

## Automated detection of brain atrophy patterns based on MRI for the prediction of Alzheimer's disease

Claudia Plant<sup>a</sup>, Stefan J. Teipel<sup>b,d,i</sup>, Annahita Oswald<sup>c</sup>, Christian Böhm<sup>c</sup>, Thomas Meindl<sup>e</sup>, Janaina Mourao-Miranda<sup>f,g</sup>, Arun W. Bokde<sup>d,h</sup>, Harald Hampel<sup>d,h</sup>, Michael Ewers<sup>d,h,\*</sup>

<sup>a</sup> Department of Neuroradiology, Technische Universität München, Munich, Germany

<sup>b</sup> Department of Psychiatry, University of Rostock, Germany

<sup>c</sup> Department for Computer Science, University of Munich, Munich, Germany

<sup>d</sup> Department of Psychiatry, Ludwig-Maximilians University, Munich, Germany

<sup>e</sup> Institute for Clinical Radiology, Department of MRI, Ziemssenstrasse 1, 80336 Munich, Germany

<sup>f</sup> Centre for Computational Statistics and Machine Learning, University College London, London, UK

<sup>g</sup> Centre for Neuroimaging Sciences, King's College London, London, UK

<sup>h</sup> Trinity College Dublin, School of Medicine, Discipline of Psychiatry & Trinity College Institute of neuroscience (TCIN) & The Adelaide and Meath Hospital Incorporating The National Children's Hospital (AMiNCH), Dublin, Ireland

<sup>i</sup> Deutsches Zentrum für Neurodegenerative Erkrankungen (DZNE), Germany

### ARTICLE INFO

#### Article history:

Received 6 August 2009

Revised 13 November 2009

Accepted 18 November 2009

Available online 2 December 2009

### ABSTRACT

Subjects with mild cognitive impairment (MCI) have an increased risk to develop Alzheimer's disease (AD). Voxel-based MRI studies have demonstrated that widely distributed cortical and subcortical brain areas show atrophic changes in MCI, preceding the onset of AD-type dementia. Here we developed a novel data mining framework in combination with three different classifiers including support vector machine (SVM), Bayes statistics, and voting feature intervals (VFI) to derive a quantitative index of pattern matching for the prediction of the conversion from MCI to AD. MRI was collected in 32 AD patients, 24 MCI subjects and 18 healthy controls (HC). Nine out of 24 MCI subjects converted to AD after an average follow-up interval of 2.5 years. Using feature selection algorithms, brain regions showing the highest accuracy for the discrimination between AD and HC were identified, reaching a classification accuracy of up to 92%. The extracted AD clusters were used as a search region to extract those brain areas that are predictive of conversion to AD within MCI subjects. The most predictive brain areas included the anterior cingulate gyrus and orbitofrontal cortex. The best prediction accuracy, which was cross-validated via train-and-test, was 75% for the prediction of the conversion from MCI to AD. The present results suggest that novel multivariate methods of pattern matching reach a clinically relevant accuracy for the a priori prediction of the progression from MCI to AD.

© 2009 Elsevier Inc. Open access under [CC BY license](http://creativecommons.org/licenses/by/3.0/).

### Introduction

Alzheimer's disease (AD) is the most frequent cause of age-related dementia. Due to the increasing proportion of elderly people in the Western societies, the prevalence of dementia is projected to double within the next three decades (Ferri et al., 2005). The reliable and early detection of AD in predementia stages such as mild cognitive impairment (MCI) is the basis for the development of preventive treatment approaches. However, especially the diagnosis of mild AD and prediction of development of AD in at-risk groups remains challenging. In addition to cerebrospinal fluid derived biomarkers (Blennow and

Hampel, 2003; Ewers et al., 2007; Hansson et al., 2006; Herukka et al., 2007; Zhong et al., 2007), neuroimaging markers have been recommended to be included in the revised NINCDS-ADRDA diagnostic standard criteria (Dubois et al., 2007) and proposed as predictors of AD (Winblad et al., 2004; Petersen et al., 2001). The best established MRI derived marker of AD, hippocampus volume, shows relatively high diagnostic accuracy for AD but clinically insufficient predictive value for the prediction of progression from MCI to AD when assessed as the sole predictor (Csernansky et al., 2005; Jack et al., 1999; Kantarci and Jack, 2003; Killiany et al., 2002; Pennanen et al., 2004; Stoub et al., 2005; Visser et al., 1999).

As an alternative to ROI based volumetry, automated morphometry and deformation-based approaches have been developed to map the pattern of structural brain changes across the entire brain (Ashburner and Friston, 2000; Good et al., 2001). A series of voxel-based morphometric studies in MCI and mild AD have shown marked volume differences not only within the hippocampus area but also

\* Corresponding author. Discipline of Psychiatry, School of Medicine and Trinity College Institute of Neuroscience, Trinity College, University of Dublin, Trinity Centre for Health Sciences, The Adelaide and Meath Hospital Incorporating The National Children's Hospital (AMiNCH), Tallaght, Dublin 24, Ireland.

E-mail address: [ewersm@tcd.ie](mailto:ewersm@tcd.ie) (M. Ewers).

distributed within cortical brain areas such as the precuneus and cingulate gyrus (Baron et al., 2001; Chetelat et al., 2002, 2005; Frisoni et al., 2002; Karas et al., 2004; Pennanen et al., 2005). However, few statistical approaches have been proposed to derive individual risk scores from such maps of atrophy for the clinical prediction of AD. Data mining approaches and pattern recognition methods provide a way to extract from millions of voxels within an MRI the minimal set of voxel values necessary to attain a sufficiently high accuracy for the prediction and diagnosis of AD. Multivariate approaches such as principal component analysis (PCA) (Friston et al., 1996), independent component analysis (McKeown et al., 1998), structural equation modeling (McIntosh et al., 1994), and support vector machine (Mourao-Miranda et al., 2005, 2006) are potential candidates but have mostly been applied to functional neuroimaging data so far. Recently, such multivariate methods have been adopted for the analysis of structural MRI to detect spatial patterns of atrophy in AD (Chen and Herskovits, 2006; Davatzikos et al., 2008; DeCarli et al., 1995; Duchesne et al., 2008a,b, 2009; Fan et al., 2008; Kloeppel et al., 2008; Misra et al., 2009; Teipel et al., 2007a; Vemuri et al., 2008). These techniques allow for deriving a single value representing the degree to which a disease-specific spatial pattern of atrophy is present in a single individual. The application of such classifiers of spatial pattern of atrophy in MCI has shown promising results for the prediction of AD (Davatzikos et al., 2008; Teipel et al., 2007a).

In the present study we applied a novel two-step approach combining a distribution free feature selection algorithm at the first stage and, at the second stage, different multivariate classifiers for case-by-case decision making. The major aims of the current study were, first, to develop a novel feature selection method to circumvent potential problems of previous approaches for feature selection including lack of statistical power due to multiple testing (Fan et al., 2008) or purely-data driven correlational patterns in unsupervised dimensionality reduction (e.g. PCA (Teipel et al., 2007a,b)). Secondly, we compared different cross-validated classifiers including support vector machine (SVM), a Bayesian classifier, and voting feature intervals (VFI) combined with unsupervised clustering algorithms to derive the minimal set of voxels for optimized prediction of diagnosis (AD vs. HC) or prediction of AD in MCI. The overall goal was to derive an optimized classification that is sensitive for the early MRI-based detection of AD.

## Materials and methods

### Subjects

32 patients with clinically probable AD, 24 patients with amnesic MCI and 18 healthy control subjects (HC) underwent MRI and clinical examinations (Table 1).

AD patients fulfilled the criteria of the National Institute of Neurological Communicative Disorders and Stroke and the Alzheimer Disease and Related Disorders Association (NINCDS-ADRDA) criteria

for clinically probable AD (McKhann et al., 1984). MCI subjects fulfilled the Mayo criteria for amnesic MCI (Petersen et al., 2001). All MCI subjects had subjective memory complaints, a delayed verbal recall score at least 1.5 standard deviations below the respective age norm, normal general cognitive function, and normal activities of daily living. Severity of cognitive impairment was assessed by the Mini-Mental-State-Examination (MMSE) (Folstein et al., 1975). Controls did not have cognitive complaints and scored within 1 standard deviation from the age adjusted norm on all subtests of the CERAD cognitive battery (Morris et al., 1989).

MCI patients received clinical follow-up examinations over approximately 2.5 years, using clinical examination and neuropsychological testing to determine which subjects converted to AD and which remained stable. All subjects were only examined if they gave their written informed consent. The study was approved by the institutional review board of the Clinic of Psychiatry at the Ludwig Maximilian University of Munich.

### MRI acquisition

MRI examinations of the brain were performed on a 1.5 T MRI scanner (Magnetom Vision, Siemens Medical Solutions, Erlangen, Germany). We acquired a high-resolution T1-weighted Magnetisation Prepared Rapidly Acquired Gradient echo (MPRAGE) 3D-sequence with a resolution of 0.55 by 0.55 by 1.1 mm<sup>3</sup>, TE = 3.9 ms, TI = 800 ms, and TR = 1,570 ms. The FOV was 240 mm and the pixel matrix was 512 × 512.

### MRI processing

The preprocessing of the scans was conducted with the statistical software package SPM2 (Wellcome Trust Centre for Neuroimaging, London, <http://www.fil.ion.ucl.ac.uk/spm/>). The high dimensional normalization of the MRI scans was processed according to a protocol that has been described in detail previously (Teipel et al., 2007a,b). First, we constructed a customized template across groups averaged across images that were normalized to the standard MNI T1 MRI template, using the low-dimensional transformation algorithm implemented in SPM2 (Ashburner and Friston, 2000; Ashburner et al., 1997). Next, one good quality MRI scan of a healthy control subject was normalized to this anatomical average image using high-dimensional normalization with symmetric priors (Ashburner et al., 1999) resulting in a pre-template image. Finally, the MRI scan in native space of the same subject was normalized to this pre-template image using high-dimensional normalization. The resulting volume in standard space served as the anatomical template for subsequent normalizations of the remaining scans. The individual anatomical scans in standard space (after low-dimensional normalization) were normalized to the anatomical template using high-dimensional image warping (Ashburner et al., 1999). These normalized images were resliced to a final isotropic voxel size of 1.0 mm<sup>3</sup>. Finally, we derived Jacobian determinant maps from the voxel-based transformation tensors. Values above 1 represent an expansion of the voxel, values below 1 a contraction of the voxel from the template to the reference brain. The resulting Jacobian determinant maps were masked for brain matter and cerebrospinal fluid (CSF) spaces using masks from the segmented template MRI (Ashburner and Friston, 1997). To obtain the brain mask, the template brain scan was segmented into grey and white matter and CSF spaces. The grey matter (GM) and white matter (WM) compartments then were combined to obtain a brain mask excluding CSF: (GM + WM)/(WM + GM + CSF). \*BRAIN with grey matter, white matter, and CSF representing the grey and white matter and CSF probabilistic maps obtained through segmentation and BRAIN representing the brain mask obtained from the brain extraction step in SPM2.

We took the logarithm of the masked maps of the Jacobian determinants (Scahill et al., 2002) and then applied a 10-mm full

**Table 1**  
Demographic variables and MMSE for the different groups.

Group	Women/men <sup>a</sup>	Age in years mean [SD] <sup>b</sup>	MMSE mean [SD] <sup>c</sup>
Healthy controls	9/9	64.8 [4.0]	29.3 [1.1]
AD patients	20/12	68.8 [8.9]	23.4 [3.0]
MCI patients	13/11	69.7 [8.5]	27.0 [1.8]

<sup>a</sup> Not different between groups,  $\chi^2 = 0.83$  with 2 df,  $p = 0.66$ .

<sup>b</sup> One-way analysis of variance (ANOVA),  $F_{2,1}^{11} = 2.2$ ,  $p = 0.114$ , two-tailed t-test AD vs. control subjects:  $t_{48} = 1.8$ ,  $p = 0.08$ , two-tailed t-test AD vs. MCI:  $t_{54} = 0.4$ ,  $p = 0.69$ , two-tailed t-test MCI vs. control subjects:  $t_{40} = 2.3$ ,  $p = 0.04$ .

<sup>c</sup> Significantly different between groups, Kruskal–Wallis ANOVA  $\chi^2 = 43.0$ ,  $p < 0.001$ , significant difference in all pair-wise comparisons using Mann–Whitney U test at  $p < 0.001$ .

width at half maximum isotropic Gaussian kernel. The masked smoothed Jacobian determinant maps were scaled to the same mean value and standard deviation using a voxel-wise z-transformation:

$$z_{i,k} = \frac{x_{i,k} - \bar{x}_k}{s_k}$$

where  $x_{i,k}$  is the FA value of voxel  $i$  in scan  $k$ ,  $\bar{x}_k$  is mean value across all  $x_i$  of scan  $k$  and  $s$  is the standard-deviation across all  $x_i$  of scan  $k$ .

#### Data mining

We applied a multi-step data mining procedure including feature selection, clustering and classification to identify the best discriminating regions in brain images.

**Notations.** Given a data set  $DS$  consisting of MRI scans of  $n$  subjects  $s_1, \dots, s_n$  labeled to a set of  $k$  discrete classes  $C = \{c_1, \dots, c_k\}$  (in our study e.g. HC and AD), we denote the class label of subject  $s_i$  by  $s_i.c$ . For each subject we have an MR image which is represented as a feature vector  $V$  composed of  $d$  voxels  $v_1, \dots, v_d$ .

#### Feature selection

First we select the most discriminating features using a feature selection criterion. We use the Information Gain (Quinlan, 1993; Hall and Holmes, 2003) to rate the interestingness of a voxel for class separation, which requires the following definitions.

**Entropy of the class distribution.** The entropy of the class distribution  $H(C)$  is defined as  $H(C) = -\sum_{c_i \in C} p(c_i) \cdot \log_2(p(c_i))$ , where  $p(c_i)$  denotes the probability of class  $c_i$ , i.e.  $\{s | s \in DS \wedge s.c = c_i\} / n$ .  $H(C)$  corresponds to the required amount of bits to tell the class of an unknown subject and scales between 0 and 1. In the case of  $k=2$ , (e.g. we consider the two classes HC and AD), if the number of subjects per class is equal for both classes,  $H(C)=1$ . In the case of unbalanced class sizes the entropy of the class distribution is smaller than one and approaches zero if there are much more instances of one class than of the other class.

**Information Gain of a voxel.** Now we can define the Information Gain of a voxel  $v_i$  as the amount by which  $H(C)$  decreases through the additional information provided by  $v_i$  on the class, which is described by the conditional entropy  $H(C|v_i)$ .

$$IG(v_i) = H(C) - H(C|v_i).$$

In the case of  $k=2$ , the Information Gain scales between 0 and 1, where 0 means that the corresponding voxel provides no information on class label of the subjects. An Information Gain of 1 means that the class labels of all subjects can be derived from the corresponding voxel without any errors.

To compute the conditional entropy, features with continuous values, as in our case, need to be discretized using the algorithm of Fayyad and Irani (1993). This method aims at dividing the attribute range into class pure intervals. The cut points are determined by the Information Gain of the split. Since a higher number of cut points always implies higher class purity but may lead to over fitting, an information-theoretic criterion based on the Minimum Description Length principle is used to determine the optimal number and location of the cut points.

#### Clustering

After feature selection, we apply a clustering algorithm to identify groups of adjacent voxels with a high discriminatory power and to remove noise. Clustering algorithms aim at deriving a partitioning of a data set into groups (clusters) such that similar objects are grouped together. We apply clustering to group voxels with similar spatial location in the brain and similar (high) IG. The density-based clustering algorithm DBSCAN (Ester et al., 1996) has been designed

to find clusters of arbitrary shape in databases with noise. In our context, clusters are connected areas of voxels having a high IG which are separated by areas of voxels of lower IG. DBSCAN has been originally designed for clustering data objects represented by feature vectors. We first briefly introduce the general definitions of DBSCAN and then elaborate on the required modifications for clustering voxels. DBSCAN employs a density threshold for clustering, which is expressed by two parameters,  $\epsilon$  specifying a volume and  $MinPts$  denoting a minimum number of objects. Formally, the density-based clustering notion of DBSCAN is defined as follows:

**Definitions of DBSCAN.** An object  $O$  is called *core object* if it has at least  $MinPts$  objects in its  $\epsilon$  range, i.e.  $|N_\epsilon(O)| \geq MinPts$ , where  $N_\epsilon(O) = \{O' | dist(O, O') \leq \epsilon\}$ . An object  $O$  is directly density-reachable from another object  $P$  with respect to  $\epsilon$  and  $MinPts$  if  $P$  is a core object and  $O \in N_\epsilon(P)$ . An object  $O$  is *density-reachable* from an object  $P$  with respect to  $\epsilon$  and  $MinPts$  if there exists a sequence of objects  $O_1, \dots, O_n$  such that  $O_1 = P$  and  $O_n = O$  and  $O_{i+1}$  is directly density-reachable with respect to  $\epsilon$  and  $MinPts$  from  $O_i$  for  $1 \leq i < n$ . Two objects  $O$  and  $P$  are *density-connected* with respect to  $\epsilon$  and  $MinPts$  if there exists an object  $Q$  such that both  $O$  and  $P$  are density-reachable from  $Q$ . A *density-based cluster* is the maximum set of density-connected objects, i.e. the transitive closure of the density reachability relation.

Therefore, it can be proven that a density-based cluster can be efficiently determined by collecting all objects which are density-reachable starting from an arbitrary core object. For an illustration of the definitions of DBSCAN see Fig. 1. For  $MinPts=3$ , a core object, a noise object, and a density-based cluster are displayed.

To adapt the algorithm to our setting, we redefine the core object property and direct density reachability as follows:

**Modified Definitions for Voxel Clustering.** Given two thresholds of Information Gain  $t_{core}$  and  $t_{border}$  and a minimum number of voxels  $MinVox$  we call a voxel  $v_i$  a core voxel if the IG of  $v_i$  is larger than  $t_{core}$  and  $v_i$  is surrounded by at least  $MinVox$  voxels having an IG of at least  $t_{border}$ .

We allow for potentially different thresholds  $t_{core} > t_{border}$  of Information Gain for core voxels and voxels at the boundaries of the clusters to require highly discriminative cluster centers and to model the natural fading of the discriminatory power in the boundary areas of the clusters. However, it is on our specific set of images not necessary to distinguish between  $t_{core}$  and  $t_{border}$  since the voxels either have a significant Information Gain value or an IG of zero. So we set  $t_{core}$  and  $t_{border}$  to the minimum IG in the data set and used  $MinVox=6$ , which means that we require a core voxel to be situated in a neighborhood of highly discriminative voxels.

#### Classification

After clustering, the selected features represent spatially coherent regions which exhibit significant differences among the groups. At this stage, classification algorithms can be applied to validate the discriminatory power of these selected clusters. Classification is a data mining (machine learning) technique used to predict group membership for data instances, which are the subjects in our application. The task of classification involves two major steps: in the so-called training phase, the classifier learns the separating information. To achieve this, some amount of instances with known class labels is required. In the test phase, the classifier predicts the class label of

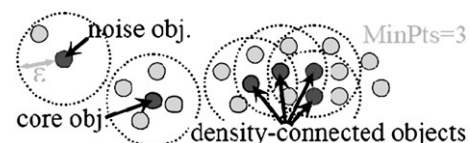
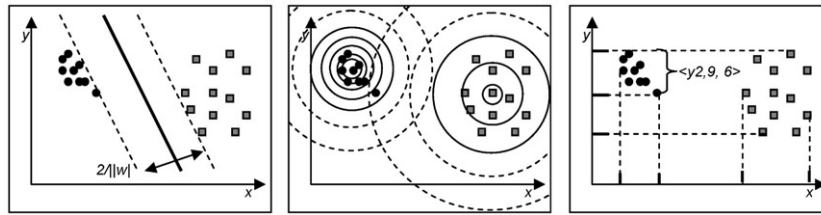


Fig. 1. Definitions of DBSCAN.





**Fig. 2.** Visualizing the different classification paradigms. Left: Support Vector Machine, Center: Bayesian Classification, Right: Voting Feature Intervals.

unlabeled instances based on the learned information. For more information on the validation of classifiers see [Appendix A](#). Among the large variety of classifiers we chose three representative approaches with very different algorithmic paradigms. For an illustration see [Fig. 2](#).

- (1) Linear Support Vector Machine (SVM). SVM aims at constructing a hyperplane separating the training examples. Among all possible hyperplanes, SVM selects the one with the maximum margin between the training examples of both classes.
- (2) Bayesian Classifier (Bayes). The fundamental idea of Bayesian classification is to model each class of the training data by a probability density function. Test objects are then assigned to most probable class.
- (3) Voting Feature Intervals (VFI). Very different to SVM and Bayes, VFI is a simple entropy-based classifier. In the training phase VFI constructs class-pure intervals for each feature and each class. Classification is performed by voting.

For all classifiers, we used the implementations of the WEKA data mining toolkit available at <http://www.cs.waikato.ac.nz/ml/weka/>.

#### Validation

To validate the data mining framework involving the steps feature selection, clustering and classification, we used two established validation techniques: Leave-one-out cross-validation and train-and-test. Both techniques rely on the idea to learn the discriminatory patterns in a training phase on the basis of one data set which is called training data set. In the subsequent test phase, the discriminatory power of the learned pattern is evaluated using a disjoint data set, i.e. the test data. The two validation techniques differ, however, in the way the training and the test data set are composed.

**Cross-validation.** Cross-validation is an established validation scheme in the case of few training examples with respect to the dimensionality of the data ([Kearns and Ron, 1997](#)). For leave-one-out cross-

validation, we divide the data set into  $n$  folds of size  $n-1$  subjects each. In each fold,  $n-1$  subjects are used for training, i.e. we perform the steps feature selection and clustering on these  $n-1$  subjects and obtain a pattern of highly selective clusters. The remaining subject is used as test object, i.e. we predict the class label of this subject by applying a classifier in the feature space defined by the clusters obtained in the training phase.

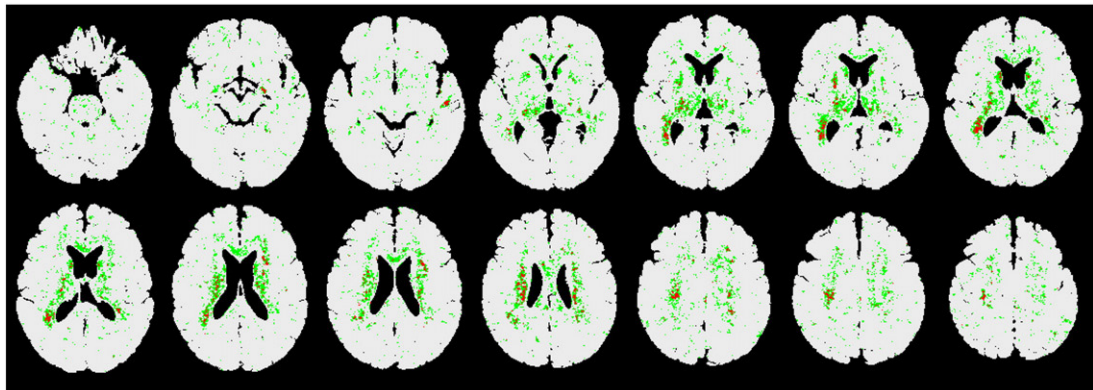
**Train-and-test.** In contrast to cross-validation, which uses disjoint partitions of a single data set for validation, the train-and-test methodology employs two fully different data sets as training and test data. We apply train-and-test validation for prediction of conversion of subjects with MCI. We use HC vs. AD as training data and MCI-MCI vs. MCI-AD as test data.

**Assessment of the classification result.** To evaluate the quality of the classification result, we report three established measures: accuracy, sensitivity and specificity. The accuracy of a classifier is defined as  $acc = \frac{|corr|}{n}$ , whereas  $|corr|$  denotes the number of correctly classified subjects. The sensitivity and specificity evaluates the performance of a classifier to identify positive and negative instances, respectively, i.e.  $sen = \frac{|TP|}{|TP + FN|}$ ,  $spec = \frac{|TN|}{|TN + FP|}$ , whereas  $|TP|$  and  $|TN|$  denotes the number of true positive and true negative instances, and  $|FP|$  and  $|FN|$  the number of false positives and negatives. Following a common convention, we consider a correctly identified Alzheimer's disease case, or a correctly predicted converter as a true positive.

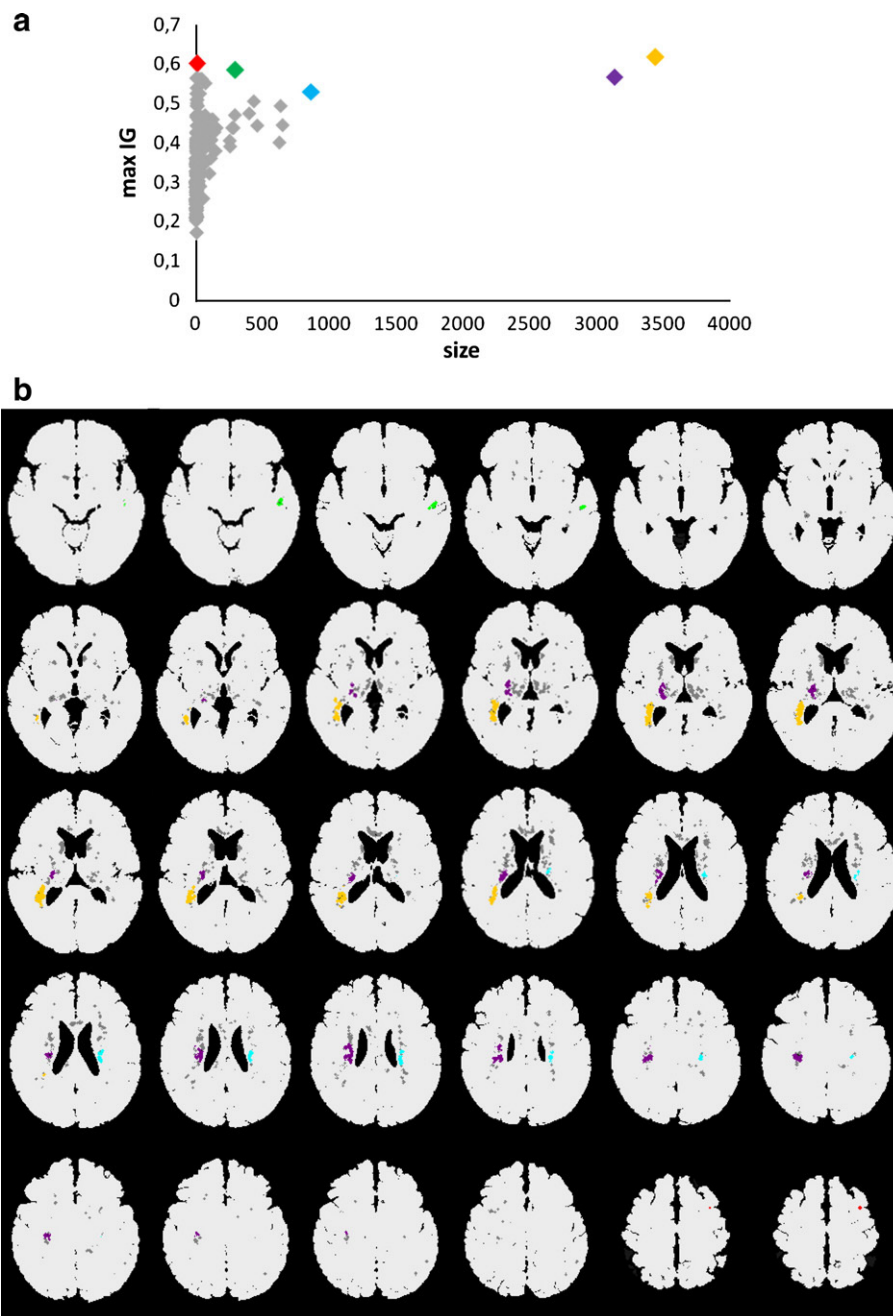
In addition to accuracy, sensitivity and specificity, we report the 95% confidence intervals of these measures as computed by the efficient-score method ([Newcombe, 1998](#)). In particular, we applied Newcombe's fourth method, which is also commonly referred to as the Wilson procedure with continuity correction.

#### Visualization

We display the spatial location of the features best discriminating the classes HC and AD, and MCI-MCI and MCI-AD, respectively. For the ease of comparison, we display in [Figs. 3, 4b and 5b](#) the features which



**Fig. 3.** Selected features for the comparison between AD vs. HC. z-coordinates in Talairach space: top row of images – 45.5, – 33.5, – 26.5, – 18.5, – 13.5, – 11.5, – 5.5; bottom row: – 3.5, 0.5, 4.5, 8.5, 13.5, 15.5, 21.5.

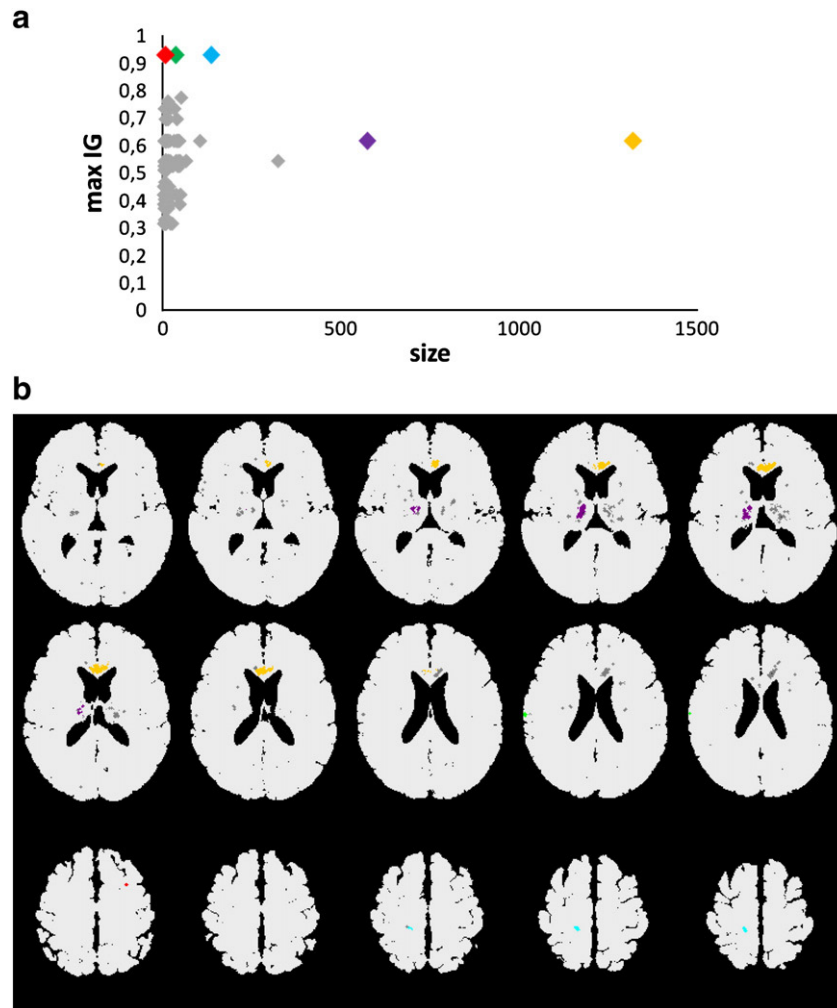


**Fig. 4.** (a) Cluster size and maximum Information Gain AD vs. HC. (b) Selected features after HC vs. AD clustering. Colors: Cluster 1 red, cluster 2 green, cluster 3: blue, cluster 4 purple cluster 5 orange. Remaining clusters gray. Displayed is every second slice starting with  $z = -31.5$  to  $22.5$ ;  $34.5$  and  $35.5$ .

are relevant for classification in all folds. Note that this is only done to obtain one common spatial map for interpretation of the best discriminating regions and the reported classification accuracies are obtained by leave-one-out cross-validation. To facilitate interpretation, we additionally highlight the most interesting clusters in different colors in Figs. 4 and 5. We were interested in clusters which are as large as possible and exhibit an IG as high as possible. Therefore, we selected those clusters in the visualization exhibiting an outstanding combination of both criteria using the *skyline* operator which has been successfully applied in many multi-criteria decision making applications, e.g. in personalized information systems (Hristidis et al., 2001) or for the selection of web services (Skoutas et al., 2008). The skyline of a data set consists of all data objects which are not dominated by any other object in the data set with respect to any possible weighting of the studied criteria. Skylines have been

studied since the 1960s and are also known as Pareto sets or admissible points (Barndorff-Nielsen and Sobel, 1966). Börzsönyi et al. (2001) proposed efficient algorithms for skyline computation. To illustrate the skyline concept, imagine a user looking for hotels which are cheap and close to the beach. The hotels are represented in a database as two-dimensional feature vectors  $\langle \text{distance}, \text{price} \rangle$ . The skyline contains all offers which might be interesting to the user. By definition of dominance it can be guaranteed for all offers in the skyline that there is no better offer, i.e. there is no other hotel which is cheaper and closer to the beach. In our context, we consider clusters of voxels which are described by the features size and IG.

Anatomical location information of the clusters was obtained with the Talairach Daemon software available at <http://www.talairach.org/> after MNI to Talairach coordinate transformation with the non-linear approach (Duncan et al., 2000, source-code



**Fig. 5.** (a) Cluster size and maximum Information Gain for MCI converter vs. MCI non-converter. (b) Skyline clusters of MCI-AD vs. MCI-MCI. Colors: cluster 1: red, cluster 2: green, cluster 3: blue, cluster 4: purple, cluster 5 orange. Displayed are some representative slices containing clusters: z-coordinates in Talairach space:  $-12.5$  to  $5.5$  and  $34.5$  to  $42.5$  (every second slice).

available at <http://imaging.mrc-cbu.cam.ac.uk/imaging/MniTalairach#head-b3a445e55dd349a8b2349acce51ab298c90685b>).

#### SPM based voxel based analysis

A univariate voxel-based analysis using SPM 8 (Wellcome Trust Centre for Neuroimaging, London; freely available at <http://www.fil.ion.ucl.ac.uk/spm/>) was conducted for the group comparisons between AD vs. HC and MCI vs. HC on the basis of the deformation maps derived as described above. The default settings in SPM8 were used, with a proportional scaling and global normalization to the mean of 50. The data were spatially smoothed with a 12 mm Gaussian kernel. A significance value of  $p < 0.001$  uncorrected was chosen for this exploratory analysis.

#### White matter rating

Age-related white matter rating was conducted according to a standardized procedure (Wahlund et al., 2001) by an experienced radiologist (T.M.) who was blinded to the diagnosis. Ratings were done on the basis of T2 weighted fluid-attenuated inversion recovery (FLAIR) images that were taken together with the T1 weighted MRI scans used for the classification experiments. FLAIR scans were available in a subset of 23 AD patients, 24 MCI subjects, and 6 healthy controls. The ratings of white matter hyperintensities (WMH) were done for different brain regions including the basal ganglia (including the striatum and globus pallidus), thalamus, and the internal and

external capsules. In addition, ratings were done for the frontal lobe, temporal lobe, parieto-occipital lobe and the infratentorial brain area within each hemisphere.

The scores were averaged across both hemispheres subsequently. The rating scale ranges from 0 (no lesions) to a maximum score of 3 (confluent lesions) (Wahlund et al., 2001).

#### Results

The mean age, MMSE and the gender distribution for AD, MCI, and HC subjects are displayed in Table 1. Nine out of 24 MCI subjects converted to AD after an average follow-up interval of 2.5 years. We applied our framework with leave-one-out cross validation on three data sets to identify highly selective brain regions for the differentiation between AD vs. HC, MCI vs. HC, and MCI converter (MCI-AD) vs. MCI non-converters (MCI-MCI). For the train-and-test

**Table 2**  
Summary of classification experiments.

Task	Comparison	Validation	Training	Test
1	AD vs. HC	Leave-one-out	n.a.	n.a.
2	MCI-MCI vs. MCI-AD	Leave-one-out	n.a.	n.a.
3	MCI vs. HC	Leave-one-out	n.a.	n.a.
4	MCI-MCI vs. MCI-AD	Train-and-Test	AD vs. HC	MCI

**Table 3**

Classification results. For all classifiers and experiments, accuracy, sensitivity and specificity are provided together with the 95% confidence intervals.

Task	SVM	Bayes	VFI
1			
Accuracy	90% [77.41, 96.26]	92% [79.89, 97.41]	78% [63.67, 88.01]
Sensitivity	96.88% [82.01, 99.84]	93.75% [77.78, 98.27]	65.63% [46.78, 80.83]
Specificity	77.78% [51.92, 92.63]	88.89% [63.93, 98.05]	100% [78.12, 100]
2			
Accuracy	95.83% [76.88, 99.78]	91.67% [71.53, 98.54]	95.83% [76.88, 99.78]
Sensitivity	88.89% [50.67, 99.42]	77.78% [40.19, 96.05]	100% [62.88, 100]
Specificity	100% [74.65, 100]	100% [74.65, 100]	93.33% [66.03, 99.65]
3			
Accuracy	97.62% [85.91, 99.88]	85.71% [70.76, 94.05]	88.1% [73.57, 95.54]
Sensitivity	95.83% [76.88, 99.78]	83.33% [61.81, 94.52]	83.33% [61.81, 94.52]
Specificity	100% [78.12, 100]	88.89% [63.93, 98.05]	94.44% [70.62, 99.71]
4			
Accuracy	50% [29.65, 70.35]	58.33% [28.99, 81.38]	75% [52.95, 89.4]
Sensitivity	55.56% [22.26, 84.66]	46.66% [22.22, 72.57]	55.56% [22.66, 84.66]
Specificity	46.47% [22.28, 72.58]	77.77% [40.19, 96.05]	86.67% [58.39, 97.66]

validation, we used the brain regions identified in the AD vs. HC for the prediction of conversion of patients with MCI. Table 2 summarizes all experiments performed and Table 3 provides a summary of the classification results for all group comparisons, which are explained below in detail. For all classification results, also the 95% confidence interval is provided in Table 3.

#### Classification of AD vs. HC

For the differentiation between AD and HC, a proportion of the voxels ranging between 97.48% and 98.04% had an Information Gain of 0, i.e. they contained no information separating the groups and were therefore excluded from further analysis.

Theoretically, combinations of these features may provide valuable information. However, due to the high dimensionality of the data, an exhaustive search for feature combinations is not applicable.

**Table 4**

Clusters AD vs. HC.

Cluster-ID	Size(voxels)	Max IG	Location	Regions
5 (orange)	3,445	0.62	41.58, 28.28, – 16.98 40.59, 33.42, – 11.34 34.65, 16.96, – 10.52 42.57, 31.32, – 14.61 34.65, 24.47, 3.84 41.58, 26.31, – 17.72 41.58, 11.02, – 12.75 37.62, 22.05, – 5.73 34.65, 12.41, – 4.41 34.65, 17.38, – 2.13 41.58, 11.94, – 13.64 23.76, – 4.36, – 9.45 24.75, – 0.61, – 12.16 26.73, 2.34, – 11.47 33.66, – 4.45, 8.05 29.7, – 6.30, 9.99 32.67, 7.31, 10.23 32.67, 8.37, 12.02 24.75, – 6.25, – 8.52 25.74, – 4.32, – 8.62 22.77, 9.92, – 15.21 25.74, 4.19, – 13.24 – 24.75, – 0.76, 4.18 – 23.76, 6.34, 10.28 – 26.73, 11.09, 8.20 – 19.8, 0.9058, – 1.30 – 49.5, – 3.13, – 23.81 – 50.49, – 1.15, – 23.07 – 33.66, – 23.51, 34.81	Frontal Lobe, Inferior Frontal Gyrus, White Matter Frontal Lobe, Middle Frontal Gyrus, Gray Matter, Brodmann area 11 Frontal Lobe, Extra-Nuclear, Gray Matter, Brodmann area 47 Frontal Lobe, Inferior Frontal Gyrus, Gray Matter, Brodmann area 11 Frontal Lobe, Inferior Frontal Gyrus, Gray Matter, Brodmann area 45 Frontal Lobe, Inferior Frontal Gyrus, Gray Matter, Brodmann area 47 Sub-lobar, Extra-Nuclear, Gray Matter, Brodmann area 13 Sub-lobar, Extra-Nuclear, Gray Matter, Brodmann area 47 Sub-lobar, Insula, Gray Matter, Brodmann area 13 Sub-lobar, Insula, Gray Matter, Brodmann area 47 Temporal Lobe, Superior Temporal Gyrus, Gray Matter, Brodmann area 38 Limbic Lobe, Parahippocampal Gyrus, Gray Matter, Amygdala Limbic Lobe, Parahippocampal Gyrus, Gray Matter, Brodmann area 34 Limbic Lobe, Subcallosal Gyrus, Gray Matter, Brodmann area 34 Sub-lobar, Claustrum, Gray Matter Sub-lobar, Lentiform Nucleus, Gray Matter, Putamen Sub-lobar, Claustrum, Gray Matter Right Cerebrum, Sub-lobar, Insula, Gray Matter, Brodmann area 13 Sub-lobar, Lentiform Nucleus, Gray Matter, Lateral Globus Pallidus Sub-lobar, Lentiform Nucleus, Gray Matter, Putamen Frontal Lobe, Inferior Frontal Gyrus, Gray Matter, Brodmann area 47 Frontal Lobe, Subcallosal Gyrus, Gray Matter, Brodmann area 34 Sub-lobar, Lentiform Nucleus, Gray Matter, Putamen Sub-lobar, Extra-Nuclear, White Matter Sub-lobar, Claustrum, Gray Matter Sub-lobar, Lentiform Nucleus, Gray Matter, Lateral Globus Pallidus Temporal Lobe, Fusiform Gyrus, Gray Matter, Brodmann area 20 Temporal Lobe, Middle Temporal Gyrus, Gray Matter, Brodmann area 21 Parietal Lobe, Postcentral Gyrus, Gray Matter, Brodmann area 2
4 (purple)	3,135	0.57		
3 (blue)	862	0.52		
2 (green)	293	0.58		
1 (red)	7	0.59		

For one randomly selected fold the range of IG value among the remaining 87,416 voxels was between 0.18 and 0.69. The minimum IG of 0.18 was relatively high, indicating that the voxels either contained a good deal of valuable information to separate the classes or are completely irrelevant.

Fig. 3 displays the spatial distribution of the voxels with non-zero IG across all folds.

For one randomly selected fold clustering reduced the 87,416 selected features to 26,228. In total, 978 clusters containing at least one core object exhibiting the maximum number of 6 neighbors were obtained. The largest cluster comprised 3,445 voxels.

Fig. 4a summarizes the cluster statistics with respect to the two most important criteria: the size of the clusters and value of IG. The anatomical locations of the skyline clusters have been highlighted with the same colors in Fig. 4b. Table 4 provides a summary of the clusters including the anatomical location. Due to space limitation, only the anatomical location of the best separating regions with an IG of at least 0.3 for the large clusters 3 to 5 are included in Table 4. The clusters were centered within the medial temporal lobe including the hippocampus, parahippocampus, amygdala, adjacent basal ganglia, the right anterior cingulate gyrus extending towards the prefrontal cortex, left insula, and claustrum (Table 4).

On the basis of the selected clusters, a classification accuracy of 92% with Bayes (sensitivity: 94%, specificity: 89%), 90% with SVM (sensitivity: 97%, specificity: 78%) and 78% with VFI (sensitivity: 66%, specificity: 100%) was obtained (cf. classification task 1 in Table 3).

#### Classification of MCI-AD vs. MCI-MCI

When applying feature selection on the brain images of the group of MCI with respect to conversion between 97.82% and 98.73% of the voxels have an Information Gain of 0.

For one randomly selected fold we obtained 74,680 features with IG greater than zero. The minimum occurring IG was 0.32. Clustering reduced the number of features to 10,775. The selected clusters separated converters and non-converters with high accuracy: 95.83%



**Table 5**  
Clusters MCI-AD vs. MCI-MCI.

Cluster-ID	Size (voxels)	Max IG	Location	Regions
5 (orange)	1,320	0.61	−1.98, 47.87, −5.59	Anterior Lobe, Culmen, Gray Matter
4 (violet)	573	0.62	15.84, −0.27, −5.45	Sub-lobar, Lenticular Nucleus, Gray Matter, Medial Globus Pallidus
			15.84, 1.66, −5.55	Sub-lobar, Lenticular Nucleus, Gray Matter, Lateral Globus Pallidus
			14.85, −7.85, −1.71	Sub-lobar, Extra-Nuclear, White Matter
			19.80, 3.69, −3.97	Sub-lobar, Lenticular Nucleus, Gray Matter, Putamen
3 (blue)	135	0.93	16.83, 14.50, 37.50	Frontal Lobe, Cingulate Gyrus, Gray Matter, Brodmann area 32
			18.81, 15.33, 34.67	Frontal Lobe, Cingulate Gyrus, White Matter
			20.79, 16.34, 35.57	Frontal Lobe, Sub-Gyrus, White Matter
			18.81, 16.26, 33.73	Limbic Lobe, Cingulate Gyrus, White Matter
			14.85, 19.39, 38.18	Limbic Lobe, Sub-Gyrus, White Matter
2 (green)	35	0.93	67.32, −0.67, 6.02	Temporal Lobe, Superior Temporal Gyrus
1 (red)	7	0.93	−27.72, −23.65, 32.04	Frontal Lobe, Sub-Gyrus, White Matter

accuracy was obtained with SVM and VFI (sensitivity SVM: 89%, VFI 100%, specificity SVM 100%, VFI 93%), cf. task 2 in Table 3. With Bayes, an accuracy of 91.67% has been obtained (sensitivity: 78% specificity: 100%). The skyline clusters for the separation of converters and non-converters are displayed in Figs. 5a and b. The corresponding anatomical regions are provided in Table 5. In total, 276 clusters were obtained.

Using AD vs. HC as training data and MCI as test data, best results have been obtained with VFI. Conversion was predicted with an accuracy of 75% (sensitivity: 56%, specificity: 87%).

When contrasting the group of subjects with MCI against the HC for one randomly selected fold a total of 37,504 characteristic features were obtained. Clustering reduced the number of features to 2,190. On the clustered data linear SVM performed best with 97.62% in accuracy (sensitivity: 96%, specificity: 100%), cf. task 3 in Table 3. With VFI we obtained accuracy of 88.1% (sensitivity: 83%, specificity: 94%), and with Bayes an accuracy of 85.71% (sensitivity: 83%, specificity: 88%). The spatial location of the best discriminating clusters was similar to that of MCI-MCI vs. MCI-AD and therefore was not displayed.

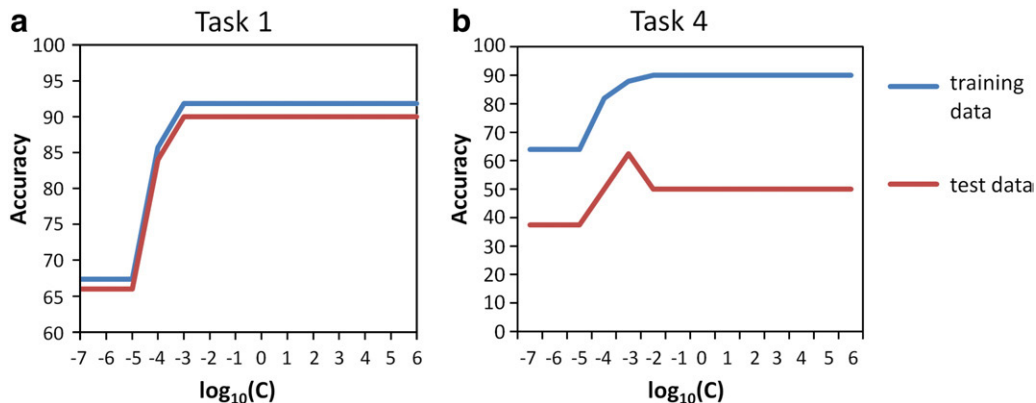
#### Parameter settings for classification

An important parameter for SVM is the complexity constant  $C$ . For the soft margin SVM, the parameter  $C > 0$  determines the trade-off between margin maximization and training error minimization. To systematically investigate the influence of the complexity constant on the classification accuracy, we repeated the experiments with various

settings for  $C$  in the range of  $\log_{10}(C) = -7$  to  $+6$ . The analysis was applied to 2 out of the 4 classification tasks: task 1 which involves the classification of AD vs. HC with leave-one-out cross-validation and task 4 which involves the prediction of conversion in MCI based on the model learned from AD vs. HC with train-and-test validation. We selected those two tasks for two reasons: SVM is outperformed by other classifiers in these settings (see Table 3) and the classification accuracy of SVM varies, depending upon the validation procedure applied in tasks 1 and 4.

The classification result of task 1 for various settings of  $C$  is displayed in Fig. 6(a). Since task 1 included a leave-one-out cross-validation, the accuracy on both training and test data, respectively, were averaged among all folds. As can be seen in Fig. 6, the parameter  $C$  had only very minor influence on the classification accuracy. The classification accuracy for the training data (90%) and the accuracy for the test data (91.84%) were constant for a wide range of parameter settings (of  $\log_{10}(C) = -3$  to 6). Only for a very small  $C$ , the classification accuracies for both the training and test data sets decreased numerically ( $\log_{10}(C) < -3$ ). For  $\log_{10}(C) < -5$  we observed a trend towards a statistically significant decrease in classification accuracy compared to the optimal level, i.e. the accuracy was 66% (95%CI = [52.15, 77.56]) for  $\log_{10}(C) < -5$  vs. 90% (95%CI [77.41, 96.26]) for  $\log_{10}(C) = -3$  to 6 when applied to the test data set.

Fig. 6b displays the result of an analogous analysis for task 4. Note that this is the most difficult classification task since the training and the test data stem from different groups of subjects. Thus, the whole data mining pipeline including feature selection, clustering and training of the SVM is applied to the training data AD vs. HC. Based on



**Fig. 6.** Effect of the parameter  $C$  on the classification accuracy of SVM in task 1 (a) and task 4 (b). For both tasks we can observe only minor influence of  $C$  for very small  $C$  ( $\log_{10}(C) < -3$ ).



the learned model, the SVM predicts the conversion of test subjects with MCI. Fig. 6b displays the accuracy for the training data, i.e. the accuracy of the SVM applied to AD vs. HC for varying settings of the complexity constant  $C$ . In addition, the accuracy for the test data is displayed, i.e. the accuracy to predict conversion in MCI subjects. We can observe two different aspects from Fig. 6b. First, the training data could be well separated using a linear kernel. Since more complex kernels (like polynomial, radial basis, etc.) lead more likely to overfitting and thus lower classification accuracy when applied to the test data, their application is indicated only in the case when the training data cannot be well separated using the linear kernel. Second, visual inspection of Fig. 6b shows that the choice of the complexity parameter  $C$  had only very minor influence on the accuracy. Within the range of  $\log_{10}(C) = -3$  to 6, the classification accuracy for the training data was constantly 90% and the accuracy for the test data was constantly 50%. Similar to the results for task 1 (see above), the classification accuracy for both the training and test data decreased only for very small  $C$  values. There was one single exception from this trend: For  $\log_{10}(C) < -3$ , we observed that the accuracy for the training data decreased from 90% to 88%, while there was a numerical increase of the accuracy from 50% to 62.5% for the test data set. However, the increase of the classification accuracy to 62.5% at  $C = 0.001$  was not significant, as it fell within the 95%CI of the classification accuracy of 50% at  $C = 1.0$  (95%CI [29.65, 70.35]). Note that the parameter setting associated with a numerical increase in prediction accuracy could not be predicted on the basis of the training data, since the accuracy on training data was lower for  $C = 0.001$  (88%) than for  $C = 1$  (90%).

#### White matter rating and SPM voxel based analysis

In order to evaluate whether the current findings may have been influenced by potential white matter damage that could lead to segmentation inaccuracies and thus bias the classification on the basis of grey matter maps, we conducted a regional rating of age-related white matter changes according to a standardized protocol (Wahlund et al., 2001). The mean scores averaged across both brain hemispheres for the basal ganglia, frontal lobe, temporal lobe, parieto-occipital lobe, and infratentorial brain area are displayed in Table 6. There were no group differences across any of the brain regions and the white matter damage was low and clinically non-significant, with the mean rating score being below 1 for each brain area in each group.

To further validate our findings, we conducted a univariate SPM8 based group comparison of the voxel based maps. Compared to the HC group, the AD subjects showed on average grey matter reductions in the basal ganglia, medial frontal gyrus, inferior parietal lobule and precuneus for the AD vs. HC comparison (Supplementary Figure 1). MCI subjects had lower grey matter primarily within the basal ganglia and the medial and superior temporal gyrus (Supplementary Figure 2). Thus, similar brain regions affected in MCI or AD were identified in both the “classical” univariate SPM-based analysis of the deformation maps and the current classification methods, lending further support for the validity of the current findings with the novel pattern recognition method.

**Table 6**

Mean rating scores of age related white matter changes and standard deviation (in brackets) for each group and different brain regions.

Group	Brain region				
	Basal ganglia	Infratentorial area	Frontal lobe	Temporal lobe	Parieto-occipital Lobe
AD	<0.1 (<0.1)	0 (0)	0.5 (0.6)	0.2 (0.4)	0.5 (0.5)
MCI	<0.1 (0.1)	0 (0)	0.6 (0.7)	0.2 (0.4)	0.5 (0.7)
HC	0 (0)	0 (0)	0.3 (0.5)	0 (0)	0.2 (0.4)

## Discussion

In the present study we demonstrated data mining methods to extract AD-typical patterns of brain atrophy using three different classifiers. We classified AD vs. HC, MCI-AD vs. MCI-MCI, and MCI vs. HC with excellent accuracy between 92% and 97.62% based upon leave-one-out validation. For the prediction of conversion from MCI to AD, the best predictive value was achieved with the VFI classifier reaching a predictive accuracy of about 75% validated in a train-and-test setting.

As a proof of concept we first established the AD-specific spatial pattern of atrophy using classifiers for the discrimination between AD and HC. Those brain regions that best discriminated AD from elderly HC included the medial temporal lobe, anterior cingulate gyrus extending towards the orbitofrontal cortex as well as the subcortical thalamic-basal ganglia brain areas, which were reliably identified using leave-one-out cross validation. These results are largely consistent with our previous PCA-based analysis (Teipel et al., 2007a), providing support for the convergent validity across different analysis methods. The pattern of atrophy detected in the current study agrees with findings of a range of previous independent MRI based (Meguro et al., 2001; Scallan et al., 2002) and neuropathological studies showing AD-typical predilection sites of pathological changes (Braak et al., 1997; Price et al., 1991, 2001). The current classification accuracy was high across different classifiers ranging between 78% and 92% after cross-validation.

For the classification of MCI-AD vs. MCI-MCI we obtained excellent results with a classification accuracy ranging from 91.97 to 95.83% with leave-one-out cross-validation. The skyline clusters are roughly a subset of the clusters identified for the separation of AD vs. HC, extending towards the temporal lobe including the superior temporal gyrus.

Using AD and HC as training data and MCI as test data, we achieved an accuracy of 50%–75% to predict conversion into AD. As expected, the performance of all classifiers declines in comparison to leave-one-out-cross-validation, since the test data originate from a data set that is expected to differ in the extent of pathological brain changes from the training data. These results fit with the findings of previous studies. In our previous study in the same patients, using a completely different multivariate approach based on PCA and canonical covariate analysis (Teipel et al., 2007a,b), we showed that the separation between MCI converters and non-converters was not significant, however, applying the feature vector of the AD vs. HC comparison to the MCI data resulted in an accuracy of 73% (Teipel et al., 2007a). Similarly, the group of Davatzikos (Misra et al., 2009) applied the AD vs. HC classifier (accuracy of 94%, (Fan et al., 2008)) to separate MCI converters vs. non-converters, finding a high sensitivity of 85.2% but a very low specificity of 36%, resulting in a classification accuracy of 48.5%, which is similar to the accuracy observed in our experiment (data were derived from the Fig. 6 in (Misra et al., 2009)).

In the current study, Bayes and VFI yielded superior results compared to the SVM approach. The leave-one-out experiments showed that the most selective regions for the discrimination of MCI-AD and MCI-MCI are roughly a subset of the most selective regions for AD vs. HC. Consequently, the training data contain many superfluous features, i.e. regions which are not selective to distinguish MCI-AD from MCI-MCI. VFI performs best on this difficult classification task, probably because this classifier is by design most robust with respect to superfluous information. The votes of these features approximately sum up to zero and thus have only minor effect on the classification accuracy. There is much more chance that random variations of the superfluous features cause overfitting in SVM.

Our pattern recognition technique detected structural changes within the anterior cingulate gyrus, the hippocampus and the basal ganglia in MCI converters, consistent with the previously found pattern of atrophy in MCI compared to HC (Pennanen et al., 2005). The

major brain regions that separated best between HC and AD included primarily the prefrontal cortex especially the inferior and middle frontal gyri, the hippocampal region and adjacent subcortical basal ganglia, as well as more posterior brain regions within the parietal lobe. The hippocampal, frontal and parietal brain regions are well documented to be affected in AD. We detected also significant changes within the subcortical brain regions of AD. These AD-specific changes cannot be accounted for by differences in WMH, since rating of the WMH on the basis of FLAIR scans showed no group differences and the severity of WMH was in general low with the mean score always being lower than 1 across the different brain regions and diagnostic groups. The validity of the current findings is further supported by the convergent results between the univariate voxel-based analysis based on SPM and the current pattern recognition method. Furthermore, the findings on the grey matter changes within the basal ganglia are consistent with our previous results in the same patients with the PCA based approach where the component that separated best between AD and HC was strongly associated with reduced volume of subcortical brain areas including the thalamus and caudate nucleus (Teipel et al., 2007a). Thus, there is considerable overlap between different methodological approaches with regard to the detection of brain regions altered in an AD specific way. Previous independent studies have shown that considerable atrophic progression is found in subcallosal basal ganglia brain structures (Ferrarini et al., 2006; Karas et al., 2003, 2004), and was demonstrated to show one of the fastest atrophy rates within the brain of AD patients (>15% per year, (Thompson et al., 2003)). Although there is strong evidence of atrophy within these brain regions in AD, less attention may have been spent on the basal ganglia, since the cognitive function of these brain areas is not well known. A recent study, however, that detected strong atrophy within the thalamus and putamen of AD patients showed an association with global cognitive performance and executive functions independently from hippocampus grey matter atrophy (de Jong et al., 2008). Thus, the basal ganglia structures may show pronounced volume reductions in AD.

We aimed to render our analysis especially sensitive towards the detection of subtle brain abnormalities by employing a non-linear supervised feature selection method that is less dependent upon sample-size restrained power due to cross-validation and train-and-test than previous analysis for dimensionality reduction (Davatzikos et al., 2008; Fan et al., 2008; Teipel et al., 2006). As an alternative to feature selection, dimensionality reduction can be achieved for example by principal component analysis and subsequently rated for class separation, using MANCOVA (Teipel et al., 2006, 2007a,b). However, these methods depend entirely upon data-driven transformations and thus do not reduce variability in an informed way. There exist few supervised versions of singular value decomposition (SVD) and independent component analysis (ICA, e.g. (Bair et al., 2006; Sakaguchi et al., 2002)) which consider the class labels during feature transformation. However, the results of these methods are difficult to interpret, since the amount of supervision is typically controlled by parameter settings. In contrast, the result of supervised feature selection is very intuitive because the interesting voxels are selected in the original image space. We decided to use the Information Gain as feature selection criterion, because it provides 1) a very general rating of the discriminatory power, 2) is highly efficient to compute and 3) has been successfully applied in a large variety of applications, e.g. in information retrieval (Mori, 2003), object recognition (Cooper and Miller, 1998) and bioinformatics (Yang et al., 2003; Plant et al., 2006). Correlation-based feature selection criteria, e.g. based on Pearson correlation, are closely related; however, the Information Gain is not restricted to linear correlations but captures any form of dependency between features and class labels. The applied feature selection technique can generally be used together with a wide variety of classifiers. We selected three classifiers which represent different

algorithmic paradigms and therefore provide a comprehensive evaluation of the discriminatory power of the selected features.

The majority of previous studies described characteristically altered brain areas in AD or MCI on a group-level (Apostolova et al., 2007; Bozzali et al., 2006; Carlson et al., 2007; Davatzikos et al., 2001). However, the diagnostic value of group-level analysis is limited. Some studies used multivariate methods which provide the potential to draw conclusions on a single-subject level but these papers do not report validated classification results (Chen and Herskovits, 2006). Recent studies reported validated classification results for the identification of AD. Duchesne et al. (2008b) proposed to apply a support vector machine classifier based on least squares optimization on a selected volume of interest consisting of Jacobian determinants resulting from spatial normalization within the temporal lobe. In contrast, we used the whole images of the subjects as single source for feature selection and classification.

Fan et al. (2007) present an approach for identification of schizophrenia relying on deformation-based morphometry and machine learning. They achieve high classification accuracy (91.8% for female subjects and 90.8% for male subjects). This approach is conceptually similar to ours since it also applies feature selection and watershed segmentation which can be regarded as some kind of clustering, before performing classification with SVM. For each voxel, a score is computed by linearly combining the discriminatory power for classification as measured by Pearson-moment correlation with the aspect of spatial consistency which is measured by intra-class correlation. Using a similar approach, Davatzikos et al., 2008 report an accuracy of 90% for the identification of MCI in a leave-one-out validation setting. Our approach also emphasizes both aspects, the discriminatory power and the spatial coherency. However, very different definitions are applied to formalize these concepts. The discriminatory power is defined by the Information Gain, with the benefit to allow for arbitrary and not only linear correlations with the class label. Spatial coherency is achieved by density-based clustering which refines the selected features to form coherent regions. The result of watershed segmentation strongly depends on suitable selection of the thresholds which is very difficult especially in the presence of noise (Gerig et al., 1992). By the application of a modified density-based clustering technique our approach allows identifying the best discriminating brain regions without requiring any parameter settings or thresholds which are difficult to estimate. Vemuri et al. (2008, 2009a,b) showed that the predictive accuracy of SVM identified brain changes can be augmented by including other biomarkers or clinical information for the detection of AD. The highest accuracy for the AD identification was obtained when combining these imaging features with covariates, including demographic information and the apolipoprotein E genotype or cerebrospinal fluid related biomarkers.

The current study had caveats that should be taken into account for the interpretation of the results. One factor to bear in mind relates to censoring effects. Particularly at shorter follow-up intervals, censoring effects are likely to increase the number of seemingly false positives, as MCI patients with a pathologic pattern in MRI may not yet have developed clinical AD during follow-up.

It should be noted that the current study is based on a limited number of patients. In order to validate the utility of the current classifiers further application to a larger multicenter data set is necessary. In smaller samples the variability of the classification accuracy based upon the classifiers may be larger and thus less reliable (see Frost and Kallis, 2009 correspondence to Kloeppel et al., 2008, 2009). The robustness of the results and potential influence of outliers has been tested in the current study by the leave-one-out validation, but may still need further testing in larger data sets. Another caveat of the current study is that the HC group was younger when compared to the MCI group. This age difference may have influenced the results. We showed, however, previously in the same data set that age and gender were not significant predictors for the

group separation based upon a PCA scores (Teipel et al., 2007a). Furthermore the focus was on distinguishing between MCI converters and non-converters who did not differ in age in the current study.

Concerning the parameter settings for classification, for SVM there are two parameter choices which may have an impact on the classification result, the choice of the kernel and the choice of the complexity constant  $C$ . Due to the high dimensionality of the solution space (i.e. the high number of variables) it is indicated to use a linear kernel. Other more complex kernel functions (such as polynomial, Gaussian, radial basis, etc.) are known to be subject to overfitting in presence of very high-dimensional spaces, i.e. good separation of the training data but deteriorated accuracy of the final classification result after validation. Therefore, more complex kernels should be used only if the training data are not well separable using the linear kernel. The results of the second tunable parameter of the SVM, the complexity constant  $C$ , show that the variation of this parameter within a wide range of values did not lead to any significant differences in the classification accuracy with either the leave one-out paradigm or within the training-test validation scheme. This is probably due to the fact that in all our experiments the training data are sufficiently separable by SVM using a linear kernel. In this case, the trade-off between margin maximization and training error minimization is of minor relevance in the optimization problem solved by SVM. Our results are consistent with those of LaConte et al who observed that the parameter  $C$  has no influence on SVM as applied to fMRI data, unless the  $C$  value is very small ( $C=0.001$ ) (LaConte et al., 2005). Vemuri et al. (2008) observed some influence of the parameter  $C$  on the classification result. For the classification of AD vs. HC based on MRI data only, best results with 85.8% in accuracy have been obtained using  $C=0.01$ . With our method we achieved a classification accuracy of 90% for this task, independent of the selection of the parameter  $C$  between 0.001 and 1,000,000. Let us note that these findings are not directly comparable for several reasons: First, the study of Vemuri et al. is based on a larger collective involving 140 subjects with AD and 140 healthy controls and a different validation strategy has been applied. In our study, we applied leave-one-out cross-validation whereas Vemuri et al. applied four-fold cross validation.

In conclusion, we showed a novel approach to identify regions of high discriminatory power for the identification of AD and the prediction of conversion to AD among MCI. Our method combines data mining techniques from feature selection, clustering and classification and provides a concise visualization of the most selective regions in the original native image space. In future work we plan to apply our framework in a large multi-centre study. The study of Kloeppel et al. (2008) demonstrated the potential of SVM classification for the identification of AD in a multi-centre setting. Applying our data mining framework to a larger sample size we expect further validation of the classification results. In addition, we expect to confirm the best discriminating regions in this sample and complement them by novel findings.

## Acknowledgments

The authors thank B. Asam, F. Jancu, L. Jertila-Aqil, and C. Sanger for technical assistance.

The study was supported by a grant from the Federal Agency of Education and Research (Bundesministerium fuer Bildung und Forschung, BMBF 01 GI 0102) to the Competence Network of Dementia (to HH, ME, and SJT), grants from Adelaide and Meath Hospital incorporating the National Children's Hospital (AMNCH) (to HH), the Health Service Executive (HSE) (to HH), Trinity College Dublin, Ireland (to HH), the Science Foundation Ireland (SFI) as part of the SFI Stokes Programme (to ALWB), a grant from the Hirnliga Foundation, Germany (to SJT), a grant from the German Center on Neurodegenerative Disorders (DZNE) within the Helmholtz Society, Germany (to SJT), Wellcome Trust (JMM).

## Appendix A

### Linear Support Vector Machine (SVM) (Platt, 1998)

As a function-based classifier, the Linear Support Vector Machine aims at constructing a hyperplane separating the training examples. Originally, the Support Vector Machine has been designed for two class problems and therefore we restrict ourselves to this case. For data sets of dimensionality  $d$  there are often many possible separating hyperplanes of dimensionality  $d-1$ , especially for high dimensional data. To predict the class of unknown objects with highest possible accuracy and to avoid over fitting, among all separating hyperplanes the one maximizing the margin between the training examples of both classes needs to be determined. Formally, a separating hyperplane can be defined by:

$$\langle w, s_i \rangle + b = 0$$

The class label of each subject  $s_i$  is determined by the signum function of the separating hyperplane. The location of the hyperplane is described by the vector  $w$  which is perpendicular to the plane and the bias  $b$  which specifies its shift from the origin of the coordinate system. To find the hyperplane providing the largest margin between both classes, only the closest instances to the plane at both sides are of interest, the so-called support vectors. If the classes are linearly separable, the maximum margin hyperplane is determined by parallel hyperplanes passing through the support vectors with maximum distance from each other (cf. Fig. 2). Since the distance between those hyperplanes equals  $2/\|w\|$ , selecting the largest margin hyperplane means minimizing  $\|w\|$  subject to the constraint of a correct classification of the training examples. This so-called primal optimization problem can be efficiently solved by dynamic programming. The optimization problem can be rewritten by expressing  $w$  in terms of scalar products of the support vectors. In this dual form kernel functions can be applied if the data is not linearly separable in the original space. An extension is the soft margin support vector machine which allows misclassified instances within the margin to counteract over fitting. For soft margin classification, there is a trade-off between minimizing  $\|w\|$  and the number of misclassified instances, i.e. between margin maximization and training error minimization. This trade-off is controlled by a parameter, the so-called complexity constant  $C$ .

### Bayesian Classifier (Bayes) (John and Langley 1995)

Bayesian classification relies on the assumption that each feature (in our application each voxel) follows a probability density function, in most approaches a Gaussian distribution is assumed. Each class can thus be characterized by a potentially different mixture model of  $d$  probability density functions. Classification is performed by assigning the object to the most probable class, i.e.

$$s_i.c = \arg \max_{c_i \in C} P(v_1, \dots, v_d | c_i) \cdot P(c_i).$$

The probability of each class  $P(c_i)$  can be interfered easily from the training data. However, it is in most applications impossible to estimate the conditional probability  $P(v_1, \dots, v_d | c_i)$ , since for each class several instances  $V = \{v_1, \dots, v_d\}$  would be needed. Therefore, the Naïve Bayesian classifier relies on the simplifying assumption that the single features are independent of each other, i.e.  $P(v_1, \dots, v_d | c_i) = \prod_{j=1}^d P(v_j | c_i)$ . The decision rule is simplified to

$$s_i.c = \arg \max_{c_i \in C} P(c_i) \cdot \prod_{j=1}^d P(v_j | c_i).$$



See Fig. 2 for an example of two classes which are modeled by Gaussian distributions. In spite of the fact that the assumption of independence does not hold in many applications including MRI data (neighboring voxels are usually highly correlated), Naïve Bayesian classifiers often show good predictive performance. The Bayesian classifier used in this study extends Naïve Bayesian classification by the application of Parzen Windows with Gaussian Kernel to estimate the distributions of continuous attributes (John and Langley, 1995). The derived distributions of the features are thus not restricted to be Gaussian which has been demonstrated to improve the performance of Bayesian classification on many real-world data sets.

#### Classification by voting feature intervals (VFI) (Demiroz and Guvenir, 1997)

This simple entropy-based classifier constructs intervals for each class and each feature and records class counts. Classification is performed by voting. During the training phase the intervals, also called concepts are constructed as follows:

For each of the  $d$  features (i.e. voxels  $v_1, \dots, v_d$ ) and for each of the  $k$  classes  $c_1, \dots, c_k$  the maximum and the minimum value of  $v_i$  in class  $c_j$  is determined. The list of  $2k$  end points is sorted and each pair of consecutive points represents an interval. Each interval can be represented as a vector  $\langle \text{lower}, \text{count}_1, \dots, \text{count}_k \rangle$  where *lower* denotes the lower bound and  $\text{count}_1, \dots, \text{count}_k$  the number of subjects of each class having an intensity value of voxel  $v_i$  within the interval. An example interval with the starting point  $y_2$  containing 9 subjects of one class and 6 subjects of the other class is visualized in Fig. 2.

To classify a subject  $s$ , for all  $d$  voxels the intervals in which they fall are determined. For each interval  $I$  and each class  $c_i$  a vote is computed as follows:

$$\text{vote}_{I, c_i, v_j} = \frac{\text{IntervalClassCount}(c_i)}{|c_i|},$$

where  $\text{IntervalClassCount}(c_i)$  denotes the number of subjects of class  $c_i$  which have an intensity of voxel  $v_j$  within the interval  $I$ . The votes are scaled between 0 and 1 and the final class prediction is computed by summing up all votes.

#### Appendix B. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.neuroimage.2009.11.046.

#### References

Apostolova, L.G., Steiner, C.A., Akopyan, G.G., Dutton, R.A., Hayashi, K.M., Toga, A.W., et al., 2007. Three-dimensional gray matter atrophy mapping in mild cognitive impairment and mild Alzheimer disease. *Arch. Neurol.* 64, 1489–1495.

Ashburner, J., Friston, K., 1997. Multimodal image coregistration and partitioning—a unified framework. *Neuroimage* 6, 209–217.

Ashburner, J., Friston, K.J., 2000. Voxel-based morphometry—the methods. *Neuroimage* 11, 805–821.

Ashburner, J., Neelin, P., Collins, D.L., Evans, A., Friston, K., 1997. Incorporating prior knowledge into image registration. *Neuroimage* 6, 344–352.

Ashburner, J., Andersson, J.L., Friston, K.J., 1999. High-dimensional image registration using symmetric priors. *Neuroimage* 9, 619–628.

Bair, E., Hastie, T., Paul, D., Tibshirani, R., 2006. Prediction by supervised principal components. *J. Amer. Stat. Ass.* 101 (473), 119–137.

Barndorff-Nielsen, O., Sobel, M., 1966. On the distribution of the number of admissible points in a vector random sample. *Theory Probab. Appl.* 11 (2), 249–269.

Baron, J.C., Chetelat, G., Desgranges, B., Perche, G., Landeau, B., de la Sayette, V., Eustache, F., 2001. In vivo mapping of gray matter loss with voxel-based morphometry in mild AD. *Neuroimage* 14, 298–309.

Blennow, K., Hampel, H., 2003. CSF markers for incipient AD. *Lancet Neurol.* 2, 605–613.

Börzsönyi, S., Kossman, D., Stocker, K., 2001. The skyline operator. *Proc. of Int. Conf. on Data Engineering*, pp. 421–430. ICDE 2001.

Bozzali, M., Filippi, M., Magnani, G., Cercignani, M., Franceschi, M., Schiatti, E., et al., 2006. The contribution of voxel-based morphometry in staging patients with mild cognitive impairment. *Neurology* 67, 453–460.

Braak, H., Griffling, K., Braak, E., 1997. Neuroanatomy of Alzheimer's disease. *Alzheimer's Research* 3, 235–247.

Carlson, N.E., Moore, M.M., Dame, A., Howieson, D., Silbert, L.E., Quinn, J.F., Kaye, J.A., 2007. Trajectories of brain loss in aging and the development of cognitive impairment. *Neurology* 70, 828–833.

Chen, R., Herskovits, E.H., 2006. Network analysis of mild cognitive impairment. *Neuroimage* 29 (4), 1252–1259.

Chetelat, G., Desgranges, B., De La Sayette, V., Viader, F., Eustache, F., Baron, J.C., 2002. Mapping gray matter loss with voxel-based morphometry in mild cognitive impairment. *Neuroreport* 13, 1939–1943.

Chetelat, G., Landeau, B., Eustache, F., Mezenge, F., Viader, F., de la Sayette, V., Desgranges, B., Baron, J.C., 2005. Using voxel-based morphometry to map the structural changes associated with rapid conversion in MCI: a longitudinal MRI study. *Neuroimage* 27, 934–946.

Cooper, M.L., Miller, M.L., 1998. Information measures for object recognition. *SPIE Proc. 3370: Algorithms for Synthetic Aperture Radar Imagery V*, pp. 637–645.

Csernansky, J.G., Wang, L., Swank, J., Miller, J.P., Gado, M., McKeel, D., et al., 2005. Preclinical detection of AD: hippocampal shape and volume predict dementia onset in the elderly. *Neuroimage* 25, 783–792.

Davatzikos, C., Genc, A., Xu, D., Resnick, S.M., 2001. Voxel-based morphometry using the RAVENS maps: methods and validation using simulated longitudinal atrophy. *Neuroimage* 14, 1361–1369.

Davatzikos, C., Fan, Y., Wu, X., Shen, D., Resnick, S.M., 2008. Detection of prodromal Alzheimer's disease via pattern classification of magnetic resonance imaging. *Neurobiol. Aging* 29 (4), 514–523.

DeCarli, C., Murphy, D.G.M., McIntosh, A.R., Teichberg, D., Schapiro, M.B., Horwitz, B., 1995. Discriminant analysis of MRI measures as a method to determine the presence of dementia of the Alzheimer type. *Psychiatry Res.* 57 (2), 119–130.

de Jong, L.W., van der Hiele, K., Veer, I.M., Houwing, J.J., Westendorp, R.G.J., Bollen, E.L.E. M., de Bruin, P.V.W., Middelkoop, H.A.M., van Buchem, M.A., van der Grond, J., 2008. Strongly reduced volumes of putamen and thalamus in Alzheimer's disease: an MRI study. *Brain* 131 (12), 3277–3285.

Demiroz, G., Guvenir, A., 1997. Classification by voting feature intervals. *Proc. 9th European Conference on Machine Learning*, pp. 85–92.

Dubois, B., Feldman, H.H., Jacova, C., Dekosky, S.T., Barberger-Gateau, P., Cummings, J., Delacourte, A., Galasko, D., Gauthier, S., Jicha, G., Meguro, K., O'Brien, J., Pasquier, F., Robert, P., Rossor, M., Salloway, S., Stern, Y., Visser, P.J., Scheltens, P., 2007. Research criteria for the diagnosis of Alzheimer's disease: revising the NINCDS-ADRDA criteria. *Lancet Neurol.* 6 (8), 734–746.

Duchesne, S., Bock, C., De Sousa, K., Frisoni, G.B., Chertkow, H., Collins, D.L., 2008a. Amnesic MCI future clinical status prediction using baseline MRI features. *Neurobiol. Aging*, in Press, Corrected Proof.

Duchesne, S., Caroli, A., Geroldi, C., Barillot, C., Frisoni, G.B., Collins, D.L., 2008b. MRI-based automated computer classification of probable AD versus normal controls. *IEEE Trans. Med. Imaging* 27, 509–520.

Duchesne, S., Caroli, A., Geroldi, C., Collins, D.L., Frisoni, G.B., 2009. Relating one-year cognitive change in mild cognitive impairment to baseline MRI features. *Neuroimage* 47 (4), 1363–1370.

Duncan, J., Seitz, R.J., Kolodny, J., Bor, D., Herzog, H., Ahmed, A., Newell, F.N., Emslie, H., 2000. A neural basis for General Intelligence. *Science* 289 (5478), 457–460.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X., 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. *Proc. 2nd Int. Conf. on Knowledge Discovery and Data Mining*, pp. 226–231.

Ewers, M., Buerger, K., Teipel, S.J., Scheltens, P., Schroder, J., Zinkowski, R.P., et al., 2007. Multicenter assessment of CSF-phosphorylated tau for the prediction of conversion of MCI. *Neurology* 69, 2205–2212.

Fan, Y., Shen, D., Gur, R.C., Gur, R.E., Davatzikos, C., 2007. COMPARE: classification of morphological patterns using adaptive regional elements. *IEEE Trans. Med. Imaging* 26 (1), 93–105.

Fan, Y., Batmanghelich, N., Clark, C.M., Davatzikos, C., 2008. Spatial patterns of brain atrophy in MCI patients, identified via high-dimensional pattern classification, predict subsequent cognitive decline. *Neuroimage* 39 (4), 1731–1743.

Fayyad, U.M., Irani, K.B., 1993. Multiinterval discretisation of continuous-valued attributes. *Proc. 13th Int'l Joint Conf. Artificial Intelligence*, pp. 1022–1027.

Ferrarini, L., Palm, W.M., Olofsen, H., van Buchem, M.A., Reiber, J.H., Admiraal-Behloul, F., 2006. Shape differences of the brain ventricles in Alzheimer's disease. *Neuroimage* 32 (3), 1060–1069.

Ferri, C.P., Prince, M., Brayne, C., Brodaty, H., Fratiglioni, L., Ganguli, M., Hall, K., Hasegawa, K., Hendrie, H., Huang, Y., Jorm, A., Mathers, C., Menezes, P.R., Rimmer, E., Sczufka, M., 2005. Global prevalence of dementia, a Delphi consensus study. *Lancet* 366 (9503), 2112–2117.

Folstein, M.F., Folstein, S.E., McHugh, P.R., 1975. Mini-mental-state: a practical method for grading the cognitive state of patients for the clinician. *J. Psychiatr. Res.* 12, 189–198.

Frisoni, G.B., Testa, C., Zorzan, A., Sabatelli, F., Beltramello, A., Soininen, H., Laakso, M.P., 2002. Detection of grey matter loss in mild AD with voxel based morphometry. *J. Neurol. Neurosurg. Psychiatry* 73, 657–664.

Friston, K., Poline, J.-P., Holmes, C.J., Frith, C.D., Frackowiak, R.S., 1996. A multivariate Analysis of PET activation studies. *Hum. Brain Mapp.* 4, 140–151.

Frost, C., Kallis, C., 2009. Reply: a plea for confidence intervals and consideration of generalizability in diagnostic studies. *Brain* 131, e103.

Gerig, G., Kubler, O., Kikinis, R., Jolesz, F.A., 1992. Nonlinear anisotropic filtering of MRI data. *IEEE Trans. Med. Imaging* 11 (2), 221–232.

Good, C.D., Johnsrude, I.S., Ashburner, J., Henson, R.N., Friston, K.J., Frackowiak, R.S., 2001. A voxel-based morphometric study of ageing in 465 normal adult human brains. *Neuroimage* 14, 21–36.



- Hall, M., Holmes, J., 2003. Benchmarking attribute selection techniques for discrete class data mining. *IEEE Trans. Knowl. Data Eng.* 15 (6), 1437–1447.
- Hansson, O., Zetterberg, H., Buchhave, P., Londos, E., Blennow, K., Minthon, L., 2006. Association between CSF biomarkers and incipient AD in patients with mild cognitive impairment: a follow-up study. *Lancet Neurol.* 5, 228–234.
- Herukka, S.K., Helisalmi, S., Hallikainen, M., Tervo, S., Soininen, H., Pirttilä, T., 2007. CSF Aβ42, Tau and phosphorylated Tau, APOE ε4 allele and MCI type in progressive MCI. *Neurobiol. Aging* 28, 507–514.
- Hristidis, V., Koudas, N., Papakonstantinou, Y., 2001. PREFER: A system for the efficient execution of multi-parametric ranked queries. *Proc. ACM SIGMOD International Conference on Management of Data*, pp. 259–270.
- Jack Jr., C.R., Petersen, R.C., Xu, Y.C., O'Brien, P.C., Smith, G.E., Ivnik, R.J., Boeve, B.F., Waring, S.C., Tangalos, E.G., Kokmen, E., 1999. Prediction of AD with MRI-based hippocampal volume in mild cognitive impairment. *Neurology* 52 (7), 1397–1403.
- John, G.H., Langley, P., 1995. Estimating continuous distributions in bayesian classifiers. *Proc. 11th Conference on Uncertainty in Artificial Intelligence*, pp. 338–345.
- Kantarci, K., Jack Jr., C.R., 2003. Neuroimaging in Alzheimer disease: an evidence-based review. *Neuroimaging Clin. N. Am.* 13, 197–209.
- Karas, G.B., Burton, E.J., Rombouts, S.A., van Schijndel, R.A., O'Brien, J.T., Scheltens, P., McKeith, I.G., Williams, D., Ballard, C., Barkhof, F., 2003. A comprehensive study of gray matter loss in patients with Alzheimer's disease using optimized voxel-based morphometry. *Neuroimage* 18 (4), 895–907.
- Karas, G.B., Scheltens, P., Rombouts, S.A., Visser, P.J., van Schijndel, R.A., Fox, N.C., Barkhof, F., 2004. Global and local gray matter loss in mild cognitive impairment and Alzheimer's disease. *Neuroimage* 23 (2), 708–716.
- Kearns, M.J., Ron, D., 1997. Algorithmic stability and sanity-check bounds for leave-one-out cross-validation. *Proc. 10th Annual Conference on Computational Learning Theory*, pp. 152–162.
- Killiany, R.J., Hyman, B.T., Gomez-Isla, T., Moss, M.B., Kikinis, R., Jolesz, F., Tanzi, R., Jones, K., Albert, M.S., 2002. MRI measures of entorhinal cortex vs hippocampus in preclinical AD. *Neurology* 58, 1188–1196.
- Kloppel, S., Stonnington, C.M., Chu, C., Draganski, B., Scahill, R.I., Rohrer, J.D., Fox, N.C., Jack Jr., C.R., Ashburner, J., Frackowiak, R.S., 2008. Automatic classification of MR scans in Alzheimer's disease. *Brain* 131, 681–689.
- Kloppel, S., Stonnington, C.M., Chu, C., Draganski, B., Scahill, R.I., Rohrer, J.D., Fox, N.C., Jack Jr., C.R., Ashburner, J., Frackowiak, R.S., 2009. A plea for confidence intervals and consideration of generalizability in diagnostic studies. *Brain* 132 (4), e102.
- LaConte, S., Strother, S., Cherkassky, V., Hu, X., 2005. Support vector machines for temporal classification of block design fMRI data. *Neuroimage* 26, 317–329.
- McIntosh, A.R., Grady, C.L., Ungerleider, L.G., Haxby, J.V., Rapoport, S.I., Horwitz, B., 1994. Network analysis of cortical visual pathways mapped with PET. *J. Neurosci.* 14, 655–666.
- McKeown, M.J., Makeig, S., Brown, G.G., Jung, T.P., Kindermann, S.S., Bell, A.J., Sejnowski, T.J., 1998. Analysis of fMRI data by blind separation into independent spatial components. *Hum. Brain Mapp.* 6 (3), 160–188.
- McKhann, G., Drachman, D., Folstein, M., Katzman, R., Price, D., Stadlan, E.M., 1984. Clinical diagnosis of AD: report of the NINCDS-ADRDA Work Group under the auspices of the Department of Health and Human Services Task Force on AD. *Neurology* 34, 939–944.
- Meguro, K., LeMestric, C., Landeau, B., Desgranges, B., Eustache, F., Baron, J.C., 2001. Relations between hypometabolism in the posterior association neocortex and hippocampal atrophy in AD: a PET/MRI correlative study. *J. Neurol. Neurosurg. Psychiatry* 71, 315–321.
- Misra, C., Fan, Y., Davatzikos, C., 2009. Baseline and longitudinal patterns of brain atrophy in MCI patients, and their use in prediction of short-term conversion to AD: results from ADNI. *Neuroimage* 44 (4), 1415–1422.
- Mori, T., 2003. A term weighting method based on information gain ratio for summarizing documents retrieved by IR systems. *Journal of Natural Language Processing* 9 (4), 3–32.
- Morris, J.C., Heyman, A., Mohs, R.C., Hughes, J.P., van Belle, G., Fillenbaum, G., Mellits, E. D., Clark, C., 1989. The consortium to establish a registry for Alzheimer's disease (CERAD). Part I. Clinical and neuropsychological assessment of Alzheimer's disease. *Neurology* 39, 1159–1165.
- Mourao-Miranda, J., Bokde, A.L., Born, C., Hampel, H., Stetter, M., 2005. Classifying brain states and determining the discriminating activation patterns: Support Vector Machine on functional MRI data. *NeuroImage* 28 (4), 980–995.
- Mourao-Miranda, J., Reynaud, E., McGlone, F., Calvert, G., Brammer, M., 2006. The impact of temporal compression and space selection on SVM analysis of single-subject and multi-subject fMRI data. *NeuroImage* 33 (4), 1055–1065.
- Newcombe, R.G., 1998. Two-sided confidence intervals for the single proportion: comparison of seven methods. *Stat. Med.* 17, 857–872.
- Pennanen, C., Kivipelto, M., Tuomainen, S., Hartikainen, P., Hanninen, T., Laakso, M.P., Hallikainen, M., Vanhanen, M., Nissinen, A., Helkala, E.L., Vainio, P., Vanninen, R., Partanen, K., Soininen, H., 2004. Hippocampus and entorhinal cortex in mild cognitive impairment and early AD. *Neurobiol. Aging* 25, 303–310.
- Pennanen, C., Kivipelto, M., Tuomainen, S., Hartikainen, P., Hanninen, T., Laakso, M.P., Hallikainen, M., Vanhanen, M., Nissinen, A., Helkala, E.L., Vainio, P., Vanninen, R., Kivipelto, M., Nononen, M., Nissinen, A., Tervo, S., Vanhanen, M., Vanninen, R., Frisoni, G.B., Soininen, H., 2005. A voxel based morphometry study on mild cognitive impairment. *J. Neurol. Neurosurg. Psychiatry* 76, 11–14.
- Petersen, R.C., Doody, R., Kurz, A., Mohs, R.C., Morris, J.C., Rabins, P.V., Ritchie, K., Rosser, M., Thal, L., Winblad, B., 2001. Current concepts in mild cognitive impairment. *Arch. Neurol.* 58, 1985–1992.
- Plant, C., Osl, M., Tilg, B., Baumgartner, B., 2006. Feature selection on high throughput SELDI-TOF mass-spectrometry data for identifying biomarker candidates in ovarian and prostate cancer. *Proc. of 6th IEEE Int. Conf. on Data Mining, Bioinformatics Workshop*, pp. 174–179.
- Platt, J., 1998. Machines using Sequential Minimal Optimization. *Advances in Kernel Methods - Support Vector Learning*.
- Price, J.L., Davis, P.B., Morris, J.C., White, D.L., 1991. The distribution of tangles, plaques and related immunohistochemical markers in healthy aging and Alzheimer's disease. *Neurobiol. Aging* 12, 295–312.
- Price, J.L., Ko, A.L., Wade, M.J., Tsou, S.K., McKeel, D.W., Morris, J.C., 2001. Neuron number in the entorhinal cortex and CA1 in preclinical Alzheimer disease. *Arch. Neurol.* 58, 1395–1402.
- Quinlan, J.R., 1993. C4.5: Programs for Machine Learning. Morgan Kaufmann.
- Sakaguchi, Y., Ozawa, S., Kotani, M., 2002. Feature extraction using supervised independent component analysis by maximizing class distance. *Proc. of the 9th International Conference on Neural Information Processing*, pp. 2502–2506.
- Scahill, R.I., Schott, J.M., Stevens, J.M., Rosser, M.N., Fox, N.C., 2002. Mapping the evolution of regional atrophy in AD: unbiased analysis of fluid-registered serial MRI. *Proc. Natl. Acad. Sci. USA* 99, 4703–4707.
- Skoutas, D., Sacharidis, D., Simitsis, A., Sellis, T.K., 2008. Serving the Sky: discovering and selecting semantic web services through dynamic skyline queries. *Proc. 2nd IEEE International Conference on Semantic Computing (ICSC)*, pp. 222–229.
- Stoub, T.R., Bulgakova, M., Leurgans, S., Bennett, D.A., Fleischman, D., Turner, D.A., deToledo-Morrell, L., 2005. MRI predictors of risk of incident Alzheimer disease: a longitudinal study. *Neurology* 64, 1520–1524.
- Teipel, S., Ewers, M., Dietrich, O., Schoenberg, S., Jessen, F., Heun, R., Freymann, N., Moller, H.J., Hampel, H., 2006. Reliability of multicenter magnetic resonance imaging. Results of a phantom test and in vivo measurements by the German Dementia Competence Network. *Nervenarzt* 77 (9), 1086–92, 1094–5.
- Teipel, S.J., Born, C., Ewers, M., Bokde, A.L., Reiser, M.F., Moller, H.J., Hampel, H., 2007a. Multivariate deformation-based analysis of brain atrophy to predict Alzheimer's disease in mild cognitive impairment. *Neuroimage* 38 (1), 13–24.
- Teipel, S.J., Stahl, R., Dietrich, O., Schoenberg, S.O., Perneczky, R., Bokde, A.L., Reiser, M.F., Moller, H.J., Hampel, H., 2007b. Multivariate network analysis of fiber tract integrity in Alzheimer's disease. *Neuroimage* 34 (3), 985–995.
- Thompson, P.M., Hayashi, K.M., de Zubicaray, G., Janke, A.L., Rose, S.E., Semple, J., Herman, D., Hong, M.S., Dittmer, S.S., Doddrell, D.M., Toga, A.W., 2003. Dynamics of gray matter loss in Alzheimer's disease. *J. Neurosci.* 23 (3), 994–1005.
- Vemuri, P., Gunter, J.L., Senjem, M.L., Whitwell, J.L., Kantarci, K., Knopman, D.S., Boeve, B.F., Petersen, R.C., Jack Jr., C.R., 2008. Alzheimer's disease diagnosis in individual subjects using structural MR images: validation studies. *NeuroImage* 39 (3), 1186–1197.
- Vemuri, P., Wiste, H.J., Weigand, S.D., Shaw, L.M., Trojanowski, J.Q., Weiner, M.W., Knopman, D.S., Petersen, R.C., Jack Jr., C.R., 2009a. MRI and CSF biomarkers in normal, MCI, and AD subjects: diagnostic discrimination and cognitive correlations. *Neurology* 73, 287–293.
- Vemuri, P., Wiste, H.J., Weigand, S.D., Shaw, L.M., Trojanowski, J.Q., Weiner, M.W., Knopman, D.S., Petersen, R.C., Jack Jr., C.R., 2009b. MRI and CSF biomarkers in normal, MCI, and AD subjects: predicting future clinical change. *Neurology* 73, 294–301.
- Visser, P.J., Scheltens, P., Verhey, F.R.J., Schmand, B., Launer, L.J., Jolles, J., Jonker, C., 1999. Medial temporal lobe atrophy and memory dysfunction as predictors for dementia in subjects with mild cognitive impairment. *J. Neurol.* 246, 477–485.
- Wahlund, L.O., Barkhof, F., Fazekas, F., Bronge, L., Augustin, M., Sjoergren, M., Wallin, A., Ader, H., Leys, D., Pantoni, L., Pasquier, F., Erkinjuntti, T., 2001. A new rating scale for age-related white matter changes applicable to MRI and CT. *Stroke* 32, 1318–1322.
- Winblad, B., Palmer, K., Kivipelto, M., Jelic, V., Fratiglioni, L., Wahlund, L.O., et al., 2004. Mild cognitive impairment—beyond controversies, towards a consensus: report of the International Working Group on Mild Cognitive Impairment. *J. Intern. Med.* 256, 240–246.
- Yang, S., Murali, T.M., Pavlovic, V., Schaffer, M., Kasif, S., 2003. RankGene: identification of diagnostic genes based on expression data. *Bioinformatics* 19 (12), 1578–1579.
- Zhong, Z., Ewers, M., Teipel, S., Burger, K., Wallin, A., Blennow, K., He, P., McAllister, C., Hampel, H., Shen, Y., 2007. Levels of beta-secretase (BACE1) in cerebrospinal fluid as a predictor of risk in mild cognitive impairment. *Arch. Gen. Psychiatry* 64 (6), 718–726.