NeuroImage 58 (2011) 785–792

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg

Effects of hardware heterogeneity on the performance of SVM Alzheimer's disease classifier

Ahmed Abdulkadir [a,b,*,1], Bénédicte Mortamet [b,1], Prashanthi Vemuri [c], Clifford R. Jack Jr. [c], Gunnar Krueger [b], Stefan Klöppel [a] and The Alzheimer's Disease Neuroimaging Initiative [2]

[a] Department of Psychiatry and Psychotherapy, Section of Gerontopsychiatry and Neuropsychology, Freiburg Brain Imaging, University Medical Center Freiburg, Freiburg, Germany
[b] Advanced Clinical Imaging Technology, Siemens Suisse SA, Healthcare Sector IM&WS-Centre d'Imagerie Biomédicale (CIBM), Lausanne, Switzerland
[c] Department of Radiology, Mayo Clinic, Rochester, MN, USA

ARTICLE INFO

Article history:
Received 19 January 2011
Revised 9 June 2011
Accepted 10 June 2011
Available online 25 June 2011

Keywords:
Magnetic resonance imaging
MRI
Support vector machines (SVM)
Alzheimer's disease
Multi-site study

ABSTRACT

Fully automated machine learning methods based on structural magnetic resonance imaging (MRI) data can assist radiologists in the diagnosis of Alzheimer's disease (AD). These algorithms require large data sets to learn the separation of subjects with and without AD. Training and test data may come from heterogeneous hardware settings, which can potentially affect the performance of disease classification.

A total of 518 MRI sessions from 226 healthy controls and 191 individuals with probable AD from the multicenter Alzheimer's Disease Neuroimaging Initiative (ADNI) were used to investigate whether grouping data by acquisition hardware (i.e. vendor, field strength, coil system) is beneficial for the performance of a support vector machine (SVM) classifier, compared to the case where data from different hardware is mixed. We compared the change of the SVM decision value resulting from (a) changes in hardware against the effect of disease and (b) changes resulting simply from rescanning the same subject on the same machine.

Maximum accuracy of 87% was obtained with a training set of all 417 subjects. Classifiers trained with 95 subjects in each diagnostic group and acquired with heterogeneous scanner settings had an empirical detection accuracy of $84.2 \pm 2.4\%$ when tested on an independent set of the same size. These results mirror the accuracy reported in recent studies. Encouragingly, classifiers trained on images acquired with homogenous and heterogeneous hardware settings had equivalent cross-validation performances. Two scans of the same subject acquired on the same machine had very similar decision values and were generally classified into the same group. Higher variation was introduced when two acquisitions of the same subject were performed on two scanners with different field strengths. The variation was unbiased and similar for both diagnostic groups. The findings of the study encourage the pooling of data from different sites to increase the number of training samples and thereby improving performance of disease classifiers. Although small, a change in hardware could lead to a change of the decision value and thus diagnostic grouping. The findings of this study provide estimators for diagnostic accuracy of an automated disease diagnosis method involving scans acquired with different sets of hardware. Furthermore, we show that the level of confidence in the performance estimation significantly depends on the size of the training sample, and hence should be taken into account in a clinical setting.

Introduction

Fully automated methods detecting presence or absence of Alzheimer's disease (AD) based on structural magnetic resonance imaging (MRI) data can help radiologists (Klöppel et al., 2008; Magnin et al., 2009; Plant et al., 2010; Vemuri et al., 2008). AD is associated with formation of extracellular amyloid immunoreactive senile plaques and tau immunoreactive neurofibrillary tangles (Braak and Braak, 1991). It is also associated with progressive atrophic changes that can be detected by structural MRI. Subjects with AD typically show patterns of gray matter (GM) atrophy involving the medial temporal lobe, particularly the hippocampus and entorhinal cortex,

* Corresponding author at: Department of Psychiatry and Psychotherapy, Section of Gerontopsychiatry and Neuropsychology, Freiburg Brain Imaging, University Medical Center Freiburg, Freiburg, Germany. Fax: +49 761 270 54160.
E-mail address: ahmed.abdulkadir@epfl.ch (A. Abdulkadir).

[1] These authors contributed equally to this work.

[2] Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (http://www.loni.ucla.edu/ADNI). 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. ADNI investigators include (complete listing available at: http://adni.loni.ucla.edu/wp-content/uploads/how_to_apply/ADNI_Authorship_List.pdf).

1053-8119/$ – see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.neuroimage.2011.06.029

among other brain regions, with simultaneous expansion of the ventricles (Baron et al., 2001; Fox et al., 1996; Jack et al., 1992; Whitwell et al., 2007). Due to the characteristic atrophy pattern, the GM is an informative biomarker to detect AD with structural MRI (Klöppel et al., 2008; Magnin et al., 2009; Vemuri et al., 2008).

An increasing number of multi-center studies aim to combine data from different scanners to increase statistical power and fields of applications. Studies suggest that data from different sites can be pooled, but at the same time that systematic inter-scanner differences can occur. Stonnington et al. (2008) compared the variation of data acquired on six distinct scanners of same vendor/type on a voxel-by-voxel level with a mass univariate test on GM probability maps and concluded that the effect of AD is significantly larger than the inter-scanner effects. On the other hand, several studies indicate that the effects of inter-scanner variability are far greater than intra-scanner variability (Huppertz et al., 2010; Moorhead et al., 2009). Similarly, bias field correction and variation in image quality such as signal to noise ratio (SNR) have an impact on the segmentation (Acosta-Cabronero et al., 2008; Klauschen et al., 2009; Shuter et al., 2008). Previous classification methods detecting presence of AD from structural MRI data indicate that performance improved when a high number of samples were used for training (Franke et al., 2010; Klöppel et al., 2009). This may entail the need to pool data from different manufacturers and hardware settings.

The Alzheimer's Disease Neuroimaging Initiative (ADNI) (Mueller et al., 2005) is a large, multi-center, multi-vendor study that acquires structural MRI of cognitively normal healthy controls (CN), mild cognitive impaired (MCI) and AD-probable (AD-p) elders. The ADNI protocols on each scanner type are adjusted such that all sites report comparable results at all times (Jack et al., 2008). Intensive quality control and the use of a phantom, assure low inter-scanner variation and high stability of the image quality (Gunter et al., 2009).

In this study we used data from 56 different sites that participated in the ADNI study to assess the change in detection performance of an AD classifier trained with images acquired either with homogenous or heterogeneous hardware. As in previous work (Klöppel et al., 2008), we used a fully automated processing pipeline and a support vector machine (SVM) classifier (Vapnik, 1998). The process that computes spatially normalized GM probability maps in a common template space from structural T1MRI images was found to outperform other approaches in a recent comparison using multi-site data from ADNI (Cuingnet et al., 2011). We set out to investigate the impact of heterogeneity of the acquisition hardware on the classifier outcome. First, as coarse measure of the performance, we computed the accuracy of classifiers trained on homogenous hardware (pure set). Then we computed the ranges of accuracies that can be expected from classifiers trained on randomly selected images from heterogeneous hardware (mixed sets) with the same sample sizes as the pure sets. These distributions were then compared to the previously observed accuracies of each pure set. Second, in order to quantify hardware-related effects we introduced the analysis of the SVM decision value. Positive values indicated AD-p and negative values indicated CN. Ideally, the decision value should depend only on the subject, not on the hardware. The further away from zero, the higher is the confidence of the classifier in its decision. With the intention to determine the minimal uncertainty of this value due to acquisition noise and pre-processing, we quantified the variation of the decision value between back-to-back scans of subjects. Then we quantified the variation of the decision value between scans of same subjects on both field strengths.

Materials

Participants and image acquisition

Our data included T1-weighted MR images from 417 individuals of which 226 were cognitively normal healthy controls (Mini-mental state examination (MMSE): 29.1 ± 1.0, age: 76.1 ± 5.0) and 191 had probable AD (MMSE: 23.3 ± 2.1, age: 75.5 ± 7.5). All images were obtained from ADNI. Inclusion criteria for participants were according to the protocol described in http://www.adni-info.org/scientists/AboutAdni.aspx#. Individuals assigned to the AD-p group met NINCDS/ADRDA criteria for probable AD (McKhann et al., 1984). We first selected all ADNI CN and AD-p subjects with a baseline MRI scan (all were scanned on 1.5 T, a subset also on 3 T). We excluded 2 AD subjects that progressed to some other dementia during follow-up. The median follow-up time for all patients was 24 months. The interquartile ranges (IQR) by field strength are listed here: 1.5 T–IQR: 24–36 months and 3 T–IQR: 24–31 months. Three subjects were further excluded because the required baseline images were not available. A total number of 417 subjects were included. The list of all images is attached in the supplementary material. T1-weighted sagittal volumes were obtained using the magnetization-prepared rapid gradient-echo (MP-RAGE) pulse sequence with imaging parameters TR = 2300 ms, TI = 900 ms, flip-angle = 9°at 3 T (and TR = 2400 ms, TI = 1000 ms, flip angle = 8°at 1.5 T) minimum full TE, sagittal slices = 160. All 1.5 T subject acquisitions used 1.25×1.25 mm^2 in-plane spatial resolution and 1.2-mm thick sagittal slices. The 3 T subject acquisitions also used 1.2-mm thick sagittal slices, but were acquired with 1.0×1.0 mm^2 in-plane spatial resolution. Back-to-back scans were acquired from each subject within each scanning session and an image analyst at Mayo clinic rated the image quality of each scan. Quality criteria included blurring/ghosting, flow and susceptibility artifacts. For the analysis based on accuracy we included the ADNI baseline scan (Timepoint 1) with the best quality rating to avoid misclassifications due to low quality, e.g. caused by motion artifacts. For the analysis of the impact when changing field strength, we included further 192 back-to-back scans with a lower or equal quality compared to the other image acquired at the same session. The ADNI structural brain imaging data can be downloaded with or without certain processing steps applied (see http://www.loni.ucla.edu/ADNI/Data/ADNI_Data.shtml). Availability of pre-processing steps depends on manufacturer and coil system (Jack et al., 2008). We included images that were corrected for system-specific image geometry distortion due to gradient non-linearity (GradWarp) and, if available, additional image intensity non-uniformity (B1 correction).We excluded subjects with diagnosed MCI to reduce biological variability, as this diagnostic group is arguably the most heterogeneous. The scanner configurations considered were (a) manufacturer, namely Siemens Healthcare, GE Healthcare and Philips Medical Systems, (b) magnetic field strength, namely 1.5 T and 3 T, and (c) coil system, namely single-channel birdcage coils (BC) and multi-channel phased-array head coils (PA). We focused on these parameters as they were explicitly taken into account during the establishment of the MRI protocols for the ADNI study (Jack et al., 2008). Other configurations like scanner software version, detailed coil configuration or coil type were not considered. Platform-specific lists of sequence parameters are available at http://www.loni.ucla.edu/ADNI/Research/Cores/.

Each of the 417 individuals had a baseline scan at 1.5 T. Among these, 101 participants had a second scan within 2 to 102 days (24 ± 15 days) in a scanner with 3 T. For the rest of the article, we will refer to the 316 images of higher quality of individuals that did not have a scan at 3 T as SOLO_1.5 T and we will refer to the two sets of 101 images from individuals that had an image at both magnetic field strengths as PAIR_1.5 T and PAIR_3.0 T respectively. All resulting 26 subgroups are listed in Supplementary Table 1. There was a trend towards age difference in two of these groups. The subgroup with lowest MMSE of the AD-p group had 22.6 ± 2.0 [18–26], and the highest MMSE score of AD-p group was 24.1 ± 2.2 [20–28] ($p = 0.03$). No significant differences in the MMSE between control groups were observed.

Information on the ADNI

Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.ucla.edu).The ADNI was launched in 2003 by the National Institute on Aging (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the Food and Drug Administration (FDA), private pharmaceutical companies and non-profit organizations, as a $60 million, 5-year public–private partnership. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessments can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). Determination of sensitive and specific markers of very early AD progression is intended to aid researchers and clinicians to develop new treatments and monitor their effectiveness, as well as lessen the time and cost of clinical trials. The Principal Investigator of this initiative is Michael W. Weiner, MD, VA Medical Center and University of California-San Francisco. ADNI is the result of efforts of many co-investigators from a broad range of academic institutions and private corporations, and subjects have been recruited from over 50 sites across the U.S. and Canada. The initial goal of ADNI was to recruit 800 adults, ages 55 to 90, to participate in the research—approximately 200 cognitively normal older individuals to be followed for 3 years, 400 people with MCI to be followed for 3 years and 200 people with early AD to be followed for 2 years. For up-to-date information, see www.adni-info.org.

Methods

Brain segmentation and registration

Image pre-processing was carried out using SPM8 software (http://www.fil.ion.ucl.ac.ek/spm). Images were automatically co-registered to a head template of a single subject, which was aligned with the prior tissue probability maps. Unified Segmentation algorithm (Ashburner and Friston, 2005) was used in combination with a high-dimensional image warping approach (Ashburner, 2007), as suggested in Klöppel et al. (2008) and validated in Cuingnet et al. (2011), to obtain spatially normalized GM probability maps. In this process, a template representing the average GM anatomy of the population was created and all images were warped into the space of this template with isotropic voxel size of 1.5 mm and spatial resolution of $121 \times 145 \times 121$ voxels. Subsequent modulation (Ashburner and Friston, 2000) was applied to ensure that the overall amount of each tissue class remained constant after spatial normalization.

Classification

An SVM is a high-dimensional pattern classification method. Given a set of samples with known group labels and a kernel function, the SVM computes a high dimensional hyperplane that separates two groups. In this study, the spatially normalized and modulated map of GM probability represents one sample with voxel values providing individual features of the local anatomy. Unlike the toy example depicted in Supplementary Fig. 1, the dimensionality of the problem of this work is not 2, but about $2 \cdot 10^6$. The number of training samples is smaller or equal to 417. Therefore the dimensionality of the problem is much larger than the number of samples. Furthermore, a large fraction of the dimensions carries no relevant information, e.g. background or brain regions not related to the disease. Dimensionality reduction strategies such as principle component analysis could be employed to reduce dimensionality without loss (and potentially even with gain) of classification performance. However, no further treatment was applied to the GM maps, because the goal of this study was to use a well-tested and accepted method (Cuingnet et al., 2011; Klöppel et al., 2008) with few parameters in order to minimize effects due to classifier-specific parameters such as dimensionality reduction methods. We used an implementation of a *C*-SVM (Boser et al., 1992; Cortes and Vapnik, 1995) by libsvm (Chang and Lin, 2001) with a linear kernel function $K_{\mathrm{LINEAR}}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \boldsymbol{x}_i^T \boldsymbol{x}_j$. Cost parameter C was fixed at the default value of 1. In preliminary test, $C > 1$ did not influence the accuracy. A mathematical formulation of the SVM classification is presented in the supplementary materials. To evaluate each model and estimate its ability to correctly classify unseen data, we performed two validation methods: leave-one-sample-out cross-validation (LOO-CV) and a validation with an independent test set. LOO-CV was adopted, as some of the pure hardware groups were too small to split them into a training and a test set. On the other hand, using LOO-CV was not possible when training was done with one hardware set, and testing with another.

In the LOO-CV each sample of the training set with total n samples was tested with a model built with all the other $n-1$ samples. This procedure is repeated until each scan has once been left out. In addition to the LOO-CV, the validation of the performance using an independent test set was performed. We refer to the evaluation of images unseen in training as testing.

Performance of disease classifiers

We subdivided the images into sets according to the hardware used to acquire them (see Supplementary Table 1). The hardware type refers to a specific combination of manufacturer, coil system and field strength. All homogeneous data sets contained acquisitions from different sites. We wanted to compare the performance of the classifiers trained on subgroups of images acquired with a pure hardware configuration, which reduced the training sample size. On the other hand, the largest possible number of subjects was desirable to get a more stable performance estimate. Furthermore, reduction to subsamples based on acquisition hardware led to different sample sizes for each hardware setting as well as an unbalanced number of AD-p and CN samples. To tackle these partly contradicting issues, we decided (a) to report the performance on equally sized groups of patients and controls, (b) to repeat this analysis at different group sizes (20, 40, 60, 80 and 95 subjects) whenever available, and (c) to require a minimum size of about 20 subjects per group, as previous work has shown a substantial decrease of performance with smaller groups (Klöppel et al., 2009).

For each group size, training and testing subsets with equal numbers of AD-p and CN was randomly selected. This was repeated 500 times to obtain mean, standard deviation and empirical confidence intervals (CI) for each group size. To allow comparison with previous studies (Cuingnet et al., 2011; Franke et al., 2010), we report performance of classifiers trained with a large sets of images.

Fig. 1 shows a schematic flowchart of the method we used to compare the accuracy of each pure setting to the expected accuracy of mixed settings of the same group size. This was based on our hypothesis that a classifier should perform better with a pure hardware setting. In the resulting analyses, the pure sets were randomly split into subsets with equal number of controls and AD-probable subjects (either 20, 40 or 60 subjects per group). Then the LOO-CV of the subset was computed. These steps were repeated 500 times to obtain the accuracy distribution including mean (x_{pure}) and standard deviation as above. The mean accuracy for each pure set was then compared to a distribution of mean accuracies obtained by computing the mean (μ) and standard deviation (σ_{mixed}) of 500 mean accuracies over 500 randomly selected mixed sets, each having the same number of CN and AD-p as the hardware-pure set. The mean accuracy of the pure set was statistically compared to the expected mean accuracy of the random sets using a z-test $z = (x_{\mathrm{pure}} - \mu)/\sigma_{\mathrm{mixed}}$.

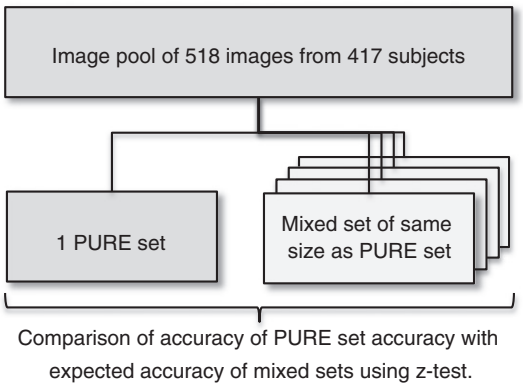

Fig. 1. Flow-chart representation of the comparison of an example computation to the mean accuracy distribution of random sets. Comparison of PURE set accuracy with expected accuracy of mixed sets using z-test.

Mutual testing across hardware sets

We mutually tested image sets by models trained on independent image sets. The list of these sets and their demographic distribution is specified in Supplementary Table 1. The aspect of field strength was investigated with 101 patients that were scanned at 1.5 T and 3 T. Additionally, we used a third, independent set of images (n = 316). We computed the LOO-CV for these three complete sets and mutually tested the sets of images. We also computed 500 times the LOO-CV and mutual testing accuracy with randomly selected reduced training sets (40CN/40AD-p) to obtain distributions of accuracies similar to the previously described assessment of the performance with different sample sizes. Thereby, we accounted for random effects that could alter the accuracy. The two image sets from the same subjects were not mutually tested. Testing of images from subjects that were in the training set would have led to an unrealistic high accuracy.

Variance of decision values

Differences in classification accuracy when the same subjects were scanned at 1.5 T and 3 T may be a true effect of field strength but may also be random. For this part of the analysis, we were interested in investigating how stable the diagnosis of a patient would be, independently of its correctness. A trained linear SVM classifier defines a N-dimensional decision boundary Ω that is defined by $\Omega = \{\boldsymbol{x} \in \mathbb{R}^N : \boldsymbol{w}^T\boldsymbol{x} + b = 0\}$, where the vector $\boldsymbol{w}$ and the scalar b are learned model parameters. Points on opposed sides have positive and negative signs respectively and hence the class label of a sample $\boldsymbol{x}_*$ is determined by the sign of the decision value $d = \boldsymbol{w}^T\boldsymbol{x}_* + b$. To assure valid interpretation of decision values and their intra-subject variability, only values computed with the same model, i.e. weight vector $\boldsymbol{w}$ and bias b were compared, and the variance was compared with the distance between the means of both classes.We quantified the error within two back-to-back scans, i.e. the minimal variation of the decision value caused by the whole acquisition and post-acquisition processing pipeline. In order to quantify the variation across hardware and isolate a systematic effect of field strength, we computed the differences between decision values within two images of the same patient acquired on two different systems, one at 1.5 T and the other at 3 T respectively. We will refer to it as a change of field strength but would like to point out that the systematic change in field strength was irregularly accompanied by a change in coil system and/or manufacturer.

The decision values were computed with two different training sets. The first was composed of 316 images (SOLO_1.5 T), since it was the largest possible data set that did not include samples from individuals that were in the testing set. This data set is supposed to give the highest accuracy. The second training set was composed of 80 randomly selected scans (40 CN and 40 AD). The testing set for both training sets was composed of 384 images from 96 participants with both back-to-back baseline scans available from 1.5 T and 3 T. Using the training set we estimated the effect of the disease on group level as well as the within-subject variance between two repeated scans acquired back-to-back and between two acquisitions of the same subject but acquired at different field strength. Furthermore we were interested in the effect of sample size on the variability.

Results

General performance

For mixed sets of equal size, and 95 probable AD subjects and 95 healthy controls in each set we empirically obtained 83.9 ± 2.3% LOO-CV accuracy (percentile 2.5% was at 79.5% and percentile 97.5% was at 88.4%). For the same subjects, but using only images acquired at 1.5 T, we empirically obtained 84.4 ± 2.2% LOO-CV accuracy (percentile 2.5% was at 80.0% and percentile 97.5% was at 88.4%). Performance on independent test sets was similar (Table 1). When changing the size of the training set while keeping balanced diagnostic groups the mean performance increased with training sample size and the variance decreased as shown in Fig. 2. The two classifiers trained with the maximum number of independent images from 417 subjects each, reach a LOO-CV accuracy of 87.5% (1.5 T only) and 86.6% (1.5 T mixed with 3 T).

Performance of hardware pure sets

We subdivided the set of all images according to the main hardware components as listed on Supplementary Table 1. The performance of randomly composed mixed sets was used as reference for estimated quality at a given group size. Table 2 reports the LOO-CV accuracy of the pure sets and the expected mean accuracy of randomly composed classifiers with the same number of CN and AD-p subjects. None of the mean accuracies of the pure sets was significantly better than the expected accuracy of mixed sets. Further LOO-CV accuracies of pure sets can be found in Supplementary Figs. 1–3. The analysis of the effect of field strength on the performance is presented in Table 3. LOO-CV accuracy with 40 CN and 40 AD-p of PAIR_1.5 T was 77.5 ± 2.4 [72.5 77.5 82.5] and the LOO-CV of PAIR_3.0 T was 76.9 ± 2.4 [72.5 76.25 81.25].

Table 1
Performance of classifiers on independent test sets as function of group size. Group sizes of training and testing set were always equal. Each row presents the summary of 500 runs in which a random subset was selected.

Group size	Acquisitions from 1.5 T only (mean accuracy ± standard deviation) [q2.5% q25% q50% q75% q97.5%]	Acquisitions from 1.5 T and 3 T (mean accuracy ± standard deviation) [q2.5% q25% q50% q75% q97.5%]
10	(72.3 ± 11.0)% [50.0 65.0 72.5 80.0 90.0]%	(71.6 ± 10.8)% [50.0 65.0 70.0 80.0 90.0]%
20	(76.2 ± 6.9)% [62.5 72.5 77.5 80.0 87.5]%	(76.5 ± 7.3)% [62.5 72.5 77.5 80.0 90.0]%
40	(80.7 ± 4.7)% [70.0 77.5 81.2 83.8 88.8]%	(80.2 ± 4.6)% [70.0 77.5 80.0 83.8 88.8]%
60	(82.9 ± 3.5)% [75.8 80.8 83.3 85.0 88.3]%	(82.3 ± 3.6)% [75.0 80.8 82.5 85.0 89.2]%
80	(83.9 ± 2.7)% [78.1 81.9 83.8 85.6 88.8]%	(83.5 ± 2.8)% [77.5 81.9 83.8 85.6 88.1]%
95	(84.5 ± 2.5)% [78.9 82.6 84.7 86.3 88.9]%	(84.2 ± 2.4)% [79.5 82.6 84.2 86.3 88.4]%

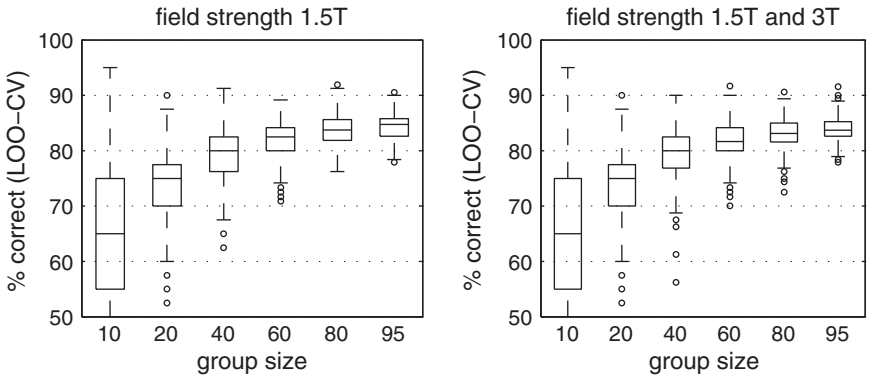

Fig. 2. Box-plots of leave-one-out cross-validation (LOO-CV) accuracy as function of group size (x-axis) obtained by 500 random permutations.

Mutual testing across hardware sets

For pure subgroups, a training and mutual testing was performed in three steps. (a) For each group, we trained a classifier. (b) We computed its LOO-CV accuracy as estimate of the detection accuracy. (c) Each classifier was tested on all image sets that did not overlap with the training set. Resulting lookup tables can be found in Supplementary Figs. 2 and 3. In these tables all images from each set were taken, whereas in Supplementary Fig. 1, subsets had equal size with equal number of AD-p and CN to allow a direct comparison. Qualitatively, one can observe that in general larger subsets perform better. However, large differences in performances can be seen in subsets of same size. Combining samples from BC and PA from the same manufacturer generally improve LOO-CV and testing accuracy (Supplementary Fig. 2). See other supplementary material for more details.

Table 2
Subgroup performance. Each image set represents a unique hardware setting. Accuracy distribution from 500 permutations within pure image set is compared to the variation of the mean from 500 randomly composed image sets from all available images.

Hardware set	Count	Group size	Pure set ($x_{pure} \pm \sigma_{xpure}$)% [p2.5% p50% p97.5%]	Mixed hardware ($\mu \pm \sigma_{xmixed}$)%	z-score $\frac{x_{pure}-\mu}{\sigma_{xmixed}}$
1.5 T | Siemens | BC	24CN/23AD	20	76.2 ± 3.8 [70.0 75.0 85.0]	73.9 ± 5.6	0.41
1.5 T | Siemens | PA	66CN/48AD	40	83.9 ± 2.6 [78.8 83.8 88.8]	79.1 ± 3.0	1.60
		20	78 ± 5.4 [67.5 78.8 87.5]	73.9 ± 3.0	1.37
1.5 T | GE | BC	31CN/28AD	20	75.8 ± 4.5 [67.5 75.0 85.0]	73.8 ± 4.7	0.43
1.5 T | GE | PA	77CN/66AD	60	79.4 ± 2.4 [74.2 79.283.3]	81.9 ± 2.6	− 0.97
		40	76.0 ± 3.6 [68.8 76.2 82.5]	79.4 ± 2.2	− 1.55
		20	69.5 ± 7.2 [52.5 70.0 82.5]	73.8 ± 2.5	− 1.72
1.5 T | Philips | BC/PA	28CN/26AD	20	72.5 ± 4.3 [65.0 72.5 80.0]	73.9 ± 4.8	− 0.29
3.0 T | Siemens | BC/PA	29CN/23AD	20	71.5 ± 4.2 [62.5 72.5 80.0]	73.6 ± 5.1	− 0.41
PAIR_1.5 T	60CN/41AD	40	77.7 ± 2.5 [72.5 77.5 82.5]	79.1 ± 3.3	− 0.42
PAIR_3.0 T	60CN/41AD	40	76.9 ± 2.5 [71.2 77.5 81.2]	79.1 ± 3.3	− 0.67

When exploring the effect of field strength, LOO-CV accuracy of 1.5 T scans in the sets PAIR_1.5 T and PAIR_3.0 T reached 80.2% and predicted the remaining 316 images equally well with accuracy of 82.0% and 82.3% respectively. The independent SOLO_1.5 T set detected the PAIR_1.5 T set with 88.1% and the PAIR_3.0 T image set with 83.2% accuracy. As shown in Table 3, subsets with 40 subjects per group from SOLO_1.5 T detected PAIR_1.5 T on average with 81.9 ± 3.0% accuracy and PAIR_3.0 T with 80.4 ± 2.7% accuracy. SOLO_1.5 T was predicted with 82.1 ± 1.4% by PAIR_1.5 T and with 81.7 ± 1.5% by PAIR_3.0 T.

Variance of decision values

In Fig. 3, the variation within back-to-back scans and across two field strengths are plotted and compared to the distributions of the decision values of each diagnostic group (Fig. 4). We observed that the variance between two back-to-back scans did neither depend on the diagnostic group, nor the field strength. Using the large sample (n = 316), the mean decision value of participants with probable AD was 0.41 and the mean of the control group was − 0.61. This indicates that on average the decision value of subjects with AD-p was 1.02 smaller than the decision value of control subjects. The standard deviation of repeated scans was 0.08 (within subjects) and the standard deviation of changing field strength was 0.20 (within subjects). The values obtained with a classifier with less training samples (n = 80) were consistently smaller. The between-group difference was equal to 0.66, the standard deviation of the error of repeated measure was 0.06 (within subjects) and the standard deviation of changing field strength was 0.15 (within subjects). The class-wise decision value distributions shown in Fig. 4 were normally distributed ($p > 0.1$) as verified with the Kolmogorov–Smirnov test. Within participants, the change of the system from 1.5 T to 3 T (Figs. 3(A+B) right panel and Figs. 4(A+B) right panel) did not introduce a systematic shift of the decision value(one-sample *t*-test $p > 0.05$). The introduced variance between repeated scans was 13.5 and 11 times smaller than the between-group difference, for the two training sizes respectively. The variation due to change in field strength was 2.5 times higher than the variation within two back-to-back scans. This error introduced by changing hardware is sufficiently large to change the decision.

The difference of the SVM-decision value between diagnostic groups was smaller when the number of training samples was decreased (n = 80) but was less variable within each group. Back-to-back differences within subjects were also smaller (Fig. 3(B)).

Discussion

We processed 710 images acquired at 56 different sites, cross-validated and tested classifiers for their detection accuracy to evaluate a possible effect of the manufacturer, magnetic field strength or coil configuration. Apart from a trend in two sets, the

Table 3
Mutual training and testing of two image sets acquired at 1.5 T and one image set acquired at 3 T. PAIR_1.5 T and PAIR_3.0 T are images from same subject. Image set SOLO_1.5 T is independent from the other two. Bold values in the diagonal are leave-one-out cross-validation accuracies (% correct). Off-diagonal values represent testing accuracies (% correct) of image sets in the same column with the training set listed in the same line. In panel A, training was performed with the full data set (316 scans), whereas in panel B, subsets with each having 40 randomly selected subjects per diagnostic group were used for each cross-validation and testing run. In B, the mean and standard deviation from 500 runs is reported.

	n	SOLO_1.5 T	PAIR_1.5 T	PAIR_3.0 T
A				
SOLO_1.5 T	316	**85.1**	88.1	83.2
PAIR_1.5 T	101	82.0	**80.2**	–
PAIR_3.0 T	101	82.3	–	**80.2**
B				
SOLO_1.5 T	316	**79.5 ± 4.0**	81.9 ± 3.0	80.4 ± 2.7
PAIR_1.5 T	101	82.1 ± 1.4	**77.5 ± 2.4**	–
PAIR_3.0 T	101	81.7 ± 1.5	–	**76.9 ± 2.4**

subjects in all sets had equal age distributions. The largest two classifiers were trained on 417 images and their cross-validation accuracies reached on average 87% which corresponds to previously reported performances (Cuingnet et al., 2011). The mean LOO-CV accuracy consistently increased, as expected, with higher numbers of training samples. When applying the classification algorithm to new data sets, the accuracy of the proposed method was reasonably high and performed comparable to several other approaches both on ADNI data as well as single site data (Cuingnet et al., 2011; Plant et al., 2010). Unlike most studies, which either have single site data or pool the entire ADNI data as training samples to validate the algorithm, our goal was to test the hardware effects on classification accuracy and required us to separate the data into smaller subgroups.

We assumed that pure hardware sets would show a better LOO-CV accuracy. Performance of individual pure sets of images varied strongly, as shown in Supplementary Figs. 1–3. Such single performance values are an uncertain estimator, and results from the permutation tests (Table 2) indicate that classifiers using images acquired with mixed hardware performed equally well. Since each pure set of images consisted of different subjects, the effect of individual anatomy on the accuracy was a covariate to the hardware effect. It is important to note that the sample sizes in each of the subgroups were different but was greater than 20, which was found to be the minimum for the proposed classification problem (Klöppel et al., 2009). Even though large sample sizes may mean more stable classifiers and better performance, the performance of each of the different hardware settings (Table 2) was found to be statistically similar.

In the comparison of PAIR_1.5 T and PAIR_3.0 T, the effect of individual anatomy was reduced to changes of aging and eventually progressive atrophy caused by disease over a period of 2 to 102 days. Because all images at 3 T were acquired after the 1.5 T scans, we expected the set of images taken at a later time point to be more or equally discriminative due to the progression of the disease in some individuals. Experimentally the opposite was observed. Classifiers trained on images acquired at 1.5 T predicted the image sets acquired at the same field strength slightly (1.6 percentage points) but significantly better. The test result of the SOLO_1.5 T set performed 6 percentage points better on the 1.5 T than on the 3 T test data (Table 3). Given that the test sets were composed of the same subjects, these differences are remarkable. However, it was probably due to chance, since the variation of the decision value was centered on zero for both diagnostic groups (Fig. 4). The higher SNR of 3 T systems compared to 1.5 T was by design used in the ADNI study to increase spatial resolution (Jack et al., 2008). Higher resolution of the images did in this case not improve performance potentially because the processing pipeline included the resampling of GM maps to 1.5 mm isotropic voxel size during spatial normalization. Reproducibility of the decision value was similarly high, for both sample sizes tested. The standard deviation of the introduced error was more than 10 times higher than the difference between the means of the diagnostic groups. Changing field strength of the scanner led to variance that was 3 times higher than the back-to-back variance. Despite a change in field strength, no systematic effect on the decision value could be observed. The small training set was not more vulnerable to changes in hardware; on the other hand, the larger training set did not decrease these kinds of errors. The difference in the decision value between groups increased with the size of the training set. The large data set pronounced differences related to the disease but also differences that are related to the acquisition process. When the number of training samples was small, adding samples from heterogeneous hardware to the training set increased the accuracy of the classifier, assumingly because benefits from a larger sample size exceed those of hardware inhomogeneity.

From these results we conclude that reproducibility of the post-acquisition pipeline is similarly high at both field strengths. The source of variation – indistinguishable with the performed analysis – are (a) scanner noise, (b) varying image quality, (c) variations in any step of the pre-processing pipeline such as segmentation, resampling or spatial normalization. Furthermore a change in hardware setting introduces variation that can shift the decision value substantially. Two possible explanations come to mind: (a) Random effects due to physiological conditions of the patient, the positioning of the head, motion, or (b) Systematic effects that are related to a specific change in system. It should, however, be kept in mind that the results of the current study cannot readily be extended to multi-center with a less stringent system of quality control. In addition, the attempts to increase the comparability between 1.5 T and 3 T data are specific to

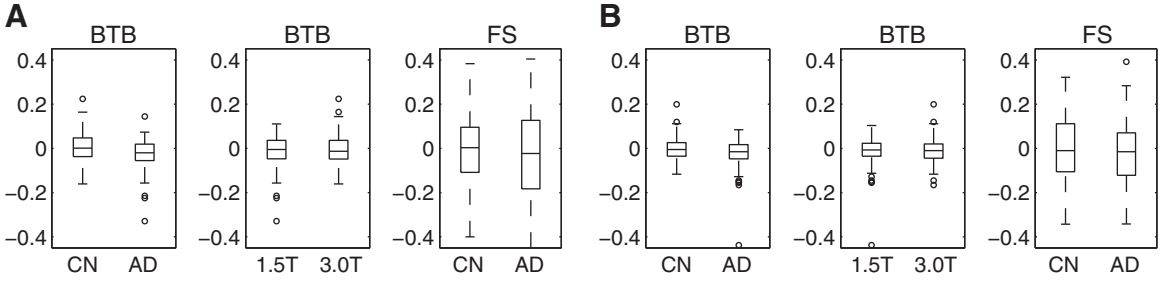

Fig. 3. Changes of the SVM decision value (y-axis) between back-to-back scans (BTB), separately for diagnostic group (first panel) and field strength (second panel), and with changes between two field strengths (thus also two systems), separately for diagnostic group (third panel) for acquisitions of 96 subjects. Change in field strength (FS) does not introduce systematic bias (one-sample *t*-test $p > 0.05$). A: Training set composed of 316 images. B: Training set composed of 80 images.

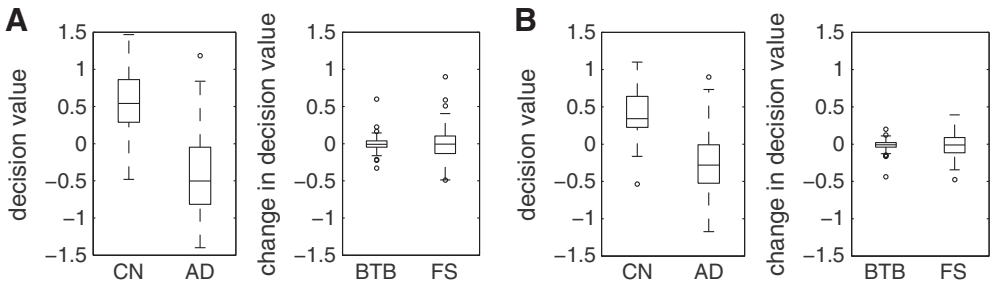

Fig. 4. Variance of decision values when comparing back-to-back scans or change of field strength compared to the effect of group. A: Training set size = 316. B: Training set size = 80. CN: healthy controls, AD: subject with probable AD, BTB: back-to-back (within subjects), FS: field strength (within subjects).

the ADNI study and allowed a successful classification across field strengths.

The results of this study have substantial implication for the clinical setting. Changing field strength introduces additional variance in the computed decision value and thus decreases accuracy, compared to repeated measures on the same scanner. It should be noted that two scanners with the identical hardware setting will not produce exactly identical results and this may also influence classification accuracy. From the practical point of view, the choice of hardware would normally influence the decision in about 5% of the cases. The obtained accuracy of about 84% presents encouraging results for automated SVM-based disease classifier with the use of images acquired at different centers in comparison to conventional clinical ante-mortem AD diagnosis, which is not 100% reliable. Specifically, approximately 30% of cognitively normal subjects will meet pathological criteria for AD at post-mortem (Morris and Price, 2001). Especially when the number of available samples from one center was small, the combination of training images from two sets often resulted in a clear improvement of performance. The results did not indicate that mixing data from different centers would lead to substential loss of classification accuracy.

Since the 95% CIs of the performance were varying as function of training sample size, and were large for small sample sizes (e.g. 62.5–90% with 20 subjects per diagnostic group), a quantification of the performance by a single estimation of the accuracy is doubtful. Reporting CI confidence intervals as in Fig. 2 strengthens the interpretability of the estimation of the classification performance and provides a measure of diagnostic confidence for clinical applications.

Supplementary materials related to this article can be found online at doi:10.1016/j.neuroimage.2011.06.029.

Acknowledgments

We would like to thank Anthony Lissot for his recommendations regarding statistical analysis.

This work was supported by the Centre d'ImagerieBioMédicale (CIBM) of the UNIL, UNIGE, HUG, CHUV, EPFL, and the Leenaards and Jeantet Foundations. This work was supported by the Siemens Schweiz AG.

Dr. Jack's and Dr. Vemuri's time was supported in part by NIH grant AG11378.

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: Abbott, AstraZeneca AB, Bayer Schering Pharma AG, Bristol-Myers Squibb, Eisai Global Clinical Development, Elan Corporation, Genentech, GE Healthcare, GlaxoSmithKline, Innogenetics, Johnson and Johnson, Eli Lilly and Co., Medpace, Inc., Merck and Co., Inc., Novartis AG, Pfizer Inc, F. Hoffman-La Roche, Schering-Plough, Synarc, Inc., as well as non-profit partners the Alzheimer's Association and Alzheimer's Drug Discovery Foundation, with participation from the U.S. Food and Drug Administration. Private sector contributions to ADNI are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of California, Los Angeles. This research was also supported by the NIH grants P30 AG010129, K01 AG030514, and the Dana Foundation.

References

Acosta-Cabronero, J., Williams, G.B., Pereira, J.M.S., Pengas, G., Nestor, P.J., 2008. The impact of skull-stripping and radio-frequency bias correction on grey-matter segmentation for voxel-based morphometry. Neuroimage 39, 1654–1665.

Ashburner, J., 2007. A fast diffeomorphic image registration algorithm. Neuroimage 38, 95–113.

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

Ashburner, J., Friston, K.J., 2005. Unified segmentation. Neuroimage 26, 839–851.

Baron, J.C., Chételat, G., Desgranges, B., Perchey, G., Landeau, B., de la Sayette, V., Eustache, F., 2001. In vivo mapping of gray matter loss with voxel-based morphometry in mild Alzheimer's disease. Neuroimage 14, 298–309.

Boser, B.E., Guyon, I.M., Vapnik, V.N., 1992. A training algorithm for optimal margin classifiers. Fifth Annual Workshop on Computational Learning Theory, Pittsburgh. ACM, pp. 144–152.

Braak, H., Braak, E., 1991. Neuropathological stageing of Alzheimer-related changes. Acta Neuropathol. 82, 239–259.

Chang, C.-C., Lin, C.-J., 2001. LIBSVM: a library for support vector machines(Software available at) http://www.csie.ntu.edu.tw/~cjlin/libsvm 2001.

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

Cuingnet, R., Gérardin, E., Tessieras, J., Auzias, G., Lehéricy, S., Habert, M.-O., Chupin, M., Benali, H., Colliot, O., 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.

Fox, N.C., Freeborough, P.A., Rossor, M.N., 1996. Visualisation and quantification of rates of atrophy in Alzheimer's disease. Lancet 348, 94–97.

Franke, K., Ziegler, G., Klöppel, S., Gaser, C., the Alzheimer's Disease Neuroimaging Initiative, 2010. Estimating the age of healthy subjects from T1-weighted MRI scans using kernel methods: exploring the influence of various parameters. Neuroimage 50, 883–892.

Gunter, J.L., Bernstein, M.A., Borowski, B.J., Ward, C.P., Britson, P.J., Felmlee, J.P., Schuff, N., Weiner, M., Jack, C.R., 2009. Measurement of MRI scanner performance with the ADNI phantom. Med. Phys. 36, 2193–2205.

Huppertz, H.-J., Kröll-Seger, J., Klöppel, S., Ganz, R.E., Kassubek, J., 2010. Intra- and interscanner variability of automated voxel-based volumetry based on a 3D probabilistic atlas of human cerebral structures. Neuroimage 49, 2216–2224.

Jack Jr., C.R., Petersen, R.C., O'Brien, P.C., Tangalos, E.G., 1992. MR-based hippocampal volumetry in the diagnosis of Alzheimer's disease. Neurology 42, 183–188.

Jack Jr., C.R., Bernstein, M.A., Fox, N.C., Thompson, P., Alexander, G., Harvey, D., Borowski, B., Britson, P.J.L., Whitwell, J., Ward, C., Dale, A.M., Felmlee, J.P., Gunter, J.L., Hill, D.L.G., Killiany, R., Schuff, N., Fox-Bosetti, S., Lin, C., Studholme, C., DeCarli, C.S., Krueger, G., Ward, H.A., Metzger, G.J., Scott, K.T., Mallozzi, R., Blezek, D., Levy, J., Debbins, J.P., Fleisher, A.S., Albert, M., Green, R., Bartzokis, G., Glover, G., Mugler, J., Weiner, M.W., 2008. The Alzheimer's Disease Neuroimaging Initiative (ADNI): MRI methods. J. Magn. Reson. Imaging 27, 685–691.

Klauschen, F., Goldman, A., Barra, V., Meyer-Lindenberg, A., Lundervold, A., 2009. Evaluation of automated brain MR image segmentation and volumetry methods. Hum. Brain Mapp. 30, 1310–1327.

Klöppel, 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.J., 2008. Automatic classification of MR scans in Alzheimer's disease. Brain 131, 681–689.

Klöppel, S., Stonnington, C.M., Chu, C., Draganski, B., Scahill, R.I., Rohrer, J.D., Fox, N.C., Ashburner, J., Frackowiak, R.S.J., 2009. A plea for confidence intervals and consideration of generalizability in diagnostic studies. Brain 132.

Magnin, B., Mesrob, L., Kinkingnéhun, S., Pélégrini-Issac, M., Colliot, O., Sarazin, M., Dubois, B., Lehéricy, S., Benali, H., 2009. Support vector machine-based classification of Alzheimer's disease from whole-brain anatomical MRI. Neuroradiology 51, 73–83.

McKhann, G., Drachman, D., Folstein, M., Katzman, R., Price, D., Stadlan, E.M., 1984. Clinical diagnosis of Alzheimer's disease: report of the NINCDS-ADRDA Work Group under the auspices of Department of Health and Human Services Task Force on Alzheimer's disease. Neurology 34, 939–944.

Moorhead, T.W.J., Gountouna, V.-E., Job, D.E., McIntosh, A.M., Romaniuk, L., Lymer, G.K.S., Whalley, H.C., Waiter, G.D., Brennan, D., Ahearn, T.S., Cavanagh, J., Condon, B., Steele, J.D., Wardlaw, J.M., Lawrie, S.M., 2009. Prospective multi-centre Voxel Based Morphometry study employing scanner specific segmentations: procedure development using CaliBrain structural MRI data. BMC Med. Imaging 9, 8.

Morris, J.C., Price, A.L., 2001. Pathologic correlates of nondemented aging, mild cognitive impairment, and early-stage Alzheimer's disease. J. Mol. Neurosci. 17, 101–118.

Mueller, S.G., Weiner, M.W., Thal, L.J., Petersen, R.C., Jack, C., Jagust, W., Trojanowski, J.Q., Toga, A.W., Beckett, L., 2005. The Alzheimer's disease neuroimaging initiative. Neuroimaging Clin. N. Am. 15, 869–877 (xi-xii).

Plant, C., Teipel, S.J., Oswald, A., Böhm, C., Meindl, T., Mourao-Miranda, J., Bokde, A.W., Hampel, H., Ewers, M., 2010. Automated detection of brain atrophy patterns based on MRI for the prediction of Alzheimer's disease. Neuroimage 50, 162–174.

Shuter, B., Yeh, I.B., Graham, S., Au, C., Wang, S.-C., 2008. Reproducibility of brain tissue volumes in longitudinal studies: effects of changes in signal-to-noise ratio and scanner software. Neuroimage 41, 371–379.

Stonnington, C.M., Tan, G., Klöppel, S., Chu, C., Draganski, B., Jack Jr., C.R., Chen, K., Ashburner, J., Frackowiak, R.S.J., 2008. Interpreting scan data acquired from multiple scanners: a study with Alzheimer's disease. Neuroimage 39, 1180–1185.

Vapnik, V.N., 1998. Statistical Learning Theory. Wiley Interscience, New York.

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, 1186–1197.

Whitwell, J.L., Przybelski, S.A., Weigand, S.D., Knopman, D.S., Boeve, B.F., Petersen, R.C., Jack Jr., C.R., 2007. 3D maps from multiple MRI illustrate changing atrophy patterns as subjects progress from mild cognitive impairment to Alzheimer's disease. Brain 130, 1777–1786.