Chapter 3.2  Automatic classification of AD and bvFTD based on cortical atrophy for single-subject diagnosis

Christiane Möller MSc,1 Yolande A L Pijnenburg MD PhD,2 Wiesje M van der Flier PhD,3,4 Adriaan Versteeg5, Betty Tijms PhD,6 Jan C de Munck PhD,4 Anne Hafkemeijer MSc,6,8 Serge A R B Rombouts PhD,6,8 Jeroen van der Grond PhD,8 John van Swieten MD PhD,5,9 Elise Dopper1,6,9 Philip Scheltens MD PhD,1 Frederik Barkhof MD PhD,3 Hugo Vrenken PhD,2,5,6 Alle Meije Wink PhD3 1Alzheimer Center & Department of Neurology, 2Department of Epidemiology & Biostatistics, 3Department of Radiology & Nuclear Medicine, 4Department of Physics & Medical Technology, 5Department of Clinical Genetics, Neuroscience Campus Amsterdam, VU University Medical Center, Amsterdam, the Netherlands, 6Institute of Psychology, Leiden University, 7Leiden Institute for Brain and Cognition, Leiden University, Leiden, the Netherlands, 8Department of Radiology, Leiden University Medical Center, Leiden, the Netherlands, 9Department of Neurology, Erasmus Medical Center, Rotterdam, the Netherlands.

Under review

Manuscript type: original research

Advances in Knowledge:
- Automated classifiers are able to discriminate between scans of patients with different forms of dementia and controls, based on gray matter (GM) patterns with high accuracy (75.3%-85.4%).
- Automated classifiers based on GM patterns can be used for single-subject diagnosis in an independent dataset with good to excellent diagnostic accuracy (AUC 0.81-0.95).
- Automated classifiers based on a generally available structural T1-weighted scans, make automated single-subject diagnosis more accessible and easy to use in daily clinical routine.

Implications for Patient Care:
- Machine learning-based categorization methods could improve the diagnostic process in the daily practice, especially in centers without experienced neuroradiologists.
- Machine learning-based categorization methods outperform classification based on a standard neuropsychological test battery.
- The application of automatic classification may be used for screening purposes in the future for high-risk groups.

Summary statement: Machine learning techniques are able to distinguish disease-specific GM atrophy between AD and bvFTD in a standard T1-weighted structural MRI scan for single-subject diagnosis.
Abstract

**Purpose:** Assessment of the diagnostic accuracy of a support vector machine (SVM) classifier for individual patients based on common T1-weighted gray matter (GM) images without extensive preprocessing, using two independent data sets.

**Materials and Methods:** The local Institutional Review Board approved the study. 84 patients with Alzheimer’s disease (AD), 51 patients with behavioral variant frontotemporal dementia (bvFTD), and 94 control subjects were divided into independent training (n=115) and test sets (n=114) with identical distributions of diagnosis and scanner type. The training set was entered in a SVM for disease specific predictions of diagnostic status between two groups based on GM patterns. Weight-values of each voxel for classification were extracted. We conducted discriminant function analyses to determine if the extracted weight-values could be used for single-subject classification in the independent test-set. Receiver operating characteristic (ROC) curves and area under the curve (AUC) were calculated for MRI classification and neuropsychological z-scores. Threshold for statistical significance was p<0.05.

**Results:** In the training set, the training accuracy of the SVM to discriminate AD from controls was 85.3%, bvFTD from controls 72.3%, and AD from bvFTD 78.7% (p<0.029). For single-subject diagnosis in the test set, the extracted weight-values discriminated 87.6% of AD and controls, 84.7% of bvFTD and controls, and 82.1% of AD and bvFTD correctly. ROC curves revealed good to excellent AUC (0.81-0.95; p<0.001). Machine learning-based categorization methods based on GM outperforms classification based on neuropsychological tests.

**Conclusions:** SVMs can be used in single-subject discrimination and help the clinician in making a diagnosis. SVMs can distinguish disease-specific GM patterns in AD and bvFTD from those of normal aging using common T1-weighted structural MRI scans.
Introduction
Alzheimer’s disease (AD) and behavioral variant frontotemporal dementia (bvFTD) are the most common causes of young onset dementia [1]. Clinical diagnostic criteria are available [2, 3], but the frequent overlap of the clinical symptoms associated with AD and bvFTD poses serious problems in the differential diagnosis. Various studies showed that cognitive tests are not accurate enough to discriminate AD from bvFTD, as bvFTD also show memory deficits and AD patients present with executive dysfunctions [4-9].

Magnetic resonance imaging (MRI), has been shown to be able to detect disease-specific macroscopic brain changes in an early disease stage. Several studies investigated gray matter (GM) changes to discriminate AD from bvFTD. These studies however typically report group-level differences for various brain structures and not for the single patient [10-12]. Furthermore, it has been shown that atrophy patterns of AD and bvFTD largely overlap e.g. frontal atrophy is seen in AD and hippocampal atrophy does not exclude a diagnosis of bvFTD; it appears in normal aging as well [9, 10, 13-15]. Moreover, especially in mild stages of the disease, cortical atrophy may not be visible by eyeballing.

Besides structural MRI, other imaging modalities, as positron emission tomography, functional or diffusion MRI, have been promising in the discrimination between AD and bvFTD. However, some of these techniques are invasive, time-consuming, require the availability of specialized scanners and are difficult to implement in a clinical setting without technical support [16, 17]. Therefore, a generally available, easy-to-use analytical imaging method that detects and quantifies more subtle changes of the brain at the single-subject level is needed to support the discrimination between AD and bvFTD.

In most memory clinics, patients are usually scanned once during dementia screening, with a standardized protocol generally including a T1-weighted 3-dimensional (3DT1) MRI sequence. This sequence is representative of the disease-specific structural changes, capable of providing similar information irrespective of different scanners and is easy to obtain. An automatic individual patient classification based on a structural 3DT1 scan at one time point could support the clinician’s diagnosis.

A support vector machine (SVM) is a machine learning technique that can categorize individual brain images by differentiation of images from two groups. These automated classifiers can be objective, quantitative and easy to implement and potentially satisfy the requirements of a diagnostic tool [18]. The available literature on SVM shares common limitations: Discriminating AD from FTD using the whole spectrum of FTD and not only the behavioral variant [16, 19-21] will be only representative for the language variants as their asymmetric atrophy will determine the classification accuracy [22]. Only using cross-validation of MR scans already known to the classifier in the discrimination of patients may result in biased estimates, especially when the sample size is small [20, 23, 24]. Therefore, we explore the diagnostic accuracy of a SVM for individual patients whose MR images belong to a test sample independent from the sample to train the classifier. The classification is based on a generally available T1-weighted GM image without extensive preprocessing. To increase generalizability we used MRI scans from different scanners and two independent data sets in a cross-sectional design.
Materials and Methods

Patients

In this study, we included 84 patients with AD, 51 patients with bvFTD, and 53 patients with subjective memory complaints who visited either the Alzheimer Center of the VU University Medical center or the Alzheimer Center of the Erasmus University Medical Center Rotterdam. All patients underwent a standardized one-day assessment including medical history, informant-based history, physical and neurological examination, blood tests, neuropsychological assessment, and MRI of the brain. Diagnoses were made in a multidisciplinary consensus meeting according to the core clinical criteria of the National Institute on Aging and the Alzheimer’s Association workgroup for probable AD [3, 25] and according to the International FTD Consortium criteria for bvFTD [2] based on the results of the one-day assessment as described above. Patients were diagnosed as having subjective memory complaints when they presented with memory complaints, but cognitive functioning was normal and criteria for MCI [26], dementia or any other neurological or psychiatric disorder known to cause cognitive decline were not met. To minimize center effects, all diagnoses were re-evaluated in a panel including clinicians from both centers. In addition, we included 41 cognitively normal controls, who were recruited by advertisements in local newspapers. They underwent an assessment including medical history, physical examination, neuropsychological assessment, and MRI of the brain comparable to the work-up of patients. Cognitively normal controls and patients with subjective memory complaints served both as controls to obtain a representative control group for the general population. Patients and controls were randomly split into equally sized independent training (n=115) and test sets (n=114) with identical distribution of diagnosis, age, sex and scanner type.

Level of education was rated on a seven-point scale [27]. Disease duration was calculated based on the time difference between date of diagnosis and the year patients caregivers noticed the first symptoms. The local medical ethics committee of both centers approved the study. All patients and controls gave written informed consent for their clinical data to be used for research purposes.

Neuropsychological assessment

To assess dementia severity we used the Mini-Mental State Examination (MMSE) [28] and to assess frontal lobe dysfunction we used the frontal assessment battery (FAB) [29]. Cognitive functioning was assessed using a standardized neuropsychological test battery covering the domains memory (immediate recall and delayed recall of Dutch version of the Rey Auditory Verbal Learning Test and total score of Visual Association Test A), language (Visual Association Test picture naming and category fluency (animals: 1 min)), attention (Trail Making Test part A (TMT A), Digit Span forward, and Letter Digit Substitution Test (LDST)), and executive functioning (Digit Span backwards, Trail Making Test part B (TMT B), letter fluency, and Stroop Color-Word card subtask [30]). For a detailed description of neuropsychological tests see Smits et al. [31]. For each cognitive task, z-scores were calculated from the raw test scores by the formula z=(x-μ)/σ, where μ is the mean and σ is the standard deviation of the performance of the control group of the whole dataset. The value z = 0 therefore reflects the average baseline test performance of the controls in a given
domain. Scores of TMT A, TMT B, and Stroop color-word card were inverted by computing \((-1)z\)-score, because higher scores imply a worse performance. Next, composite \(z\)-scores were calculated for each cognitive domain by averaging \(z\)-scores with the MEAN function in SPSS. Composite \(z\)-scores were calculated when at least one neuropsychological task was available in each cognitive domain.

**MR image acquisition and review**

Imaging at the VUmc was carried out on two 3T scanners (Signa HDxt, GE Healthcare, Milwaukee, WI, USA and TF PET/MR, Philips Medical Systems, Cleveland, Ohio, USA), using an 8-channel head coil with foam padding. Patients and controls from the Erasmus University Medical Center Rotterdam were all scanned at the Leiden University Medical Center (LUMC) on a 3T scanner (Achieva, Philips Medical Systems, Best, the Netherlands) using an 8-channel head coil. The scan protocol included a whole-brain near-isotropic 3DT1-weighted sequence. At the VUmc this was a fast spoiled gradient echo sequence (FSPGR; repetition time TR 7.8 ms, echo time TE 3 ms, inversion time TI 450 ms, flip angle 12°, 180 sagittal slices, voxel size 0.98x0.98x1 mm, total scan time 4.57 minutes) or a turbo field echo sequence (T1TFE; TR 7 ms, TE 3 ms, flip angle 12°, 180 sagittal slices, voxel size 1x1x1 mm, total scan time 6.14 minutes). At the LUMC this was a turbo field echo sequence (T1TFE; TR 9.8 ms, TE 4.6 ms, flip angle 8°, 140 transversal slices, voxel size 0.88x0.88x1.2 mm, total scan time 4.57 minutes). In addition, the MRI protocols included a 3D Fluid Attenuated Inversion Recovery (FLAIR) sequence, dual-echo T2-weighted sequence, and susceptibility weighted imaging (SWI) which were reviewed for brain pathology other than atrophy by an experienced radiologist.

**MR processing**

DICOM images of the 3DT1-weighted sequence from the Signa HDxt were corrected for gradient nonlinearity distortions. All scans were converted to Nifti format. The linear transformation matrix to MNI space was calculated using FSL-FLIRT [32] and used to place the image coordinate origin \((0,0,0)\) on the anterior commissure by using the Nifti s-form. The structural 3DT1 images were then analyzed using the voxel-based morphometry toolbox (VBM8; version 435; University of Jena, Department of Psychiatry) in Statistical Parametric Mapping (SPM8; Functional Imaging Laboratory, University College London, London, UK) implemented in MATLAB 7.12 (MathWorks, Natick, MA). Data of the training- and test set were preprocessed separately with VBM8 to avoid any bias [33]. The first module of the VBM8 Toolbox ("Estimate and Write") segments the 3DT1 volumes into GM, white matter (WM) and cerebrospinal fluid (CSF), applies a registration to MNI space (affine) and subsequently a non-linear deformation. The non-linear deformation parameters are calculated via the high dimensional DARTEL algorithm and the MNI 152 template. Remaining non-brain tissue was removed by the ‘light clean-up’ option. Tissue classes were normalized in alignment with the template with the ‘non-linear only’ option which allows comparing the absolute amount of tissue corrected for individual brain size. The correction is applied directly to the data, which makes a head-size correction to the statistical model redundant. In the second module, images were smoothed using an eight mm
full width at half maximum (FWHM) isotropic Gaussian kernel. All images were visually inspected after every processing step.

**Support vector machine – pattern recognition**

For the pattern recognition analysis we used the "Pattern Recognition for Neuroimaging Toolbox" (PRoNTo) [33], implemented in MATLAB 7.12 following the standard descriptions of the manual (http://www.mlnl.cs.ucl.ac.uk/pronto). As a first step – the training step – classification of the patients in the training set was done by using PRoNTo. Only the training set was used in PRoNTo to learn the patterns of the MRI scans of the 3 groups. The normalized, modulated, smoothed GM images of all subjects in the training-set were used as inputs. We used a custom mask, which was made by the mean of all smoothed GM segmentations and binarized at a threshold of 0.2. We used a binary SVM to classify (1) AD from controls, (2) bvFTD from controls, and (3) AD from bvFTD, all with leave-one out cross-validation. To estimate how much each voxel contributes to the classification task, we calculated the voxel-wise ‘discrimination maps’ [33]. The weights are the model parameters learned by the SVM projected back to the input space. The weight-value can be positive or negative, where a positive value represents a higher weighted average for class one, a negative value represents a higher weighted average for class two. These maps are shown for each classifier in Figure 1. For illustrative purposes the weight maps are thresholded at 30% of the maximum positive and negative weight values, in line with Mourao-Miranda [34]. The performance of the classifier in the training set was expressed in balanced accuracy (class-specific accuracy), sensitivity and specificity. Permutation testing was used to derive a p-value to determine whether the balanced accuracy exceeded chance levels (50%). Class labels were permuted 1000 times.

**Support vector machine – classification of new subjects**

As a second step – the test step –, we used the discrimination maps computed in PRoNTo on a new independent set of patients. To test whether the learned weight values from the training set could classify unseen subjects, we extracted the weights of every voxel of the weight map over all folds. These weights were then multiplied with the normalized, modulated, smoothed GM images of each subject from the independent test set. This integrated product of the weight map and the smoothed GM image per subject was averaged and transferred to SPSS for further analyses. This was done separately for the classification between AD-controls, bvFTD-controls, and AD-bvFTD.

**Statistical analyses**

SPSS version 20.0 for Windows was used for statistical analysis. Differences between groups were assessed using univariate analysis of variance (ANOVA), Kruskal-Wallis tests and χ² tests, where appropriate. Composite cognitive domain z-scores were compared using multivariate analysis of variance (MANOVA) with Bonferroni posthoc tests and age, sex, educational level and disease duration as covariates. To determine the discriminative power of the neuropsychological examination in the test set, we conducted a stepwise discriminant function analysis between AD and bvFTD wit leave-one out cross-validation. As predictors we entered the four z-domains memory,
language, attention and executive functioning. Additionally, we created a receiver operation characteristic (ROC) and calculated the area under the curve (AUC). To determine the performance of the SVM for single-subject classification in the independent test set, we conducted three discriminant function analyses between two groups with leave-one-out cross validation. As predictor we entered the averaged integrated product of the weight map and the smoothed GM image per subject from the corresponding classification of the SVM (e.g. for the discriminant analysis between AD and controls, the average integrated product from weight map "AD vs. controls" was used for prediction). Additionally, we created a receiver operation characteristic (ROC) and calculated the area under the curve (AUC). Statistical significance was set at p<0.05.

Results

Demographics

Demographic data for training and test set are summarized in Table 1. As the two sets were split based on age, sex, scanner type and diagnosis, the two sets did not differ in these variables. Furthermore, there were no differences in disease duration and MMSE.

Groups based on diagnosis within training and test set did not differ in level of education, sex, scanner type and disease duration (Table 2). AD and bvFTD patients had lower scores on the FAB compared to controls. AD patients were older than controls and had the lowest MMSE scores. The MMSE scores of both patient groups showed that patients were in a mild stadium of the disease.

In both datasets, AD patients performed worse on memory compared to bvFTD and controls, and both patient groups had lower scores on the language domain compared to controls. In the training set, AD and bvFTD patients performed worse on attention and executive functioning compared to controls, whereas in the test set only AD patients performed worse than controls on attention and had the worst scores in executive functioning compared to bvFTD and controls.

In the test set the four composite cognitive z-domains discriminated 81.0% of AD patients and controls correctly, with correct classification of 23 AD patients (62.2%) and 45 controls (95.7%). Memory had the highest loading. The ROC curve revealed an excellent AUC for the memory domain (0.95; p<0.001). For the discrimination between bvFTD and controls, the four composite cognitive z-domains discriminated 80.6% of all cases correctly, with correct classification of 15 bvFTD patients (60.0%) and 43 controls (91.5%). Language had the highest loading. The ROC curve revealed an excellent AUC for the language domain (0.92; p<0.001). For the discrimination between AD and bvFTD patients, the four composite cognitive z-domains discriminated 65.7% of all cases correctly, with correct classification of 33 AD patients (78.6%) and 11 bvFTD patients (44.0%). Memory had the highest loading. The ROC curve revealed a fair AUC for the memory domain (0.74; p=0.002).
Support vector classification

Training accuracy

Performance of the binary SVM is summarized in Figure 2. The classifier discriminated AD patients from controls with 85.3% balanced accuracy (p=0.001). Sensitivity for classification of AD patients was 83.3% (p=0.001) and specificity 87.2% (p=0.001).

A correct distinction between bvFTD patients and controls, was made in 72.3% (balanced accuracy, p=0.001). Sensitivity for classification of bvFTD patients was 61.5% (p=0.001) and specificity 83% (p=0.029).

The classifier discriminated AD from bvFTD with 78.7% balanced accuracy (p=0.001). Sensitivity for classification of AD patients was 88.1% (p=0.001) and specificity was 69.2% (p=0.001).

Generalizability of the classifiers for single-subject diagnosis

Results are summarized in Figure 3. In the independent test set the extracted weights discriminated 87.6% of AD patients and controls correctly, with correct classification of 36 AD patients (85.7%) and 42 controls (89.4%). The ROC curve revealed an excellent AUC for the extracted weights (0.95; p<0.001). For the discrimination between bvFTD patients and controls, the extracted weights predicted 84.7% of all cases correctly, with correct classification of 15 bvFTD patients (60%) and 46 controls (97.9%). The ROC curve revealed a good AUC for the extracted weights (0.87; p<0.001). For the discrimination between AD and bvFTD patients, the extracted SVM weights predicted 82.1% of all cases correctly, with correct classification of 39 AD (92.9%) and 16 bvFTD patients (64%). The ROC curve revealed a good AUC for the extracted weights (0.81; p<0.001).

Discussion

In this study we showed that it is possible to discriminate between patients with different forms of mild dementia and between dementia and controls, based on GM patterns with an automated classifier with high accuracy (75.3%-85.4%). Crucially, we have also demonstrated that automated classifiers can be used in single-subject diagnosis, as its diagnostic accuracy in an independent dataset was good to excellent (AUC 0.81-0.95).

The accuracy levels in our study are comparable with other studies using SVM in differentiation of AD and FTLD [17-20, 35]. The slightly higher accuracy values found in two other studies can be explained by the use of all subtypes of FTLD. It is possible that the specific atrophy patterns of the language variant of the FTLD spectrum increased the diagnostic power [18, 20]. Besides that, AD and FTD patients in those previous studies had lower scores on the MMSE than in our study, which are indicative of a later disease stage and presumably more disease related GM atrophy, facilitating discrimination. In our study, only patients with bvFTD were included, all patient groups had MMSE scores indicative for mild dementia, as well as comparable age and disease duration, rendering these confounding effects less likely. Furthermore, as the SVM identified brain regions, which contribute highly to the classification, are in
agreement with results of literature on GM atrophy in AD and bvFTD, we are confident that our results are valid.

The most important difference of this study in comparison to other studies using SVM is, that we used two independent sets of patients to test the generalizability and performance of the classifier. One set, the training set, was used in PRoNTo to learn the atrophy patterns and assign weights to the different patient groups. The learned weights from the training set were then used to classify new patients who were not used in the training phase. As each prediction is done independently for each test subject, this corresponds to a single-subject classification. To our knowledge, other studies using automated classifiers do not use an independent test set and therefore using the cross-validation framework within the classifier to classify patients. This has the disadvantage that classified patients have also been used to train the classifier and are ‘known’ to the machine. This could bias the results.

To our knowledge, there is only one other study that used a separate training and test set to evaluate the predictive power of a SVM [35]. The accuracy levels of our study are comparable and demonstrate the robustness of the performance of the automated classifier in independent datasets. Our study extends these results by testing larger sample sizes and fully independent training and test sets. We also used whole brain information instead of a region-of-interest (ROI) approach. The use of a ROI approach may lead to higher accuracy rates but [35, 36] restrain the implementation of the SVM approach in the daily practice, as extraction of ROIs is time-consuming. Another problem of the ROI based approach is the limited generalizability of the trained automatic classifier using single-center data when applied to new data sets acquired on different scanners.

When we compare our results with other classification studies using SVM with other modalities, we found slightly higher accuracy rates for SPECT (84-88%) [37], FDG-PET (80-82.9%) [16], and a study on FDG-PET with higher accuracy rates [38]. Even though PET imaging may reveal higher accuracy rates in some studies, it is still not available in most hospital and requires specialized technical support. It may be possible to increase SVM accuracy by adding biomarkers, however for this study we aimed to use only one modality, which is widely available mainstream hospitals. Furthermore, adding modalities does not necessarily lead to higher accuracy rates. A recent study added CBF measurements GM densities, and reported little improvements in diagnostic value for the combined data [17].

While MRI-based methods alone might not show best performance for all applications, our results indicate that the relative ease of data acquisition and applying automated diagnosis methods and the non-invasive nature of MRI make it a useful diagnostic measure for the discrimination between AD and bvFTD, especially in centers, where PET imaging and/or CSF measures are not available.

Although the clinical consensus criteria for bvFTD and AD have been improved, underlying pathology can only be predicted on clinical grounds with limited accuracy. Especially in mild stages of the disease, bvFTD and AD show overlapping clinical features and an enormous heterogeneity within both diagnoses [39, 40]. Various studies show that neuropsychological tests are not accurate enough to discriminate
AD from bvFTD [41], as bvFTD also show memory deficits and AD patients present with executive dysfunctions [5-9, 31, 42, 43]. MMSE and FAB have also been shown to not reliably differentiate AD from bvFTD [44]. We confirm these findings by showing that the diagnostic accuracy of the automated classifier based on MRI outperformed the diagnostic accuracy of neuropsychological tests.

Although several studies examining cross-sectional and longitudinal effects in volumes of brain regions have shown significant group differences between AD, bvFTD, and controls, the ability to detect structural patterns that enable accurate single-subject predictions will ultimately assist the diagnostic process in the daily practice. Our study focused on making automated single-subject diagnosis more accessible, taking into account multiple factors of the daily clinical practice, such as scans from different MR scanners and a control group consisting of healthy elderly controls and people with subjective memory complaints. Therefore, the method we described has potential in achieving more accurate dementia diagnosis in clinical practice.

It is possible that accuracy rates of the classifier may differ in patient samples from a different memory clinic because of heterogeneity within diagnoses. Patients of our test set were selected based on previous diagnoses, which may have improved classification accuracy. We do not expect lower accuracy rates in other memory clinics, as patients with an unclear clinical presentation are often referred to our clinic for a second opinion. At the other end of the spectrum, we predict that in a single-clinic setting with one scanner, and by using healthy controls not SMC subjects, accuracy rates will be even higher than in our study. Another possible limitation is that, we only used binary classifiers, which means that a test case not belonging to one of the two groups will be incorrectly assigned to one of these. Multi-class classifiers software for MRI are currently not widely available. Future releases of available machine learning models will facilitate multi-class classifications, and thereby improve the diagnostic usefulness. However, the current finding indicates that a binary classifier can assist the diagnostic process by predicting different dementia subtypes based on localized changes in GM density. It could be argued that valuable classification power is lost due to the whole brain approach and well-placed ROIs would improve categorization as some brain areas may be more informative about class membership than others. However, a disadvantage of a ROI based approach is that it might not be as generalizable as a classifier that takes into account whole brain information. Nevertheless, the aim of our study was to achieve optimal classification based on whole brain information to build an optimal classifier in a way that it could be easily used in daily practice.

Our results are encouraging and describe a role for computerized diagnostic methods in clinical practice. The analytical technique presented here is able to distinguish disease-specific GM atrophy between AD and bvFTD in a standard T1-weighted structural MRI scan for single-subjects. The differentiation based on GM outperforms the classification based on neuropsychological tests. A goal of machine learning based automated MR image analysis is higher sensitivity and specificity of ante-mortem diagnosis than is currently possible. A study by Klöppel et al. [19]...
showed that computer-based diagnosis is equal to or better than that achieved by radiologists. Together with our results, it is conceivable that in the future machine learning-based categorization methods could improve diagnosis, especially in centers without experienced neuroradiologists and without other supporting diagnostic measures. An important next step will be the application of automatic classification for screening purposes. If machine learning discrimination is sensitive enough to classify subtle differences, early screening of high-risk groups could be easily implemented.

Acknowledgements
The VUmc Alzheimer Center is supported by Alzheimer Nederland and Stichting VUMC Fonds. The study was supported by the Netherlands Organization for Scientific Research (NWO). Research of the VUmc Alzheimer Center is part of the neurodegeneration research program of the Neuroscience Campus Amsterdam. The clinical database structure was developed with funding from Stichting Dioraphte. This project is funded by the Netherlands Initiative Brain and Cognition (NIHC), a part of the Netherlands Organization for Scientific Research (NWO) under grant numbers 056-13-014 and 056-13-010. Dr. Wiesje van der Flier is recipient of the Alzheimer Nederland grant (Influence of age on the endophenotype of AD on MRI, project number 2010-002).
# Tables and Figures

## Table 1. Demographics

<table>
<thead>
<tr>
<th></th>
<th>Training set</th>
<th>Test set</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>N</strong></td>
<td>115</td>
<td>114</td>
</tr>
<tr>
<td><strong>Scanners</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GE</td>
<td>84 (73%)</td>
<td>83 (73%)</td>
</tr>
<tr>
<td>PETMR</td>
<td>9 (8%)</td>
<td>9 (8%)</td>
</tr>
<tr>
<td>Philips</td>
<td>22 (19%)</td>
<td>22 (19%)</td>
</tr>
<tr>
<td><strong>Diagnosis</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AD</td>
<td>42 (37%)</td>
<td>42 (37%)</td>
</tr>
<tr>
<td>bvFTD</td>
<td>26 (22%)</td>
<td>25 (22%)</td>
</tr>
<tr>
<td>Controls</td>
<td>47 (41%) (20 HC, 27 SMC)</td>
<td>47 (41%) (21 HC, 26 SMC)</td>
</tr>
<tr>
<td><strong>Age</strong></td>
<td>62.7 ± 7.5</td>
<td>62.7 ± 6.7</td>
</tr>
<tr>
<td><strong>Sex, f</strong></td>
<td>36 (31%)</td>
<td>36 (31%)</td>
</tr>
<tr>
<td><strong>MMSE</strong></td>
<td>25.2 ± 4.4</td>
<td>25.2 ± 4.3</td>
</tr>
<tr>
<td><strong>Disease duration, months</strong></td>
<td>42.6 ± 32.4</td>
<td>43.2 ± 31.0</td>
</tr>
</tbody>
</table>

Values presented as mean ± standard deviation or n (%). Differences between groups for demographics were assessed using ANOVA. Kruskall-Wallis tests and χ² tests where appropriate. Key: SMC: Subjective memory complaints; MMSE: Mini-Mental State Examination

## Table 2. Demographics within training- and test-set based on diagnosis

<table>
<thead>
<tr>
<th></th>
<th>Training set</th>
<th>Test set</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>N</strong></td>
<td>42</td>
<td>26</td>
</tr>
<tr>
<td><strong>Scanner</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GE</td>
<td>31 (74%)</td>
<td>18 (69%)</td>
</tr>
<tr>
<td>PETMR</td>
<td>3 (7%)</td>
<td>3 (12%)</td>
</tr>
<tr>
<td>Philips</td>
<td>8 (19%)</td>
<td>5 (19%)</td>
</tr>
<tr>
<td><strong>Age</strong></td>
<td>64.9 ± 7.1</td>
<td>62.1 ± 7.8</td>
</tr>
<tr>
<td><strong>Sex, f</strong></td>
<td>13 (31%)</td>
<td>9 (35%)</td>
</tr>
<tr>
<td><strong>Disease duration, months</strong></td>
<td>36.6 ± 22.1</td>
<td>44.6 ± 40.3</td>
</tr>
<tr>
<td><strong>Level of education</strong></td>
<td>4.7 ± 1.5</td>
<td>4.9 ± 1.3</td>
</tr>
<tr>
<td><strong>MMSE</strong></td>
<td>21.9 ± 4.5</td>
<td>24.8 ± 3.4</td>
</tr>
<tr>
<td><strong>FAB</strong></td>
<td>12.4 ± 4.0</td>
<td>12.7 ± 4.3</td>
</tr>
<tr>
<td><strong>Memory</strong></td>
<td>-4.5 ± 3.0</td>
<td>-1.4 ± 1.6</td>
</tr>
<tr>
<td><strong>Language</strong></td>
<td>-1.3 ± 1.3</td>
<td>-1.0 ± 1.1</td>
</tr>
<tr>
<td><strong>Attention</strong></td>
<td>-1.6 ± 1.5</td>
<td>-1.2 ± 1.6</td>
</tr>
<tr>
<td><strong>EF</strong></td>
<td>-1.9 ± 1.7</td>
<td>-1.7 ± 1.7</td>
</tr>
</tbody>
</table>

Values presented as mean ± standard deviation or n (%). Differences between groups for demographics were assessed using ANOVA. Kruskall-Wallis tests and χ² tests where appropriate. Cognitive composite z-domains were calculated of the available z-scores of each test by the MEAN function in SPSS and compared by MANOVA with Bonferroni posthoc tests and age, sex, educational level and disease duration as covariates.

Key: MMSE: Mini-Mental State Examination; FAB: Frontal Assessment Battery; EF: Executive functioning. a different from controls (p<0.05), b different from AD (p<0.05)
Figure 1. Weight maps for each classifier at a threshold of 30% of the maximum positive and negative weight values, superimposed onto a standard brain template from FSL (MNI152_T1_1mm) showing areas of the brain most vital for discriminating the two groups. Red-Yellow: negative values indicative for class 1. Blue-Light blue: positive values indicative for class 2. (A) AD vs. controls, (B) bvFTD vs. controls, (C) AD vs. bvFTD.
Figure 2. Performance of support vector machine classification of training set data. (A) AD vs. controls, (B) bvFTD vs. controls, (C) AD vs. bvFTD.
Figure 3. Discriminating performance in test-set of averaged integrated product of weight map from the training-set and smoothed GM images from the test-set. Scatterplots showing discrimination between two groups. The ROC curve illustrates the performance of the binary classifier. (A) 87.6% of AD patients and controls were classified correctly, with correct classification of 36 AD patients (85.7%) and 42 controls (89.4%). The ROC curve revealed an excellent AUC for the extracted weights (0.95; p<0.001). (B) 84.7% of FTD patients and controls were classified correctly, with correct classification of 15 bvFTD patients (60%) and 46 controls (97.9%). The ROC curve revealed a good AUC for the extracted weights (0.87; p<0.001). (C) 82.1% of AD and bvFTD patients were classified correctly, with correct classification of 39 AD (92.9%) and 16 bvFTD patients (64%). The ROC curve revealed a good AUC for the extracted weights (0.81; p<0.001).
Reference List


