Ensemble support vector machine classification of dementia using structural MRI and Mini-Mental State Examination

Lauge Sørensen[a,b,*], Mads Nielsen[a,b], for the Alzheimer's Disease Neuroimaging Initiative[☆]

[a]*Department of Computer Science, University of Copenhagen, DK-2100 Copenhagen Ø, Denmark*
[b]*Biomediq A/S, DK-2100 Copenhagen Ø, Denmark*

Abstract

Background: The International Challenge for Automated Prediction of MCI from MRI Data offered independent, standardized comparison of machine learning algorithms for multi-class classification of normal control (NC), mild cognitive impairment (MCI), converting MCI (cMCI), and Alzheimer's disease (AD) using brain imaging and general cognition.
New Method: We proposed to use an ensemble of support vector machines (SVMs) that combined bagging without replacement and feature selection. SVM is the most commonly used algorithm in multivariate classification of dementia, and it was therefore valuable to evaluate the potential benefit of ensembling this type of classifier.
Results: The ensemble SVM, using either a linear or a radial basis function (RBF) kernel, achieved multi-class classification accuracies of 55.6% and 55.0% in the challenge test set (60 NC, 60 MCI, 60 cMCI, 60 AD), resulting in a third place in the challenge. Similar feature subset sizes were obtained for both kernels, and the most frequently selected MRI features were the volumes of the two hippocampal subregions left presubiculum and right subiculum. Post-challenge analysis revealed that enforcing a minimum number of selected features and increasing the number of ensemble classifiers improved classification accuracy up to 59.1%
Comparison with Existing Method(s): The ensemble SVM outperformed single SVM classifications consistently in the challenge test set.
Conclusions: Ensemble methods using bagging and feature selection can improve the performance of the commonly applied SVM classifier in dementia classification. This resulted in competitive classification accuracies in the International Challenge for Automated Prediction of MCI from MRI Data.

Keywords: Alzheimer's disease; computer-aided diagnosis; ensemble support vector machine; mild cognitive impairment; mini-mental state examination; structural MRI

1. Introduction

The combination of image analysis and machine learning to construct structural magnetic resonance imaging (MRI) biomarkers of dementia is an active research area (Falahati et al., 2014; Rathore et al., 2017; Arbabshirani et al., 2017). Many different methods have been proposed and evaluated with promising results, however, there is a need for standardized comparisons. Several studies have empirically compared different methods (Cuingnet et al., 2011; Aguilar et al., 2013; Sabuncu et al., 2015) providing some insight as to which MRI features and/or which multivariate methods are beneficial. More recently, challenges in dementia classification have been organized (Simmons et al., 2014; Bron et al., 2015) providing diverse, independent, standardized comparisons. The International Challenge for Automated Prediction of MCI from MRI Data (Sarica et al., 2016), henceforth referred to as "the challenge", offered an opportunity to compare different machine learning methods using precomputed MRI features and mini-mental state examination (MMSE) scores supplied by the challenge organizers. The challenge relied on data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) (Petersen et al., 2010), and a notable characteristic, in comparison with previous challenges, were the multi-class classification of normal control (NC), Alzheimer's disease (AD), mild cognitive impairment that did not convert to AD at follow-up (MCI), and MCI that converted to AD at follow-up (cMCI) as evaluation metric.

This paper presents our algorithm submitted for the challenge. The algorithm used an ensemble of support vector machines (SVMs), i.e., a combination of several differently trained SVMs. An SVM is the most commonly used multivariate method in MRI-based dementia classification (Falahati et al., 2014; Rathore et al., 2017; Arbabshirani et al., 2017), and the classifier has also been widely and successfully applied in studies using data from the ADNI cohort (Weiner et al., 2015). Ensemble clas-

[☆]Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

[*]Corresponding author at: University of Copenhagen, Department of Computer Science, Universitetsparken 1, DK-2100 Copenhagen Ø, Denmark.
E-mail address: lauges@diku.dk (L. Sørensen).

Table 1: Characteristics of the challenge datasets.

	n	Age mean (SD)	Sex % male	MMSE score mean (SD)
Challenge learning set				
NC	60	72.3 (5.7)	50.0	29.1 (1.1)
MCI	60	72.2 (7.5)	46.7	28.3 (1.6)
cMCI	60	73.0 (7.3)	58.3	27.2 (1.9)
AD	60	74.8 (7.4)	48.3	23.4 (2.1)
Challenge test set				
NC	40	74.9 (5.6)	45.0	29.0 (1.1)
MCI	40	72.4 (8.1)	57.5	27.6 (1.9)
cMCI	40	71.7 (6.3)	62.5	27.6 (1.8)
AD	40	73.1 (8.2)	57.5	22.7 (2.0)

sification methods such as the ones that use different subsets of the data, e.g., bagging (Breiman, 1996), or different feature subsets, e.g., the random subspace method (Ho, 1998), may in many cases improve classification performance over a single classifier (Kuncheva, 2014), and ensemble SVMs have previously been successfully applied for dementia classification using different types of MRI measurements and ensemble methods (Shen et al., 2012; Chincarini et al., 2011; Varol et al., 2012; Simpson et al., 2013).

The proposed ensemble method was inspired by the random forest algorithm that uses a combination of bagging and random feature subsets (Breiman, 2001). In particular, we combined bagging without replacement with sequential forward feature selection (SFS) to obtain feature subsets optimal for the SVM classifier. To the best of our knowledge, this is a novel way of constructing the SVM ensemble. Previous feature subset ensemble SVM studies, both within MRI-based dementia classification and within other application areas, were either purely feature subset-based using some form of feature selection or ranking (Chincarini et al., 2011; Varol et al., 2012), random subspace (Waske et al., 2010; Xia et al., 2016), or a combination of selection/ranking and random subspace (Nanni, 2006; Lienemann et al., 2007; Kuncheva et al., 2010; Chen et al., 2014), or combined bagging and feature subsets either using ranking (Shen et al., 2012), random subspace (Tao et al., 2006) or recursive feature elimination based on linear SVM weights (Anaissi et al., 2016). The last is non-trivial to extend to non-linear SVM kernels.

We experimented with using both a linear kernel and a radial basis function (RBF) kernel in the SVMs, and these two configurations were submitted for the challenge. A detailed analyses of the classification results and of the selected feature subsets is presented for the ensembles submitted to the challenge, in addition to a post-challenge analysis of the performance of different feature subset methods and ensemble sizes.

2. Materials and methods

2.1. Data

The challenge used data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (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 MRI, positron emission tomography, other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of MCI and early AD. For up-to-date information, see www.adni-info.org.

The challenge organizers selected a total of 400 subjects from ADNI; 100 NC, 100 MCI, 100 cMCI, and 100 with AD. The subjects were split in a learning set with 240 observations and a test set with 160 observations (Table 1). The subject selection and data set definition procedures are described on the challenge website (Sarica et al., 2016). Information about time to follow-up diagnosis, used to determine MCI or cMCI, was not provided for the challenge data.

2.2. Features

The available features in the challenge consisted of 426 T1-weighted structural MRI measures computed using the cross-sectional pipeline of the FreeSurfer software package (version 5.3) (Fischl and Dale, 2000; Fischl et al., 2002), the age and sex of the subjects, and their baseline MMSE score. The challenge organizers performed all MRI processing and made the resulting MRI measures available to the challenge participants. Among the available MRI measures, we selected 33 brain volumetric measures, 14 hippocampal subregional volumetric measures, 66 regional cortical thickness measures, and the volume of white matter hypointensities. In addition, we computed 10 lobar cortical thickness measures as the mean of the individual regional cortical measures representing each lobe according to the grouping defined by Schmansky et al. (2017). See Table 2 for a detailed specification of the 124 MRI features considered in this study.

The supplied hippocampal subregional volumetric measures and regional cortical thickness measures contained unrealistically large values in some cases. An automatic MRI feature pre-processing step was therefore implemented to bring the order of magnitude to a realistic range (e.g., such that a mean cortical thickness of 2000.0 mm became 2.0 mm). This step was performed prior to the computation of the 10 lobar cortical thickness measures.

FreeSurfer's estimate of the intra-cranial volume (ICV) was also provided among the MRI measurements, and it was included in the feature vector to allow the algorithm to automatically select it if beneficial.

The MMSE score was part of the information used to obtain the clinical diagnosis in ADNI (Petersen et al., 2010) which in turn served as the label in the challenge. We therefore, in addition to the raw MMSE score, made an encoded version using the ADNI thresholds as follows: MMSE $< 24 : 0$ (we know this an AD subject); MMSE ≥ 24 and MMSE $\leq 26 : 1$ (this is a gray zone); MMSE $> 26 : 2$ (we know this is not an AD subject).

The final feature vector was 128-dimensional and consisted of the 124 MRI features, MRI ICV, sex, baseline MMSE score, and encoded baseline MMSE score.

Table 2: Overview of MRI features.

MRI feature category	n	ROI(s)
Brain volumetry	33	l/r accumbens area, l/r amygdala, l/r caudate, l/r cerebellum cortex, l/r choroid plexus, anterior/central/mid anterior/mid posterior/posterior corpus callosum, l/r hippocampus, optic chiasm, l/r pallidum, l/r putamen, l/r thalamus proper, l/r ventral DC, 3rd ventricle, 4th ventricle; l/r inferior lateral ventricle, l/r lateral ventricle, whole brain
Hippocampal sub-regional volumetry	14	l/r CA1, l/r CA2+CA3, l/r CA4+dentate gyrus, l/r fimbria, l/r hippocampal fissure, l/r presubiculum, l/r subiculum
Cortical lobar thickness	10	l/r cingulate cortex, l/r frontal lobe, l/r occipital lobe, l/r parietal lobe, l/r temporal lobe
Cortical regional thickness	66	l/r banks of the superior temporal sulcus, l/r caudal anterior cingulate, l/r caudal middle frontal, l/r cuneus, l/r entorhinal, l/r fusiform, l/r inferior parietal, l/r inferior temporal, l/r isthmus cingulate, l/r lateral occipital, l/r lateral orbitofrontal, l/r lingual, l/r medial orbitofrontal, l/r middle temporal, l/r parahippocampal, l/r paracentral, l/r pars opercularis, l/r pars orbitalis, l/r pars triangularis, l/r pericalcarine, l/r postcentral, l/r posterior cingulate, l/r precentral, l/r precuneus, l/r rostral anterior cingulate, l/r rostral middle frontal, l/r superior frontal, l/r superior parietal, l/r superior temporal, l/r supramarginal, l/r frontal pole, l/r temporal pole, l/r transverse temporal
White matter hypointensities	1	White matter

Abbreviations: l/r; left/right.

2.3. Algorithm

The algorithm consisted of an ensemble of SVMs (Cortes and Vapnik, 1995), each trained using different random subsamples of the Challenge learning set, and each relying on a different subset of the features according to a sequential forward feature selection (SFS) procedure (Jain et al., 2000). The ensemble classification was performed using majority voting among the SVMs. Subject age was accounted for using an age-dependent normalization of the MRI features, and MMSE was always included in the feature set of each individual SVM.

2.3.1. Accounting for age

Age was accounted for using a previously proposed normalization procedure (Sørensen et al., 2017) in which each MRI feature was z-score transformed using a mean and standard deviation that were dependent on the age of the subject. Contrary to Sørensen et al. (2017), we did not perform the normalization per class in order to avoid a four-fold increase of candidate MRI features for feature selection. The normalization parameters were estimated using the entire Challenge learning set, and were subsequently applied to all MRI features in the Challenge learning and Challenge test sets.

2.3.2. Feature selection and SVM training

An SFS procedure with the multi-class (NC vs. MCI vs. cMCI vs. AD) classification accuracy (CA) of an SVM as objective function was used to select the optimal feature subset among the candidate features. The MMSE score was always included in the feature subset because it contributed to the clinical diagnosis of the ADNI subjects (Petersen et al., 2010) and was therefore expected to carry important discriminative information. The SFS procedure was applied by splitting the Challenge learning set stratified by class label into a training and validation set, and, based on the current feature subset, the SVM was trained using the training set while the validation set was used to compute the objective function (i.e., the multi-class CA of SVM). After splitting, both training set and validation set features were z-score normalized using the means $\boldsymbol{\mu}$ and standard deviations $\boldsymbol{\sigma}$ of the training set features to ensure that MRI features and non-MRI features were in the same range. SVM training at each stage in the SFS procedure involved determination of hyper-parameters using grid search in an inner cross-validation loop, followed by training of the SVM using the optimal hyper-parameter values and the entire training set. We denote the final SVM, trained using the optimal feature subset according to the SFS procedure, f. The associated mapping from candidate feature set to optimal feature subset is denoted as s.

2.3.3. Ensemble construction

The individual classifiers in the ensemble were independently trained by repeatedly running the full SFS procedure described in Section 2.3.2 using random permutations of the Challenge learning set. This effectively corresponded to training of each SVM using a random subsample of the Challenge learning set without replacement and an optimal feature subset, and it resulted in a set of trained SVMs $\{f_i\}$, associated mappings to feature subsets $\{s_i\}$, and vectors of means $\{\boldsymbol{\mu}_i\}$ and standard deviations $\{\boldsymbol{\sigma}_i\}$ of the training set features. The ensemble construction procedure is illustrated in Figure 1.

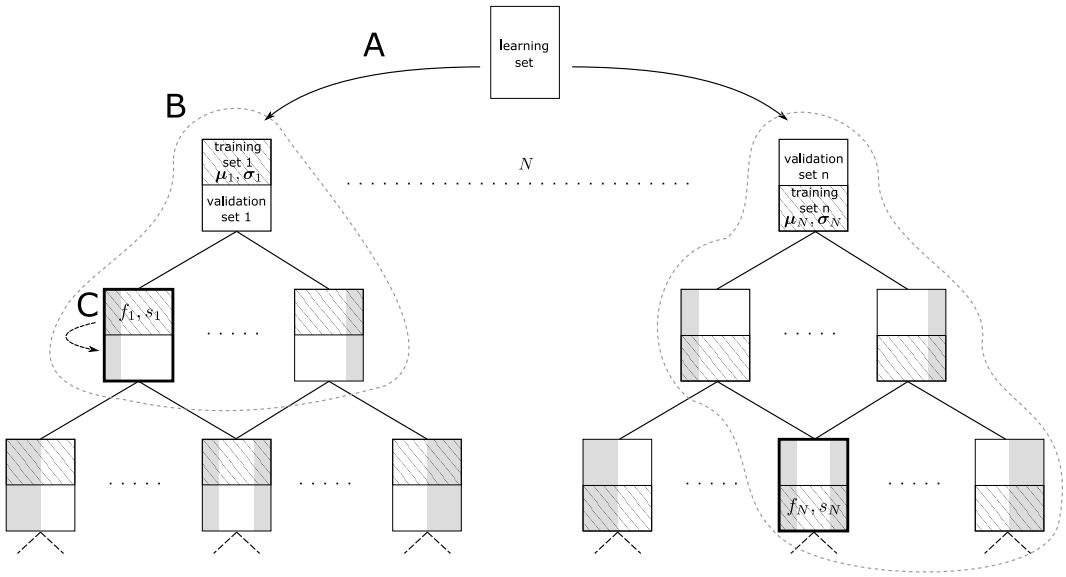

Figure 1: Illustration of the ensemble SVM construction which combined bagging without replacement with SFS for feature subset selection. A: the learning set (rows as observations, columns as features) was split at random in two non-overlapping halves stratified by class label, a training set and a validation set. The training set feature means $\boldsymbol{\mu}$ and standard deviations $\boldsymbol{\sigma}$ were used for z-score normalization of both the training and the validation sets. B: a full SFS procedure was run using the z-score normalized training and validation sets resulting in a mapping s from the full feature set to a feature subset, and an associated trained multi-class SVM f that used that feature subset. C: at each node of the SFS search path in the graph of possible feature combinations, a multi-class SVM was trained using the training set, and the multi-class CA of the trained SVM applied to score the validation set provided the SFS objective function. In the illustrated example, the SFS procedure terminated at the first feature in the first split (the search path is illustrated as a dashed gray line, the data column for the selected feature is gray, and the vertex at which the SFS procedure terminated is marked by a thick black boundary), and it terminated at the combination of the first and last features in the Nth split. The final ensemble consisted of the feature subset mappings $\{s_1, \ldots, s_n\}$ and associated trained SVMs $\{f_1, \ldots, f_n\}$, and the training set feature means $\{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_n\}$ and standard deviations $\{\boldsymbol{\sigma}_1, \ldots, \boldsymbol{\sigma}_n\}$ for normalization.

The combination of data subsets and feature subsets in ensemble SVM construction was inspired by the random forest algorithm (Breiman, 2001). However, the proposed method used bagging without replacement for data subset selection and SFS for optimal feature subset selection. A general empirical study demonstrated that bagging without replacement performed better than with replacement in situations with imbalanced and noisy data (Khoshgoftaar et al., 2011). The data is balanced on this study, however, the classes are noisy due to the use of clinical labels and lack of definite ground truth, and due to the clinical overlap between the classes. The use of SFS for feature subset selection instead of random selection allows for feature subsets that are optimal for the SVM while still providing relatively fast ensemble construction in comparison with using computationally more involved feature selection algorithms.

2.3.4. Ensemble classification

The final classification of a feature vector $\mathbf{x}$ from the Challenge test set was obtained using majority voting among the SVMs in the ensemble

$$f_{\text{ensemble}}(\mathbf{x}) = \underset{\omega_j}{\text{argmax}} \sum_{i=1}^{N} g(f_i(s_i((\mathbf{x} - \boldsymbol{\mu}_i) \oslash \boldsymbol{\sigma}_i)), \omega_j),$$

where N is the size of the ensemble, ω_j is the jth class label, $\oslash$ denotes element-wise division, and $g(\cdot, \cdot)$ is an indicator function defined as

$$g(y, \omega) = \begin{cases} 1 & \text{if } y = \omega \\ 0 & \text{if } y \neq \omega \end{cases},$$

where y is the predicted label.

2.4. Experimental setup

We used the LIBSVM library (version 3.22) (Chang and Lin, 2011) for SVM training and classification. This meant that the multi-class classification was performed using the "one-against-one" approach. The ensemble method, including the SFS procedure, was our own in-house Matlab (R2015b) code. The following parameters and settings were used in the experiments:

- SVM kernel: {linear, RBF}
- SVM hyper-parameter grid:
  - $C \in \{2^{-3}, 2^{-1}, \ldots, 2^{3}\}$
  - $\gamma \in \{2^{-5}, 2^{-3}, \ldots, 2^{5}\}$ (only for the RBF kernel)
- Number of inner cross-validation folds for SVM hyper-parameter search: 3
- Number of SVMs in the ensemble: $N = 25$
- Proportion of observations per class in training / validation set in the SFS procedure: 75% / 25%

The random number generator for Challenge learning set permutation was initialized using the same random seed for each of the two constructed ensembles (corresponding to using either the linear or the RBF kernel) for comparability.

Table 3: Test set classification accuracies and true positive fractions. The accuracies used for final ranking in the challenge are marked in bold font.

	4-class CA[a]		3-class CA[b]		TPF_{NC}	TPF_{MCI}	TPF_{cMCI}	TPF_{AD}
	A+C	C	A+C	C	C	C	C	C
Ensemble SVM								
linear	37.0	**55.6**	50.6	68.8	70.0	20.0	42.5	90.0
RBF	35.6	**55.0**	48.0	68.1	72.5	15.0	42.5	90.0
Individual SVMs[c]								
linear	35.4 (1.4)	54.3 (2.3)	48.6 (2.1)	67.4 (2.3)	65.7 (6.9)	23.0 (6.4)	40.6 (7.5)	87.7 (3.7)
RBF	35.1 (1.4)	52.9 (2.0)	48.3 (2.6)	65.6 (3.0)	63.3 (7.6)	22.0 (8.3)	40.5 (7.1)	85.6 (3.6)

Abbreviations: A+C, Artificial+Challenge test set; C, Challenge test set; CA, classification accuracy; TPF_i, true positive fraction for the ith class.
[a] NC vs. MCI vs. cMCI vs. AD.
[b] NC vs. (MCI ∪ cMCI) vs. AD.
[c] Mean (SD) classification accuracy of the 25 individual SVMs in the ensemble.

2.5. Challenge structure

The challenge organizers generated an additional 340 artificial test observations that were joined with the real test observations in the Challenge test set to form a combined test set of 500 observations. This combined test set was used in the online part of the competition that was hosted on the Kaggle competition platform (Sarica et al., 2016). We term this combined set the Artificial+Challenge test set. The Artificial+Challenge test set was split in half in a public and a private test set. During the online part of the challenge between December 21, 2016 and June 1, 2017, teams could submit an attempt each day and get the result on the public test set. When the challenge ended June 1, 2017, the performance of one attempt selected by each team was evaluated on the private test set. Based on the results of the online part of the challenge, two attempts were selected for final evaluation and ranking based on the Challenge test set, i.e., base on the real test data only. The labels of the test data were released to the participants post-challenge for use in manuscript writing.

Table 4: Challenge test set confusion matrices for the ensemble SVMs. Rows correspond to predicted class and columns correspond to the true class.

Linear kernel				
	NC	MCI	cMCI	AD
NC	28	16	6	0
MCI	7	8	11	0
cMCI	5	10	17	4
AD	0	6	6	36

RBF kernel				
	NC	MCI	cMCI	AD
NC	29	20	7	0
MCI	5	6	10	0
cMCI	6	11	17	4
AD	0	3	6	36

3. Results and discussion

3.1. Classification results

CAs were computed for the Artificial+Challenge test set and for the Challenge test set. Both the performance of the ensemble SVM and of the individual SVMs in the ensemble were evaluated. In the case of the individual SVMs, the mean and standard deviation of the CAs of the individual SVMs were computed. The CA for the 3-class classification problem of NC vs. MCI vs. AD, obtained by joining the MCI and cMCI classifications, was also evaluated to enable comparison with previous challenges that considered this 3-class problem (Bron et al., 2015; Simmons et al., 2014). All the CAs are reported in Table 3.

The CAs for the test set including the artificial data were markedly lower than for the test set containing only the real data across all classifiers, indicating a systematic difference between the artificial and real data. The ensemble SVM achieved higher CA than the mean CA of the individual SVMs in the ensemble in all cases except for 3-class CA using an RBF kernel, which showed similar performance, demonstrating the benefit of ensembling. The CAs for the 3-class problem containing the joint MCI and cMCI classes were more than 12 percent points higher compared with the 4-class problem considered in the challenge. This gap would likely increase had the SVMs been trained directly for the 3-class problem.

The performance of the ensembles on the Challenge learning set was evaluated using the out-of-bag estimate of the CA (Kuncheva, 2014), and it was 70.8% for both kernel types. In comparison with the Challenge test set CAs (Table 3, 55.6% and 55.0%), there were more than 15 percentage point differences, which could indicate that training was not saturated and that performance could be increased by enlarging the training data.

The 3-class CAs between 65.6% and 68.8% were substantially higher than the CAs reported in the Computer-Aided Diagnosis of Dementia (CADDementia) challenge where the high-

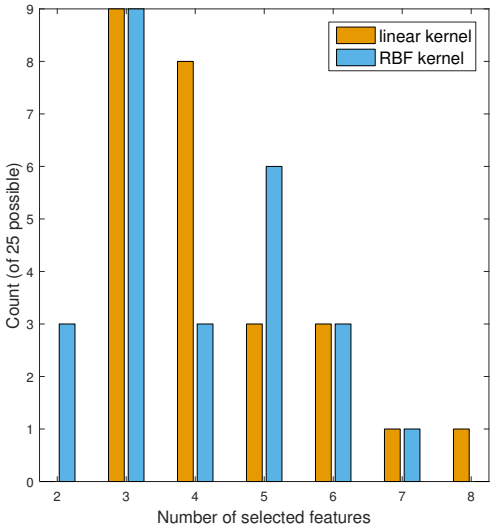

Figure 2: Number of features selected in the 25 SFS procedures in the ensemble construction.

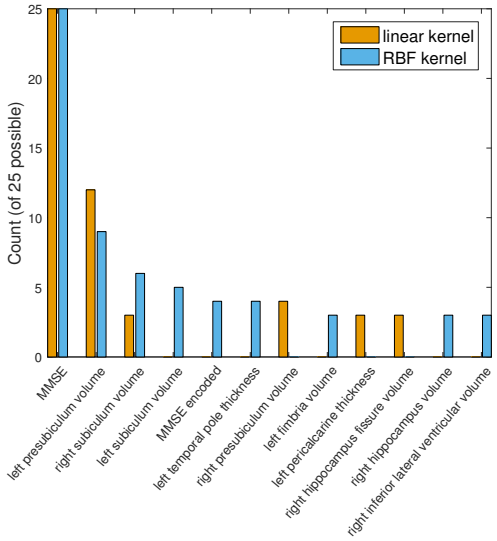

Figure 3: Specific features selected in the 25 SFS procedures in the ensemble construction.

est CA was 63.0% (Bron et al., 2015), and the Alzheimer's Disease Big Data (ADBD) DREAM challenge #1 where the highest CA was 60.2% (Simmons et al., 2014). This was likely because MMSE was included as a feature in the present challenge. Indeed, disregarding MMSE and re-running the algorithms resulted in 3-class CAs of 61.9% and 62.5% for the linear and the RBF SVM ensembles. Another factor was likely that training and test set originated from different cohorts in the two mentioned challenges (Simmons et al., 2014; Bron et al., 2015). In a recent study by Zhu et al. (2016) that used both MRI and positron emission tomography image features, a within-ADNI 3-class CA of 73.0% and a 4-class CA of 62.0% were reported for the same classification problems as considered in the present challenge and study.

Confusion matrices were computed for the Challenge test set using the classifications of the ensemble SVMs (Table 4). The (mis-)classification tendencies were the same for both the linear and the RBF kernel. NC and AD were perfectly separated, AD was almost perfectly classified and only in a few cases misclassified as cMCI, NC was in approximately 30% of the cases misclassified as either MCI or cMCI, and the MCI and cMCI classes were badly classified with MCI true positive fractions of 20% and 15% for the linear and RBF kernel and cMCI true positive fractions of 42.5% for both kernels. The largest class overlaps were between NC and MCI, and between MCI and cMCI. The most comparable algorithms (SVM with volumetric or volumetric and cortical thickness MRI features) in the CADDementia challenge also showed a tendency of good discrimination between NC and AD, and more misclassifications between NC and MCI than between MCI and AD (Bron et al., 2015).

3.2. Selected features

The number of features selected in the 25 SFS procedures in the ensemble construction were plotted (Figure 2). Note that baseline MMSE, which was apriori specified to always be included in the feature set, was counted among these. The feature set size distributions were relatively similar for the two SVM kernels with a tendency of set sizes between 3 and 6 features. The RBF kernel resulted in a feature set size of 2 in some cases whereas the smallest size for the linear kernel was 3. The largest feature set size of 8 appeared in one case for the linear kernel.

Left presubiculum volume was the most frequently selected feature followed by the right subiculum volume, and both features were selected for both kernels (Figure 3). Less frequently selected MRI features included other hippocampal subregional volumetric measures, the right hippocampal volume and the volume of the neighboring right inferior lateral ventricle, left temporal pole cortical thickness, and left pericalcarine cortical thickness. Note that baseline MMSE was always included apriori and that its 100% occurrence therefore was not a result of the SFS procedure. Figure 3 does not depict features the were selected less than 3 times. A total of 41 and 34 different features were selected once for the linear and RBF kernel, and 8 and 2 were selected twice.

As expected, measurements of the hippocampus were important features for the considered classification problem. It is well known from pathological studies that the hippocampus is affected in AD (West et al., 1994; Braak and Braak, 1997), and hippocampal volumetry has previously demonstrated both separation between NC, MCI, and AD (Convit et al., 1997) and between MCI and cMCI (Jack et al., 1999). Moreover, the presubiculum/subiculum is the subregion of the hippocampus where Aβ-amyloid protein, a major pathological hallmark of AD, is deposited the earliest in AD (Braak and Braak, 1997), and this part of the hippocampus suffers significant neuronal loss in AD (West et al., 1994).[1] In line with the high frequency of selection of the presubiculum/subiculum in this study, a pre-

[1]Note that there are differences in the subregion definitions between FreeSurfer (Van Leemput et al., 2009) and the definitions in West et al. (1994);

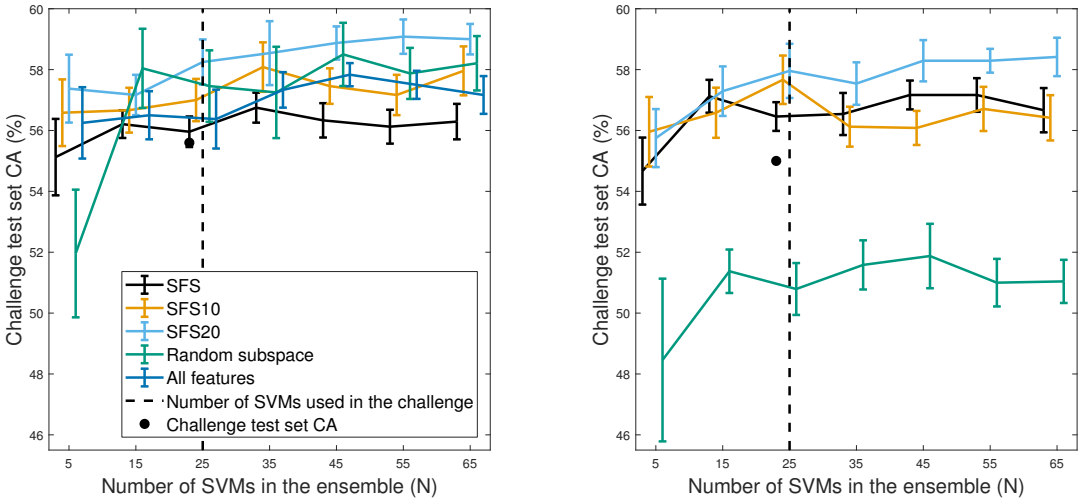

Figure 4: Challenge test set CA for different feature subset methods as a function of the number of SVMs in the ensembles. Computed as the mean of 15 ensembles. Error bars mark the 95% confidence interval. (a) Linear kernel. (b) RBF kernel.

vious study using FreeSurfer hippocampal subregion volumetry found the presubiculum and subiculum to best discriminate NC vs. AD and NC vs. MCI, and that these subregions were the only ones that could discriminate MCI vs. AD (Carlesimo et al., 2015).

In addition to the subtle information captured by hippocampal subregional volumetry, other types of hippocampus features that target subtle information could be considered in an attempt to improve the classification performance. For example, the shape of the hippocampus (Gerardin et al., 2009; Achterberg et al., 2014) or textural patterns of the intensities within the hippocampus (Chincarini et al., 2011; Sørensen et al., 2016). In the CADDementia challenge, a mix of volumetry, cortical thickness measures, and shape and/or intensity texture was used by all of the top-three performing algorithms (Bron et al., 2015). In the ADBD DREAM challenge #1, shape features were provided alongside volumetric and cortical thickness measures for many brain regions (Simmons et al., 2014). However, it was not reported which of these features were used by the winning algorithm.

3.3. *Post-challenge analysis of feature subset methods and ensemble size*

The Challenge test set labels, released post-challenge, were used to investigate the effect of different feature subset methods within the algorithm as a function of the number of base classifiers in the ensemble. A range of $[5, 15, \ldots, 65]$ SVMs was considered. For each number of base classifiers, 15 ensembles were constructed using different Challenge learning set permutations and applied to the Challenge test set, and the mean of the resulting 15 CAs was computed . The same learning set permutations were used for all the investigated feature subset methods, and they all used bagging without replacement. The following five feature subset methods were evaluated:

- *All features:* Using all 128 features.
- *Random subspace:* Random subspace using $128/2 = 64$ features (Ho, 1998).
- *SFS:* The method used in the challenge.
- *SFS10:* The method used in the challenge with the additional requirement that at least 10 features should be selected. In cases when no candidate features improved the CA from the previous iteration of the SFS procedure, the best candidate was still added to the set of selected features.
- *SFS20:* Same as SFS10, but with the requirement that at least 20 features should be selected.

SFS10 and SFS20 were investigated because it was observed that using all 128 features worked better than using SFS for the linear SVM, indicating that the SFS was stuck in local minima, and enforcing a larger feature set size may overcome some of these.

The results are shown in Figure 4. The curve for RBF SVM using all features, that peaked at 43.1% for 15 base classifiers, was not shown for better visualization of the other curves. Increasing the number of base classifiers would only have benefited marginally for the method used in the challenge (i.e. for SFS). However, using another feature subset method would have. Both random subspace with the heuristic of using 50% of the features, enforcing at least 10 or 20 selected features in SFS, or simply using all available features, showed improved performance for the linear SVM ensemble, and for the RBF ensemble, enforcing at least 20 features selected in SFS improved performance. The highest obtained CAs were obtained using SFS20, with 59.1% for the linear kernel at 55 base classifiers and 58.4%

the presubiculum in FreeSurfer roughly corresponds to the subiculum in West et al. (1994), and part of the subiculum in FreeSurfer overlaps with CA1 in West et al. (1994).

for the RBF kernel at 65 base classifiers. These results require confirmation in an independent dataset.

The clear benefit of enforcing a minimum amount of features selected in the SFS procedure indicated that some local minima was overcome by considering several features simultaneously during the selection step. Feature selection approaches designed to handle such situations, such as floating search methods (Pudil et al., 1994), may further improve performance without incurring a significant increase in the computational complexity of the method.

It is further noted that the ensemble instance used for final ranking in the challenge, which was dependent on the particular Challenge learning set permutations used, appeared to be among the more unlikely cases for the RBF kernel. The RBF ensemble Challenge test set CA of 55.0 % was below the corresponding estimated mean CA across 15 ensembles of size 25 SVMs and outside the 95% confidence interval.

4. Conclusions

An ensemble of SVMs was proposed for multi-class classification of NC vs. MCI vs. cMCI vs. AD using structural MRI features and MMSE score. The ensemble algorithm combined bagging without replacement and SFS for optimal feature subset selection. Experiments were conducted using both linear and RBF kernels in the SVMs achieving CAs of 55.6% and 55.0% in the International Challenge for Automated Prediction of MCI from MRI Data. This was better than average single SVM classifications and resulted in a third place in the challenge. Ensemble methods can be used to improve the performance of the commonly applied SVM algorithm in dementia classification providing competitive classification performance.

Disclosures

L. Sørensen report no disclosures. M. Nielsen is shareholder in Biomediq A/S.

Acknowledgments

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; 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 Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

References

Achterberg, H.C., van der Lijn, F., den Heijer, T., Vernooij, M.W., Ikram, M.A., Niessen, W.J., de Bruijne, M., 2014. Hippocampal shape is predictive for the development of dementia in a normal, elderly population. Hum Brain Mapp 35, 2359–2371.

Aguilar, C., Westman, E., Muehlboeck, J.S., Mecocci, P., Vellas, B., Tsolaki, M., Kloszewska, I., Soininen, H., Lovestone, S., Spenger, C., Simmons, A., Wahlund, L.O., 2013. Different multivariate techniques for automated classification of MRI data in Alzheimer's disease and mild cognitive impairment. Psychiatry research 212, 89–98.

Anaissi, A., Goyal, M., Catchpoole, D.R., Braytee, A., Kennedy, P.J., 2016. Ensemble feature learning of genomic data using support vector machine. PloS one 11, e0157330.

Arbabshirani, M.R., Plis, S., Sui, J., Calhoun, V.D., 2017. Single subject prediction of brain disorders in neuroimaging: promises and pitfalls. NeuroImage 145, 137–165.

Braak, H., Braak, E., 1997. Frequency of stages of Alzheimer-related lesions in different age categories. Neurobiol Aging 18, 351–357.

Breiman, L., 1996. Bagging predictors. Mach Learn 24, 123–140.

Breiman, L., 2001. Random forests. Mach Learn 45, 5–32.

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., Pinto, M., Meireles, J.R., Garrett, C., Bastos-Leite, A.J., Abdulkadir, A., Ronneberger, O., Amoroso, N., Bellotti, R., Cárdenas-Pena, D., Álvarez Mezao, A.M., Dolph, C.V., Iftekharuddin, K.M., Eskildsen, S.F., Coupé, P., Fonov, V.S., Franke, K., Gaser, C., Ledig, C., Guerrero, R., Tong, T., Gray, K.R., Moradi, E., Tohka, J., Routier, A., Durrleman, S., Sarica, A., Fatta, G.D., Sensi, F., Chincarini, A., Smith, G.M., Stoyanov, Z.V., Sørensen, L., Nielsen, M., Tangaro, S., Inglese, P., Wachinger, C., Reuter, M., van Swieten, J.C., Niessen, W.J., Klein, S., for the Alzheimer's Disease Neuroimaging Initiative, 2015. Standardized evaluation of algorithms for computer-aided diagnosis of dementia based on structural MRI: the CADDementia challenge. NeuroImage 111, 562–579.

Carlesimo, G.A., Piras, F., Orfei, M.D., Iorio, M., Caltagirone, C., Spalletta, G., 2015. Atrophy of presubiculum and subiculum is the earliest hippocampal anatomical marker of Alzheimer's disease. Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring 1, 24–32.

Chang, C.C., Lin, C.J., 2011. LIBSVM: A library for support vector machines. ACM Trans Intell Syst Technol 2, 27:1–27:27.

Chen, Y., Zhao, X., Lin, Z., 2014. Optimizing subspace SVM ensemble for hyperspectral imagery classification. IEEE J Sel Topics Appl Earth Observ in Remote Sens 7, 1295–1305.

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

Convit, A., De Leon, M., Tarshish, C., De Santi, S., Tsui, W., Rusinek, H., George, A., 1997. Specific hippocampal volume reductions in individuals at risk for Alzheimer's disease. Neurobiol Aging 18, 131–8.

Cortes, C., Vapnik, V., 1995. Support-vector networks. Mach Learn 20, 273–297.

Cuingnet, R., Gerardin, E., Tessieras, J., Auzias, G., Lehéricy, S., Habert, M.O., Chupin, M., Benali, H., Colliot, O., for the Alzheimer's Disease Neuroimaging Initiative, 2011. Automatic classification of patients with Alzheimer's disease from structural MRI: a comparison of ten methods using the ADNI database. NeuroImage 56, 766–781.

Falahati, F., Westman, E., Simmons, A., 2014. Multivariate data analysis and machine learning in Alzheimer's disease with a focus on structural magnetic resonance imaging. J Alzheimers Dis 41, 685–708.

Fischl, B., Dale, A.M., 2000. Measuring the thickness of the human cerebral cortex from magnetic resonance images. Proc Natl Acad Sci USA 97, 11050–11055.

Fischl, B., Salat, D.H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., van der Kouwe, A., Killiany, R., Kennedy, D., Klaveness, S., Montillo, A., Makris, N., Rosen, B., Dale, A.M., 2002. Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. Neuron 33, 341–355.

Gerardin, E., Chételat, G., Chupin, M., Cuingnet, R., Desgranges, B., Kim, H.S., Niethammer, M., Dubois, B., Lehéricy, S., Garnero, L., Eustache, F., Colliot, O., for the Alzheimer's Disease Neuroimaging Initiative, 2009. Multidimensional classification of hippocampal shape features discriminates Alzheimer's disease and mild cognitive impairment from normal aging. NeuroImage 47, 1476–1486.

Ho, T.K., 1998. The random subspace method for constructing decision forests. IEEE Trans Pattern Anal Mach Intell 20, 832 – 844.

Jack, C., Petersen, R., Xu, Y., O'Brien, P., Smith, G., Ivnik, R., Boeve, B., Waring, S., Tangalos, E., Kokmen, E., 1999. Prediction of AD with MRI-based hippocampal volume in mild cognitive impairment. Neurology 52, 1397–403.

Jain, A., Duin, R., Mao, J., 2000. Statistical pattern recognition: a review. IEEE Trans Pattern Anal Mach Intell 22, 4–37.

Khoshgoftaar, T.M., Hulse, J.V., Napolitano, A., 2011. Comparing boosting and bagging techniques with noisy and imbalanced data. IEEE Trans Syst, Man, Cybern A, Syst,Humans 41, 552–568.

Kuncheva, L.I., 2014. Combining pattern classifiers: methods and algorithms. John Wiley & Sons, Inc. 2nd edition.

Kuncheva, L.I., Rodriguez, J.J., Plumpton, C.O., Linden, D.E.J., Johnston, S.J., 2010. Random subspace ensembles for fMRI classification. IEEE Trans Med Imag 29, 531–542.

Lienemann, K., Plötz, T., Fink, G., 2007. On the application of SVM-ensembles based on adapted random subspace sampling for automatic classification of NMR data, in: Haindl M., Kittler J., R.F. (Ed.), Multiple Classifier Systems, Springer, Berlin, Heidelberg. pp. 42–51.

Nanni, L., 2006. An ensemble of classifiers for the diagnosis of erythemato-squamous diseases. Neurocomputing 69, 842–845.

Petersen, R.C., Aisen, P.S., Beckett, L.A., Donohue, M.C., Gamst, A.C., Harvey, D.J., Jack, Jr, C., Jagust, W.J., Shaw, L.M., Toga, A.W., Trojanowski, J.Q., Weiner, M.W., 2010. Alzheimer's Disease Neuroimaging Initiative (ADNI): clinical characterization. Neurology 74, 201–209.

Pudil, P., Novovičová, J., Kittler, J., 1994. Floating search methods in feature selection. Pattern Recognit Lett 15, 1119–1125.

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

Sabuncu, M.R., Konukoglu, E., for the Alzheimer's Disease Neuroimaging Initiative, 2015. Clinical prediction from structural brain MRI scans: a large-scale empirical study. Neuroinformatics 13, 31–46.

Sarica, A., Cerasa, A., Quattrone, A., Calhoun, V., for the Alzheimer's Disease Neuroimaging Initiative, 2016. A machine learning neuroimaging challenge for automated diagnosis of mild cognitive impairment. URL https://inclass.kaggle.com/c/mci-prediction. Accessed December 29, 2017.

Schmansky, N., Desikan, R., Stevens, A., Nguyen, K., Moreau, A., 2017. FreeSurferWiki - cortical parcellation. URL https://surfer.nmr.mgh.harvard.edu/fswiki/CorticalParcellation. Accessed December 29, 2017.

Shen, K.K., Fripp, J., Mériaudeau, F., Chételat, G., Salvado, O., Bourgeat, P., for the Alzheimer's Disease Neuroimaging Initiative, 2012. Detecting global and local hippocampal shape changes in Alzheimer's disease using statistical shape models. NeuroImage 59, 2155–2166.

Simmons, A., Klein, A., Logsdon, B., Fardo, D., Suver, C., Bare, C., Stolovitzky, G., Kauwe, J., Mangravite, L., Peters, M., Tustison, N., Dobson, R., Ghosh, S., Friend, S., Newhouse, S., Maxwell, T., Norman, T., 2014. Alzheimers Disease Big Data DREAM Challenge #1. URL https://www.synapse.org/#!Synapse:syn2290704. Accessed December 29, 2017.

Simpson, I.J.A., Woolrich, M.W., Andersson, J.L.R., Groves, A.R., Schnabel, J.A., 2013. Ensemble learning incorporating uncertain registration. IEEE Trans Med Imag 32, 748–756.

Sørensen, L., Igel, C., Liv Hansen, N., Osler, M., Lauritzen, M., Rostrup, E., Nielsen, M., for the Alzheimer's Disease Neuroimaging Initiative and the Australian Imaging Biomarkers and Lifestyle flagship study of ageing, 2016. Early detection of Alzheimer's disease using MRI hippocampal texture. Hum Brain Mapp 37, 1148–1161.

Sørensen, L., Igel, C., Pai, A., Balas, I., Anker, C., Lillholm, M., Nielsen, M., for the Alzheimer's Disease Neuroimaging Initiative and the Australian Imaging Biomarkers and Lifestyle flagship study of ageing, 2017. Differential diagnosis of mild cognitive impairment and Alzheimer's disease using structural MRI cortical thickness, hippocampal shape, hippocampal texture, and volumetry. NeuroImage Clin. 13, 470–482.

Tao, D., Tang, X., Li, X., Wu, X., 2006. Asymmetric bagging and random subspace for support vector machines-based relevance feedback in image retrieval. IEEE Trans Pattern Anal Mach Intell 28, 1088–1099.

Van Leemput, K., Bakkour, A., Benner, T., Wiggins, G., Wald, L.L., Augustinack, J., Dickerson, B.C., Golland, P., Fischl, B., 2009. Automated segmentation of hippocampal subfields from ultra-high resolution in vivo MRI. Hippocampus 19, 549–557.

Varol, E., Gaonkar, B., Erus, G., Schultz, R., Davatzikos, C., 2012. Feature ranking based nested support vector machine ensemble for medical image classification, in: Proc IEEE Int Symp Biomed Imaging, pp. 146–149.

Waske, B., van der Linden, S., Benediktsson, J.A., Rabe, A., Hostert, P., 2010. Sensitivity of support vector machines to random feature selection in classification of hyperspectral data. IEEE Trans Geosci Remote Sens 48, 2880–2889.

Weiner, M.W., Veitch, D.P., Aisen, P.S., Beckett, L.A., Cairns, N.J., Cedarbaum, J., Green, R.C., Harvey, D., Jack, C.R., Jagust, W., Luthman, J., Morris, J.C., Petersen, R.C., Saykin, A.J., Shaw, L., Shen, L., Schwarz, A., Toga, A.W., Trojanowski, J.Q., for the Alzheimer's Disease Neuroimaging Initiative, 2015. 2014 update of the Alzheimer's Disease Neuroimaging Initiative: a review of papers published since its inception. Alzheimers Dement 11, e1–120.

West, M.J., Coleman, P.D., Flood, D.G., Troncoso, J.C., 1994. Differences in the pattern of hippocampal neuronal loss in normal ageing and Alzheimer's disease. Lancet 344, 769–772.

Xia, J., Chanussot, J., Du, P., He, X., 2016. Rotation-based support vector machine ensemble in classification of hyperspectral data with limited training samples. IEEE Trans Geosci Remote Sens 54, 1519–1531.

Zhu, X., Suk, H.I., Lee, S.W., Shen, D., 2016. Canonical feature selection for joint regression and multi-class identification in Alzheimer's disease diagnosis. Brain Imaging Behav 10, 818–828.