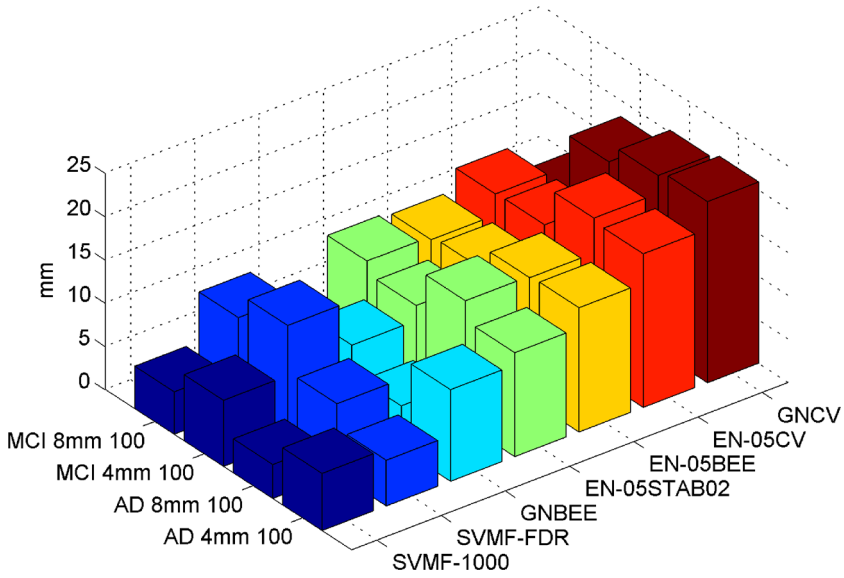

Fig. 3 The average mHDs in millimeters when $N_C = 100$. Note how mHD values were similar in the AD vs. NC and MCI vs. NC problems for the embedded and stability selection methods, but for the filter FS methods, the mHD values were higher for the more difficult MCI vs. NC problem

to the standard massively univariate hypothesis testing with FDR based multiple comparisons correction. In terms of the mHD values, the BEE based parameter selection was more stable than the CV-based parameter selection with any embedded method ($p < 10^{-5}$ always). The variability of the mHD and Dice values of GNCV with 4 mm data was far greater than for other methods. The reason was the same as for the excess variability in the classification accuracy, namely, that GNCV was sensitive to the slight variations in the subject characteristics.

As hypothesized earlier correlation between the average number of voxels selected (noF) and the average Dice coefficient across methods was strong: it varied from 0.67 to 0.98 across the eight conditions (two classification problems, two N_C, and two voxel sizes) and was, as an example, 0.83 for AD vs. NC with 4 mm data and $N_c = 100$. Also, the negative correlation between average NoF and average mHD was strong: it varied from -0.51 to -0.86 across eight conditions (-0.70 for AD vs. NC with 4 mm data and $N_c = 100$). Hence, the dense voxel selection produced more stable feature sets. The average NoF and ΔACC were not found be correlated. The correlation between them averaged across conditions (computed by the z-transform method Kenny 1987) was -0.11 (two-sided $p = 0.46$ according to the test outlined by Kenny (1987)). Thus, it appears that increasing or decreasing the number of voxels selected resulted in no improvement to the variability of classification accuracy.

Not surprisingly, we observed no correlations between the average classification accuracy and either average Dice coefficient or the average mHD. Instead, we observed a significant correlation between average mHD and ΔACC. The correlation averaged by the z-transform method (Kenny 1987) over eight conditions was 0.39 which is significant ($p < 0.005$) according to the test outlined by Kenny (1987). However, the variability in the correlation coefficient was high (from -0.04 to 0.97) between the conditions, with the value 0.97 stemming from the MCI vs. NC problem with 4 mm data and $N_C = 50$, where the embedded methods suffered from the high variability. Also, it needs to be noted that similar correlation was not observed between the average Dice coefficient and ΔACC.

The Fig. 4 shows the probability of the voxel being selected in one split-half but not in the other. The comparison of this Figure to Fig. 2 reveals that the voxels that were probable to be selected were also the most likely to be selected differently between two independent replications.

Discussion

We have presented a comparative analysis of FS methods for whole brain voxel-based classification analysis of structural neuroimaging data. The methods were compared with respect to their classification accuracy and its variation due to independent subject samples as well as the stability of the selected features between different subject samples. We focused on two related and well studied problems: AD vs. NC classification and MCI vs. NC classification with the ADNI data. The compared FS and classification methods included filter-based FS followed by SVM based classification, standard embedded FS methods (LASSO and Elastic-net), stability selection followed by SVM classification, and neuroimaging specific embedded FS (GraphNet). Further, with embedded FS methods, we analyzed two different model selection criteria, non-parametric cross-validation and parametric Bayesian error estimation.

Comparisons of different classification methods on AD related classification tasks have been presented, for

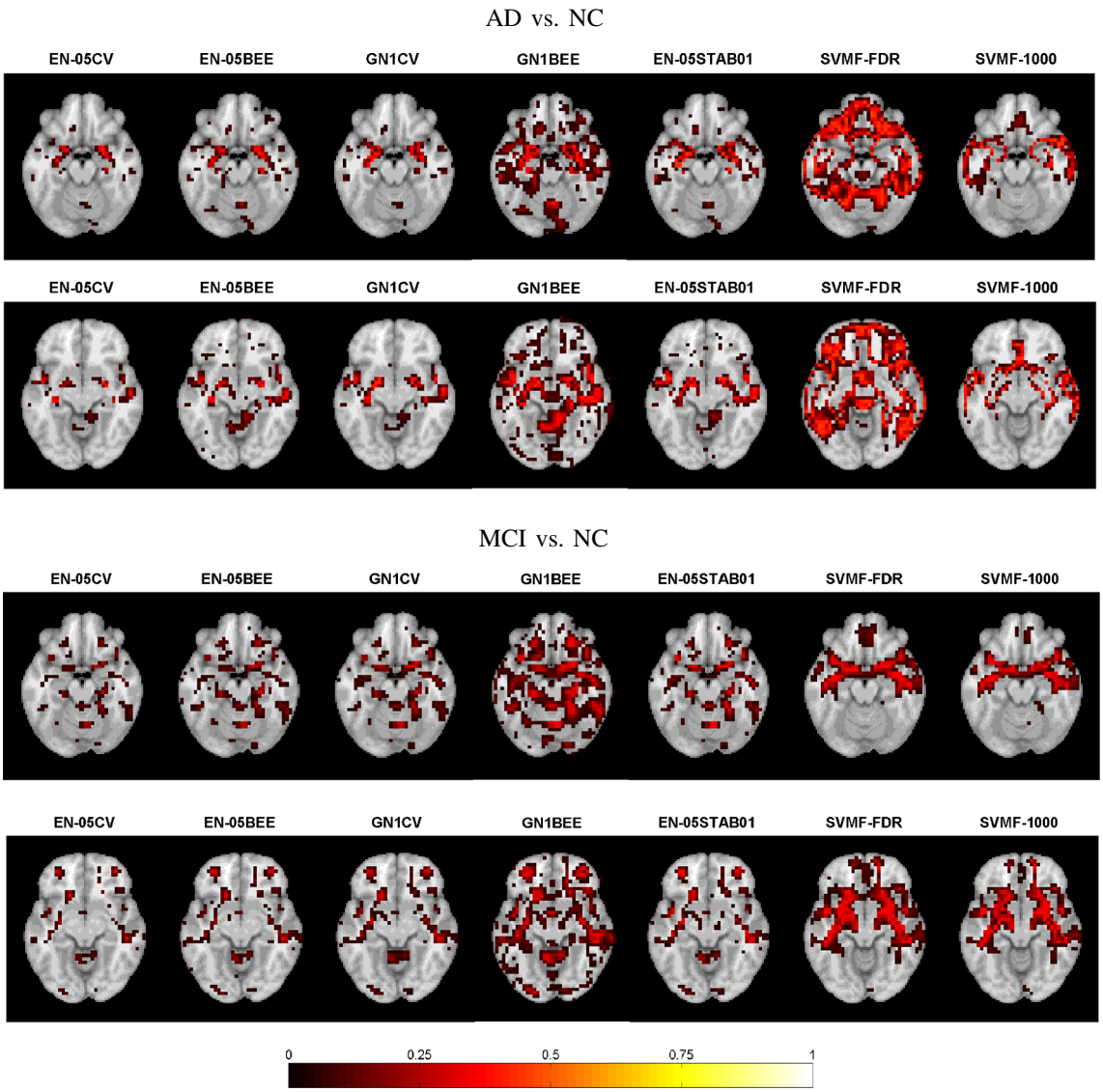

Fig. 4 The probability of voxels for being selected in one split-half while not for the other one over 1000 replications ($N_c = 100$ and 4 mm data was used). Axial slices at MNI coordinates $z = -18$ *mm* (showing Hippocampus, top row) and $z = -10$ *mm* (showing Hippocampus and temporal lobes,bottom row) are shown

instance, by Bron et al. (2015), Cuingnet et al. (2011) and Sabuncu et al. (2015). As these comparative studies have used the classification accuracy, or related quantities, on the whole test sample as the figure of merit, they do not address the questions related the variability of the classifiers with respect to subject sample, which was the focus of this work. Rasmussen et al. (2012) studied the selection of regularization parameters for the embedded FS in fMRI using an NPAIRS framework and concluded that the selection regularization parameters should not be based solely on the classification performance if the interpretation of the resulting classifiers is the final goal. The questions we have addressed are related but different, namely, how do the variability among the subject pool alter the classification accuracy and features set selected and if some feature selection methods are better than others in terms of the generalization performance.

Chu et al. (2012) studied different FS techniques combined with SVMs (filters and recursive feature elimination) on ADNI structural MRI data and concluded that the FS does not have positive influence on the classification accuracy. Our results concerning the classification accuracy match with those of Chu et al. (2012) in the AD vs. NC classification, where the performance of SVM-ALL (which does not use any feature selection) was at the same level as with the classifiers incorporating feature selection. Also, more generally, the variation due to subject sample was more important than the variation due to selected

classification method with the AD vs. NC problem. This is also in line with Chu et al. (2012). On the contrary, embedded FS methods outperformed the SVM based methods with the MCI vs. NC problem, particularly when the training set was large enough, and the performance improvement with a large training set was several times larger than the variability in the classification accuracy due to subject sample. This indicates that data-driven FS can improve the classification accuracy. Note that Chu et al. (2012) did not find FS to be useful for the MCI vs. NC problem. However, this seems to be due to the fact that they studied only filter based FS methods and recursive feature elimination and these do not work as well as the embedded FS methods for this problem according to our results (see also Kerr et al. (2014) for similar conclusions). We did not find significant differences in the classification performance between the imaging specific embedded technique (GraphNet) and a more general embedded technique (Elastic Net). Interestingly, the performance of the stability selection was similar to SVM-ALL, indicating that it did not provide similar gains in classification accuracy as more traditional embedded FS methods.

The variability of the classification accuracy due to subject sample (ΔACC) was almost the same for all methods within the same problem with few exceptions (particularly GraphNet with CV). Not surprisingly, the variability increased with decreasing number of subjects and increasing the problem difficulty (the variability was greater in the MCI vs. NC problem than in the AD vs. NC problem). Instead, the voxel size did not have statistically significant effect on ΔACC. In general, ΔACC measures were a positive surprise, compared to the variability reported in Glick (1978), Dougherty et al. (2010), and although also this work has demonstrated that classification accuracy has a non-zero variance that must be taken into account, the variance was on a tolerable level with the sample sizes studied in this work. The GraphNet with the CV based model selection resulted in higher ΔACC values than the other methods in certain circumstances. This was the problem of model selection as the GraphNet equipped with the parametric BEE model selector did not suffer from the same problem. Otherwise, we did not observe the BEE model selection to differ from the CV based model selection in terms of the classification accuracy or ΔACC. However, as the BEE is several times faster to compute than the CV error estimate (see Huttunen and Tohka 2015), the BEE model selection criterion is attractive for neuroimaging purposes.

The selected feature sets were not particularly stable when the stability was assessed with the Dice index which measures the set similarity without considering the spatial distances between the voxels. Especially, with embedded methods reproducibility of the feature sets as measured with Dice index was poor. The filter based methods produced more stable feature sets. Surprisingly, while the stability selection improved the Dice measure over the traditional model selection methods focusing on the prediction accuracy, the improvement was smaller than expected as the stability selection tries to select models that are maximally stable. However, the stability selection considers each voxel independently that might not be optimal in neuroimaging applications and which may explain rather low Dice values. When accounting for the spatial nature of the data with modified Hausdorff distance (Dubuisson and Jain 1994), the FS stability appeared in a better light. For example, for AD vs. NC problem with 4 mm data and $N_c = 100$, the mHD values varied from 0.614 voxels to 5.192 voxels and for several methods mHD was below 12 mm which can be considered tolerable.

There was a strong linear relation between the sparsity of the classifier and instability of the features, measured either with Dice index or the modified Hausdorff distance. Generally, the more dense the models were the more reproducible they were; this phenomenon has also been noticed in the context of fMRI classification analysis (Rasmussen et al. 2012). Especially this is clearly seen when comparing SVMF-1000 (selecting 1000 features) to SVMF-125 (selecting 125 features). However, selecting more features did not result in less variation in the classification accuracy let alone in a better classification accuracy. Likewise, we did not observe the average classification accuracy and feature stability measures to be correlated. However, we found correlations between ΔACC and the modified Hausdorff distance, which indicates that the feature variability, when quantified with a measure taking spatial nature of the data into account, explained at least some of the variability in the classification accuracy.

Different types of feature selection techniques (filters vs. embedded methods and stability selection) seemingly disagreed on which voxels to select, especially in the MCI vs. NC problem. This is interesting, because filter based methods are (in a sense) equivalent to standard massively univariate analysis, where voxel-wise statistical maps are constructed considering each voxel independently and then thresholded while accounting for multiple comparisons. While the two approaches are different and in many ways complementary, the improved predictive performance of the embedded feature selection methods for the MCI vs. NC problem offers additional evidence that multivariate classification methods could be a useful addition for neuroscientific interpretation, supporting similar conclusions in Jimura and Poldrack (2012), Davis et al. (2014), Khundrakpam et al. (2015), Mohr et al. (2015). In this respect, it is important to bear in mind that machine learning produces so-called backward models and the classifier weights (or selected voxels) have a different meaning than the parameter estimates in the forward models produced by a standard mass-univariate

analysis (Haufe et al. 2014). Especially, truly multivariate feature selection can select features that are not by themselves diagnostic but control for various nuisance factors (Kerr et al. 2014; Haufe et al. 2014).

An application specific finding was that the age-removal procedure (Moradi et al. 2015) improved the classification performance with every classifier. Although the performance improvement did not reach significance according to the corrected repeated t-test, the average t over all the classifiers was significantly different from zero, verifying the findings in the AD vs. NC classification of Dukart et al. (2011) and in the MCI-to-AD conversion prediction of Moradi et al. (2015). The rationale for age-removal stemmed from strong evidence of overlapping effects of normal aging and dementia on brain atrophy (Fjell et al. 2013; Dukart et al. 2011). We note that there was no stratification according to age or gender when dividing the data into two sets A_i and B_i. This was because we wanted reproduce the normal variability between different subject samples: a research group rarely has the possibility to exactly reproduce demographics of the sample acquired by a different research group in a different centre. Obviously, in addition to age, there might be other confounds (such as personal health parameters studied in Franke et al. 2014), whose removal from MRI could improve the classification accuracy and a recent study (Klöppel et al. 2015) jointly removed the effects of age, gender and intracranial volume for the diagnosis of dementia.

An obvious limitation of this study is that we have considered only dementia related applications of machine learning within brain MRI. While we have made a specific effort to avoid using application related information in the classifier design (except for age removal), it is still not clear how well the findings of this study generalize to the studies of other brain diseases. Also, the ADNI study has stringent inclusion/exclusion criteria (Petersen et al. 2010), for example depressed subjects were excluded, and it might be that the variabilities in the classification accuracy reported in this study might underestimate the variabilities in the classification accuracies in more heterogeneous, community based samples.

Conclusions

The question that this work addressed was how much classification accuracy and selected features in machine learning analysis of MRI depend on the subject sample. This question is important as the machine learning analysis is increasingly used in brain imaging and it is essential to know how reliable and reproducible these analyses are. The results in this paper support the use of advanced machine learning techniques in anatomical neuroimaging, but also raise serious concerns related to certain methods and underline the need of care when interpreting the machine learning results. In brief, the main specific findings of this study were: 1) the embedded feature selection methods (GraphNet and Elastic Net) resulted in higher generalization performance than the filter based ones or stability selection in the MCI vs. NC problem; 2) the variability in classification accuracy due to independent samples did not typically depend on the feature selection method and was at an acceptable level; 3) the removal of the age confound improved the classification performance; 4) the feature stability was not correlated with the average classification performance, but a slight correlation with the stability of classification performance was observed.

Information Sharing Statement

The MRI brain image dataset used in this paper was obtained from the Alzheimers Disease Neuroimaging Initiative (ADNI, RRID:nif-0000-00516) which is available at http://www.adni-info.org. The in-house implementations of the GraphNet and stability selection methods are available at https://github.com/jussitohka. The mat files containing the detailed results of the computational analysis will be available at request.

Acknowledgments Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimers Association; Alzheimers Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. 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.

This project has received funding from the Universidad Carlos III de Madrid, the European Unions Seventh Framework Programme for research, technological development and demonstration under grant agreement nr 600371, el Ministerio de Economía y Competitividad (COFUND2013-40258) and Banco Santander.

We also acknowledge CSC – IT Center for Science Ltd., Finland, for the allocation of computational resources.

Compliance with Ethical Standards

Conflict of interests No conflicts of interest exist for any of the named authors in this study.

References

Ashburner, J., & Friston, K. (2005). Unified segmentation. *Neuroimage*, *26*(3), 839–851.

Baldassarre, L., Mourao-Miranda, J., & Pontil, M. (2012). Structured sparsity models for brain decoding from fmri data. In *Pattern Recognition in NeuroImaging (PRNI), 2012 International Workshop on* (pp. 5–8): IEEE.

Bouckaert, R.R., & Frank, E. (2004). Evaluating the replicability of significance tests for comparing learning algorithms. In *Advances in knowledge discovery and data mining* (pp. 3–12): Springer.

Bron, E.E., Smits, M., van der Flier, W.M., Vrenken, H., Barkhof, F., Scheltens, P., Papma, J.M., Steketee, R.M., Orellana, C.M., Meijboom, R., & et al. (2015). Standardized evaluation of algorithms for computer-aided diagnosis of dementia based on structural mri: The caddementia challenge. *NeuroImage*, *111*, 562–579.

Carroll, M.K., Cecchi, G.A., Rish, I., Garg, R., & Rao, A.R. (2009). Prediction and interpretation of distributed neural activity with sparse models. *NeuroImage*, *44*(1), 112–122.

Casanova, R., Whitlow, C.T., Wagner, B., Williamson, J., Shumaker, S.A., Maldjian, J.A., & Espeland, M.A. (2011b). High dimensional classification of structural mri alzheimer's disease data based on large scale regularization. Frontiers in neuroinformatics 5.

Chang, C.C., & Lin, C.J. (2011). Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *2*(3), 27.

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

Cuadra, M.B., Cammoun, L., Butz, T., Cuisenaire, O., & Thiran, J.P. (2005). Comparison and validation of tissue modelization and statistical classification methods in t1-weighted mr brain images. *IEEE Transactions on Medical Imaging*, *24*(12), 1548–1565.

Cuingnet, R., Gerardin, E., Tessieras, J., Auzias, G., Lehéricy, S., Habert, M.O., Chupin, M., Benali, H., & Colliot, O. (2011). Automatic classification of patients with alzheimer's disease from structural mri: a comparison of ten methods using the adni database. *Neuroimage*, *56*(2), 766–781.

Cuingnet, R., Glaunès, J.A., Chupin, M., Benali, H., & Colliot, O. (2013). Spatial and anatomical regularization of svm: a general framework for neuroimaging data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(3), 682–696.

Dalton, L.A., & Dougherty, E.R. (2011). Bayesian minimum mean-square error estimation for classification error—part II: The Bayesian MMSE error estimator for linear classification of Gaussian distributions. *IEEE Trans Signal Process*, *59*(1), 130–144.

Davis, T., LaRocque, K.F., Mumford, J.A., Norman, K.A., Wagner, A.D., & Poldrack, R.A. (2014). What do differences between multi-voxel and univariate analysis mean? how subject-, voxel-, and trial-level variance impact fmri analysis. *NeuroImage*, *97*, 271–283.

Dice, L.R. (1945). Measures of the amount of ecologic association between species. *Ecology*, *26*(3), 297–302.

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

Dougherty, E.R., Sima, C., Hanczar, B., & Braga-Neto, U.M. (2010). Performance of error estimators for classification. *Current Bioinformatics*, *5*(1), 53.

Dubuisson, M.P., & Jain, A.K. (1994). A modified hausdorff distance for object matching. In *Pattern Recognition, 1994. Vol. 1-Conference A: Computer Vision & Image Processing., Proceedings of the 12th IAPR International Conference on*, (Vol. 1 pp. 566–568): IEEE.

Dukart, J., Schroeter, M.L., & Mueller, K. (2011). Age correction in dementia–matching to a healthy brain. *PloS one*, *6*(7), e22–193.

Fiot, J.B., Raguet, H., Risser, L., Cohen, L.D., Fripp, J., & Vialard, F.X. (2014). Longitudinal deformation models, spatial regularizations and learning strategies to quantify alzheimer's disease progression. *NeuroImage: Clinical*, *4*, 718–729.

Fjell, A.M., McEvoy, L., Holland, D., Dale, A.M., Walhovd, K.B., & et al. (2013). Brain changes in older adults at very low risk for alzheimer's disease. *The Journal of Neuroscience*, *33*(19), 8237–8242.

Franke, K., Ziegler, G., Klöppel, S., & Gaser, C. (2010). Estimating the age of healthy subjects from t1-weighted mri scans using kernel methods: Exploring the influence of various parameters. *Neuroimage*, *50*(3), 883–892.

Franke, K., Ristow, M., Gaser, C., Initiative, A.D.N., & et al. (2014). Gender-specific impact of personal health parameters on individual brain aging in cognitively unimpaired elderly subjects. *Frontiers in Aging Neuroscience*, *6*(94).

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*(1), 1–22.

Gaser, C. (2009). Partial volume segmentation with adaptive maximum a posteriori (map) approach. *NeuroImage*, *47*, S121.

Gaser, C., Franke, K., Klöppel, S., Koutsouleris, N., Sauer H, & Initiative, A.D.N. (2013). Brainage in mild cognitive impaired patients: Predicting the conversion to alzheimer's disease. *PloS one*, *8*(6), e67–346.

Genovese, C.R., Lazar, N.A., & Nichols, T. (2002). Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *Neuroimage*, *15*(4), 870–878.

Glick, N. (1978). Additive estimators for probabilities of correct classification. *Pattern Recognition*, *10*(3), 211–222.

Grosenick, L., Greer, S., & Knutson, B. (2008). Interpretable classifiers for fmri improve prediction of purchases. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *16*(6), 539–548.

Grosenick, L., Klingenberg, B., Katovich, K.B.K., & Taylor, J.E. (2013). Interpretable whole-brain prediction analysis with graphnet. *NeuroImage*, *72*, 304–321.

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *The Journal of Machine Learning Research*, *3*, 1157–1182.

Hastie, T., Rosset, S., Tibshirani, R., & Zhu, J. (2004). The entire regularization path for the support vector machine. *The Journal of Machine Learning Research*, *5*, 1391–1415.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning*, 2nd: Springer series in statistics.

Haufe, S., Meinecke, F., Görgen, K., Dähne, S., Haynes, J.D., Blankertz, B., & Bießmann, F. (2014). On the interpretation of weight vectors of linear models in multivariate neuroimaging. *Neuroimage*, *87*, 96–110.

Huttunen, H., & Tohka, J. (2015). Model selection for linear classifiers using bayesian error estimation. *Pattern Recognition*, *48*, 3739–3748.

Huttunen, H., Manninen, T., & Tohka, J. (2012). *Mind reading with multinomial logistic regression: Strategies for feature selection*, (pp. 42–49). Helsinki, Finland: Federated Computer Science Event.

Huttunen, H., Manninen, T., Kauppi, J.P., & Tohka, J. (2013a). Mind reading with regularized multinomial logistic regression. *Machine Vision and Applications*, *24*(6), 1311–1325.

Huttunen, H., Manninen, T., & Tohka, J. (2013b). Bayesian error estimation and model selection in sparse logistic regression. In *2013 IEEE International Workshop on Machine Learning for Signal Processing (MLSP)* (pp. 1–6): IEEE.

Inza, I., Larrañaga, P., Blanco, R., & Cerrolaza, A.J. (2004). Filter versus wrapper gene selection approaches in dna microarray domains. *Artificial Intelligence in Medicine*, *31*(2), 91–103.

Jimura, K., & Poldrack, R.A. (2012). Analyses of regional-average activation and multivoxel pattern information tell complementary stories. *Neuropsychologia*, *50*(4), 544–552.

Kenny, D. (1987). *Statistics for the Social and Behavioral Sciences*: Little Brown.

Kerr, W.T., Douglas, P.K., Anderson, A., & Cohen, M.S. (2014). The utility of data-driven feature selection: Re: Chu et al. 2012. *NeuroImage*, *84*, 1107–1110.

Khundrakpam, B.S., Tohka, J., & Evans, A.C. (2015). Prediction of brain maturity based on cortical thickness at different spatial resolutions. *NeuroImage*, *111*, 350–359.

Klöppel, S., Peter, J., Ludl, A., Pilatus, A., Maier, S., Mader, I., Heimbach, B., Frings, L., Egger, K., Dukart, J., & et al. (2015). Applying automated mr-based diagnostic methods to the memory clinic: A prospective study. *Journal of Alzheimer's Disease*, *47*, 939–954.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *International Joint Conference on Artificial Intelligence (IJCAI95)*, (Vol. 14 pp. 1137–1145).

Lazar, N.A., Luna, B., Sweeney, J.A., & Eddy, W.F. (2002). Combining brains: a survey of methods for statistical pooling of information. *Neuroimage*, *16*(2), 538–550.

Meinshausen, N., & Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *72*(4), 417–473.

Michel, V., Gramfort, A., Varoquaux, G., Eger, E., & Thirion, B. (2011). Total variation regularization for fmri-based prediction of behavior. *IEEE Transactions on Medical Imaging*, *30*(7), 1328–1340.

Mohr, H., Wolfensteller, U., Frimmel, S., & Ruge, H. (2015). Sparse regularization techniques provide novel insights into outcome integration processes. *NeuroImage*, *104*, 163–176.

Moradi, E., Gaser, C., & Tohka, J. (2014). Semi-supervised learning in mci-to-ad conversion prediction - when is unlabeled data useful *IEEE Pattern Recognition in Neuro Imaging*, 121–124.

Moradi, E., Pepe, A., Gaser, C., Huttunen, H., & Tohka, J. (2015). Machine learning framework for early mri-based alzheimer's conversion prediction in mci subjects. *NeuroImage*, *104*, 398–412.

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

Nadeau, C., & Bengio, Y. (2003). Inference for the generalization error. *Machine Learning*, *52*(3), 239–281.

Pajula, J., Kauppi, J.P., & Tohka, J. (2012). Inter-subject correlation in fmri: method validation against stimulus-model based analysis. *PloS one*, *7*(8), e41–196.

Petersen, R., Aisen, P., Beckett, L., Donohue, M., Gamst, A., Harvey, D., Jack, C., Jagust, W., Shaw, L., Toga, A., & et al. (2010). Alzheimer's disease neuroimaging initiative (adni) clinical characterization. *Neurology*, *74*(3), 201–209.

Rajapakse, J.C., Giedd, J.N., & Rapoport (1997). Statistical approach to segmentation of single-channel cerebral mr images. *IEEE Transactions on Medical Imaging*, *16*(2), 176–186.

Rasmussen, P.M., Hansen, L.K., Madsen, K.H., Churchill, N.W., & Strother, S.C. (2012). Model sparsity and brain pattern interpretation of classification models in neuroimaging. *Pattern Recognition*, *45*(6), 2085–2100.

Retico, A., Bosco, P., Cerello, P., Fiorina, E., Chincarini, A., & Fantacci, ME. (2015). *Predictive models based on support vector machines: Whole-brain versus regional analysis of structural mri in the alzheimer's disease*: Journal of Neuroimaging (in press).

Rondina, J.M., Hahn, T., De Oliveira, L., Marquand, A.F., Dresler, T., Leitner, T., Fallgatter, A.J., Shawe-Taylor, J., & Mourao-Miranda, J. (2014). Scors–a method based on stability for feature selection and mapping in neuroimaging. *IEEE Transactions on Medical Imaging*, *33*(1), 85–98.

Ryali, S., Supekar, K., Abrams, D.A., & Menon, V. (2010). Sparse logistic regression for whole-brain classification of fmri data. *NeuroImage*, *51*(2), 752–764.

Sabuncu, M.R., Konukoglu, E., Initiative, A.D.N., & et al. (2015). Clinical prediction from structural brain mri scans: A large-scale empirical study. *Neuroinformatics*, *13*, 31–46.

Strother, S.C., Anderson, J., Hansen, L.K., Kjems, U., Kustra, R., Sidtis, J., Frutiger, S., Muley, S., LaConte, S., & Rottenberg, D. (2002). The quantitative evaluation of functional neuroimaging experiments: the npairs data analysis framework. *NeuroImage*, *15*(4), 747–771.

Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society Series B*, *58*, 267–288.

Tohka, J., Zijdenbos, A., & Evans, A. (2004). Fast and robust parameter estimation for statistical partial volume models in brain mri. *Neuroimage*, *23*(1), 84–97.

Van Gerven, M.A., Cseke, B., De Lange, F.P., & Heskes, T. (2010). Efficient bayesian multivariate fmri analysis using a sparsifying spatio-temporal prior. *NeuroImage*, *50*(1), 150–161.

Weiner, M., Veitch, D.P., Aisen, P.S., Beckett, L.A., Cairns, N.J., & et al. (2012). The alzheimer's disease neuroimaging initiative: A review of paper published since its inception. *Alzheimers & Dementia*, *8*(1), S1–S68.

Ye, J., Farnum, M., Yang, E., Verbeeck, R., Lobanov, V., Raghavan, N., Novak, G., Dibernardo, A., & Narayan, V. (2012). Sparse learning and stability selection for predicting mci to ad conversion using baseline adni data. *BMC Neurology*, *12*(46), 1–12.

Zijdenbos, A.P., Dawant, B.M., Margolin, R.A., & Palmer, A.C. (1994). Morphometric analysis of white matter lesions in mr images: method and validation. *IEEE Transactions on Medical Imaging*, *13*(4), 716–724.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B*, *67*(2), 301–320.