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

<table>
<thead>
<tr>
<th>Journal:</th>
<th>Radiology</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manuscript ID:</td>
<td>Draft</td>
</tr>
<tr>
<td>Manuscript Type:</td>
<td>Original Research</td>
</tr>
<tr>
<td>Manuscript Categorization Terms:</td>
<td>Adults &lt; 1. SUBJECT MATTER, MR-Imaging &lt; 2. MODALITIES/TECHNIQUES, Head/Neck &lt; 4. AREAS/SYSTEMS, Brain/Brain Stem &lt; 5. STRUCTURES, Dementia &lt; 6. TOPICS, Technology Assessment &lt; 7. METHODOLOGY</td>
</tr>
</tbody>
</table>
Automatic classification of AD and bvFTD based on cortical atrophy for single-subject diagnosis

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.
- 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 with leave-one-out cross-validation to determine if the extracted weight-values could be used for single-subject classification in the independent test-set, created receiver operating characteristic (ROC) curves and calculated the area under the curve (AUC). Threshold for statistical significance was p<0.05.

**Results:** In the training-set, the 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).

**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.

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 morphometric gray matter (GM) changes to discriminate AD from bvFTD. These studies typically report group-level differences for various brain structures (4-6). 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 (7, 8).

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.

Automatic individual patient classifications 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 (9).

The available literature on SVM shares common limitations: Discriminating AD from FTD using the whole spectrum of FTD and not only the behavioral variant (7, 10-12) will be only
representative for the language variants as their asymmetric atrophy will determine the classification accuracy (13). Using only cross-validation in the discrimination may result in biased estimates, especially when the sample size is small (11, 14, 15). The use of a region-of-interest (ROI) approach or different imaging modalities for the discrimination (16, 17) 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.

Therefore, we explore the diagnostic accuracy of a SVM for individual patients 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, 18) 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 (19), 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. Before inclusion in the present study, they were screened for memory complaints, family history of dementia, drugs- or alcohol abuse, major psychiatric disorder, and neurological or cerebrovascular diseases. 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 and scanner type.

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.

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 (20) 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 (21). 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 classification

For the pattern recognition analysis we used the “Pattern Recognition for Neuroimaging Toolbox” (PRoNTo) (21), implemented in MATLAB 7.12 following the standard descriptions
of the manual (http://www.mlnl.cs.ucl.ac.uk/pronto). The normalized, modulated, smoothed GM images of all subjects in the training dataset 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’ (21). 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 (22). 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.

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 and multiplied them 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 $\chi^2$ tests, where appropriate.

To determine the performance of the SVM for single-subject classification in an 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. There were no differences in age, sex, scanner type, disease duration, MMSE score, and diagnosis. To get more insight in the different datasets and if patient groups were comparable, we examined the demographic data of training set and the test set based on diagnosis: There were differences in sex, scanner type and disease duration. In both datasets, AD patients were older than controls and had the lowest MMSE scores (Table 2).

Support vector classification

Training-set

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 scans of patients with different forms of 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 of patients and controls 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 (9, 11, 16, 23). 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 (9, 11). As it is clinically relevant to make a clear distinction between the language variant and bvFTD, a combination of these forms based on atrophy patterns is undesirable. 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 bvFTD was used and patient groups had a 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.

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 (16). 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 ROI approach, making our approach easier to implement in daily clinical routine.

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. The method we described, clearly has potential in achieving more accurate dementia diagnosis in clinical practice. Admittedly, the processing and preparation of a training dataset is time consuming, but once a training dataset is available, the spatial normalization and classification of any new scan can be performed instantaneously.

A possible limitation of this study is that we did not have post-mortem data available, so the possibility of misdiagnosis cannot be excluded. Nevertheless, we used an extensive standardized work-up and all AD patients fulfilled clinical criteria of probable AD, 42 patients fulfilled the criteria for probable bvFTD and nine for possible bvFTD. All diagnoses were re-evaluated in a panel including clinicians from both centers to minimize sample effects. 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.

Discrimination with pattern recognition models is based on whole brain information, rather than on individual regions or voxels. Therefore, all voxels contribute to the classification and no conclusions should be drawn about a particular subset of voxels in isolation. However, some brain areas may be more informative about class membership than others. It could be argued that valuable classification power is lost due to the whole brain approach and well-placed ROIs would improve categorization. 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 indicate that machine learning techniques can aid the clinical diagnosis. 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. 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 (23) 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. 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.
Reference List


### Table 1. Demographics

<table>
<thead>
<tr>
<th></th>
<th>Training set</th>
<th>Test set</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>115</td>
<td>114</td>
</tr>
<tr>
<td>Scanners</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>Diagnosis</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>Age</td>
<td>62.7 ± 7.5</td>
<td>62.7 ± 6.7</td>
</tr>
<tr>
<td>Sex, f</td>
<td>36 (31%)</td>
<td>36 (31%)</td>
</tr>
<tr>
<td>MMSE</td>
<td>25.2 ± 4.4</td>
<td>25.2 ± 4.3</td>
</tr>
<tr>
<td>Disease duration, months</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></td>
<td>AD</td>
<td>bvFTD</td>
</tr>
<tr>
<td>N</td>
<td>42</td>
<td>26</td>
</tr>
<tr>
<td>Scanner</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>Age</td>
<td>64.9 ± 7.1&lt;sup&gt;a&lt;/sup&gt;</td>
<td>62.1 ± 7.8</td>
</tr>
<tr>
<td>Sex, f</td>
<td>13 (31%)</td>
<td>9 (35%)</td>
</tr>
<tr>
<td>MMSE</td>
<td>21.9 ± 4.5&lt;sup&gt;a&lt;/sup&gt;</td>
<td>24.8 ± 3.4&lt;sup&gt;a,b&lt;/sup&gt;</td>
</tr>
<tr>
<td>Disease duration, months</td>
<td>36.6 ± 22.1</td>
<td>44.6 ± 40.3</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 \( \chi^2 \) tests where appropriate. Key: MMSE: Mini-Mental State Examination. <sup>a</sup> different from controls (p<0.05), <sup>b</sup> 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.
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.

176x173mm (96 x 96 DPI)
Figure 2. Performance of support vector machine classification of training set data. (A) AD vs. controls, (B) bvFTD vs. controls, (C) AD vs. bvFTD.

132x385mm (96 x 96 DPI)
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.

138x226mm (96 x 96 DPI)