Published in final edited form as:

Neuroimage. 2008 February 1; 39(3): 1186–1197.

Alzheimer's Disease Diagnosis in Individual Subjects using Structural MR Images: Validation Studies

Prashanthi Vemuri¹, Jeffrey L. Gunter¹, Matthew L. Senjem¹, Jennifer L. Whitwell¹, Kejal Kantarci¹, David S. Knopman², Bradley F. Boeve², Ronald C. Petersen², and Clifford R. Jack Jr¹

¹Department of Radiology, Mayo Clinic Rochester, MN

²Department of Neurology, Mayo Clinic Rochester, MN

Abstract

OBJECTIVE—To develop and validate a tool for Alzheimer's disease (AD) diagnosis in individual subjects using support vector machine (SVM) based classification of structural MR (sMR) images.

BACKGROUND—Libraries of sMR scans of clinically well characterized subjects can be harnessed for the purpose of diagnosing new incoming subjects.

METHODS—190 patients with probable AD were age- and gender-matched with 190 cognitively normal (CN) subjects. Three different classification models were implemented: Model I uses tissue densities obtained from sMR scans to give STructural Abnormality iNDex (STAND)-score; and Models II and III use tissue densities as well as covariates (demographics and Apolipoprotein E genotype) to give adjusted-STAND (aSTAND)-score. Data from 140 AD and 140 CN were used for training. The SVM parameter optimization and training was done by four-fold cross validation. The remaining independent sample of 50 AD and 50 CN were used to obtain a minimally biased estimate of the generalization error of the algorithm.

RESULTS—The CV accuracy of Model II and Model III aSTAND-scores was 88.5% and 89.3% respectively and the developed models generalized well on the independent test datasets. Anatomic patterns best differentiating the groups were consistent with the known distribution of neurofibrillary AD pathology.

CONCLUSIONS—This paper presents preliminary evidence that application of SVM-based classification of an individual sMR scan relative to a library of scans can provide useful information in individual subjects for diagnosis of AD. Including demographic and genetic information in the classification algorithm slightly improves diagnostic accuracy.

Keywords

support vector machines; classification; diagnosis; Alzheimer's

Corresponding author: Clifford R. Jack Jr, MD Department of Radiology Mayo Clinic and Foundation 200 1st St SW, Rochester, MN 55905 Fax: 507 284 9778 Tel : 507–284–2511 Email: jack.clifford@mayo.edu.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

INTRODUCTION

Diagnostic criteria for Alzheimer's disease (AD) are currently based on clinical and psychometric assessment. As the search for effective therapies to arrest or slow the progression of Alzheimer's disease (AD) intensifies, there is a need to develop better diagnostic tools. Development of such tools should also aid in measuring the efficacy of new therapies. Neuroimaging and specifically magnetic resonance imaging (MRI) has been shown to be a surrogate for early diagnosis of AD, particularly in subjects clinically classified as amnesic mild cognitive impairment (aMCI) which in most patients is a precursor to AD (Petersen et al., 1995). Degenerative histological changes occur long before the disease is clinically detectable (Gomez-Isla et al., 1996), perhaps decades, and thus imaging can be expected to help identify the early onset.

Studies demonstrating cross sectional inter-group differences associated with AD and MCI are now common place. Algorithmic complexity varies from conceptually simple measurement of volumes in manually or semi-manually delineated a priori regions of interest (ROI) (Barnes et al., 2004; Jack et al., 1999; Testa et al., 2004), to mathematically complex voxel-wise modeling of tissue density changes e.g. voxel or tensor based morphometry (VBM, TBM) (Bozzali et al., 2006; Freeborough and Fox, 1998a; Hirata et al., 2005; Thompson et al., 2007). By their nature, ROI based analyses are spatially limited and do not make use of all the available information contained in a 3-dimensional image data set. The most widely used voxel-based analytic techniques (Ashburner and Friston, 2000) are agnostic of disease-specific information. They are also designed to perform only group-wise comparisons and thus are unsuitable for evaluating the disease state of an individual subject. Because of these disadvantages, investigators have recently turned their attention towards multivariate analysis and machine learning based algorithms for distinguishing patients from cognitively normal (CN) subjects (Alexander and Moeller, 1994; Csernansky et al., 2005; Davatzikos et al., 2005; Fan et al., 2005; Freeborough and Fox, 1998b; Lao et al., 2004). These techniques use the entire 3D MRI data to form a disease model against which individual subjects may be compared.

Machine learning and computer-aided diagnostics (CAD) have been of growing interest in the field of medical imaging (Metz, 1999). Broadly speaking, a supervised machine learning algorithm is "trained" to produce a desired output from a set of input (training) data. As such, the trained algorithm may be treated as a "black box" encapsulating knowledge gleaned from the training data whose inputs are useful for producing the expected outcome. The supervised machine learning algorithm used in this paper is the support vector machine (SVM) (Vapnik, 1998).

The data used in this paper for each patient includes a structural MR (sMR) scan, ApolipoproteinE (APOE) genotype information and commonly available demographic details: age and gender. Demographic information was included because brain volume varies with both age and gender (Jack et al., 1997; Lemaitre et al., 2005; Shiino et al., 2006; Smith et al., 2006). In addition, old age is the strongest known risk factor for typical late onset AD (Evans et al., 1989). The information about APOE was added because there is a well established positive risk for AD associated with the presence of the $\epsilon 4$ allele while $\epsilon 2$ is protective (Bickeboller et al., 1997; Craft et al., 1998; Farrer et al., 1997). Various imaging studies have shown that APOE genotype, age, and gender co-vary with imaging measures of brain morphology (Bigler et al., 2000; Fleisher et al., 2005).

The aims of the paper were: (1) to optimize and train a model for AD vs. CN classification based on a library of MRI scans from 280 clinically well-characterized subjects as well as demographic and genetic information; (2) to measure the diagnostic sensitivity and specificity of the SVM algorithm in an independent test set of 50 AD and 50 CN. A hierarchical model

construction to classify AD and CN was followed, starting with sMR and then adding demographics and APOE information in a stepwise fashion. Our overall goal was preliminary validation of this algorithmic approach for the diagnosis of AD.

THEORY

SVM Classifier

Consider a set of m training datasets $\{\mathbf{x}_k, y_k\}$ ($k=1,2,\dots,m$ in our case $m=280$), where each training vector $\mathbf{x}_k$ has p input features and the desired output is $y_k \in \{-1, +1\}$ where $+1$ corresponds to AD and -1 corresponds to CN. A simple two dimensional example for our application with real data for 20 CN and 20 AD patients taken from the Mayo Clinic Alzheimer's Disease Research Center study is plotted in Fig. 1(a). The figure shows the expected difference between the two classes where AD patients have larger ventricular and lower hippocampal volume compared to CN. In this example, a line separating the classes is defined by $f(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} + w_0$, where $\mathbf{w}$ defines the boundary. The classification rule is given by $G(\mathbf{x}) = \text{sign}(f(\mathbf{x}))$, such that a test dataset $\mathbf{x}_k$ is declared an AD if $G(\mathbf{x}_k) > 0$ and CN if $G(\mathbf{x}_k) \leq 0$. In a SVM formulation, the solution of $\mathbf{w}$ satisfies $y_k \cdot (\mathbf{w} \cdot \mathbf{x}_k + w_0) \geq 0$ subject to the minimization of $\|\mathbf{w}\|^2$. In addition there is a soft-margin formulation which allows data that cannot be cleanly separated using a decision boundary to improve the generalization of the algorithm using C which is the user defined parameter/penalty term. The formulation as well as solution of SVMs can be found in (Cristianini and Shawe-Taylor, 2000; Hastie et al., 2001; Vapnik, 1998). The SVM classifier using soft-margin formulation ($C=100$) for the problem shown in Fig. 1(a) is presented in Fig. 1(b). The data points falling on or within the margin or the opposite side of the margin are called support vectors and are the only datasets that are required to define the boundary and are shown as filled symbols in Fig. 1(b).

To achieve better classification, the datasets are mapped into a higher dimensional space by a kernel function ϕ which is a non-linear transformation that translates the linear boundary into a non-linear boundary and exploits information about the inner product of data inputs. The most commonly used kernel functions are polynomial and Gaussian. Fig. 1(c) shows a third degree kernel SVM with $C = 100$ for the same example. The output of the polynomial kernel is shown in Fig. 1(d). For higher dimensional input vectors, the decision line becomes a linear hyperplane. The linear hyperplane SVM is known as a linear SVM or equivalently a SVM with a polynomial kernel of degree 1. In our initial experiments, both Gaussian as well as linear kernels gave comparable results because of which we decided to use the linear kernel throughout the paper.

It is important to note that the decision boundary is at SVM output zero. Input vectors outside the margin (i.e. -1 and 1) are correctly classified with high probability so there is little or no difference between input vectors yielding outputs of 3 and 5 – both should be classified as AD. Moreover, these far away points are not support vectors and were not used in constructing the surface.

MATERIALS AND METHODS

Subjects

One hundred ninety subjects that fulfilled clinical criteria for probable AD (McKhann et al., 1984) were age and gender matched to 190 CN subjects. By design there was no difference in age or gender distribution across subject groups. All subjects had been prospectively recruited into the Mayo Clinic Alzheimer's Disease Research Center (ADRC), or the Alzheimer's disease Patient Registry (ADPR), and were identified from the ADRC/ADPR database. These longitudinal studies include independent nursing, neurological, and psychometric evaluations. Each participant's information is reviewed by a panel of neurologists, neuropsychologists, and

research nurses to assign a consensus clinical diagnosis. The diagnosis of dementia was made according to DSM-III-R criteria (American Psychiatric Association, 1987). The diagnosis of AD was also made according to established criteria (McKhann et al., 1984). Potential subjects for this study that fulfilled clinical criteria for AD were excluded if they had hemispheric cerebral infarctions on MRI or any secondary clinical features suggesting contributions from non-AD dementias. The criteria for normal subjects were: no active neurological or psychiatric conditions; no cognitive complaints; normal neurological exam; no psychoactive medications; if a prior neurological or psychiatric condition was present, they must have returned to normal. If the subject developed cognitive impairment at any point in time during the available follow-up, they were excluded from the study. Note that clinical diagnostic classification of the subjects was made without knowledge of the results of imaging data analyses generated for this study. Cognitive status was assessed using the Mini-Mental Status Examination (MMSE) (Folstein et al., 1975) and the Clinical Dementia Rating scale (CDR) (Hughes et al., 1982). Two AD patients had MMSE score of 30; one of them perhaps due to the familiarity with the test (for several years). APOE genotype was obtained using standard techniques.

MRI Data Acquisition

MRI studies were performed on 12 different 1.5 Tesla GE-SIGNA MRI scanners (GE Medical Systems, Waukesha, WI) using a standard transmit-receive volume head coil over a period 11 years (1995–2006). All scanners undergo a standardized quality control calibration procedure every morning which monitors geometric fidelity over a 200 mm volume along all 3 cardinal axes, signal to noise ratio, and transmit gain. Despite these quality control features we also matched the date/year that the scans were performed between the AD and CN groups in an attempt to control for any temporal fluctuations associated with different scanner software platform versions. Subject images were obtained using a standardized imaging protocol that included a coronal T1-weighted 3-dimensional volumetric spoiled gradient echo (SPGR) sequence with the following scan parameters: FOV= 24×18.5 or 22×16.5 cm, in-plane matrix= 256×192 , 1.6 mm partition thickness and 124 contiguous partitions, flip angle= 25° , TR=23 sec, TE=6–10 ms, bandwidth= ± 16 kHz.

Division of Data into Training and Testing

When a classification algorithm is developed, it is important to know that the classifier works well enough to be useful for the application. It is easy to design an optimistically biased (low error, over-trained) algorithm. If, for example, the same data is used for model selection (i.e. optimization of C and γ or n in our case), model training, and validation, an obvious risk is the creation of a machine which does not generalize. Such a model may be of limited effectiveness for classifying novel data. To eliminate such circular logic for this study we performed validation of the algorithm on independent/unseen data. The 190 AD and 190 CN datasets were randomly divided into two sets of 140 AD, 140 CN and 50 AD, 50 CN. The training datasets need not be exactly age and gender matched if age and gender are given as input to the SVM. One set of 280 scans was used for training (feature selection, model selection and model optimization) and the second set of 100 scans was used to give a minimally biased estimate of the true diagnostic performance of the algorithm. The training and test participant characteristics are shown in Table 1.

Classification Algorithm for AD and CN

The algorithm development for the classification algorithm was done in a hierarchical fashion. Model I involved the development and validation of a SVM based algorithm for classification using information from sMR scans of 280 training datasets. The output of this model was a structural abnormality index (STAND)-score. We refer to the output of a model where the sMR

based information is adjusted by additional subject-specific information as adjusted structural abnormality index (aSTAND)-scores.

We expected to find age and gender dependence in the STAND-scores. Therefore, the Model II aSTAND-score is an extension of Model I by adding demographic information (age and gender) to the information from sMR scans. Since APOE genotype is known to increase ($\epsilon 4$ allele) or decrease ($\epsilon 2$ allele) risk of developing AD relative to the $\epsilon 3$ allele, this genetic information was included in Model III along with information from sMR scans and demographic information. This output is the Model III aSTAND-score. The development of the algorithm was done in a stepwise fashion to understand the improvement in the classifier by adding covariates. The algorithm was developed in MATLAB (The Math Works, Natwick, MA) and the SVM-KM Toolbox (Canu et al., 2005). Throughout the paper it will be assumed that the tissue input variables to the SVM are individually normalized to zero mean and standard deviation one for all 280 training datasets. The mean and standard deviation estimates which are used for normalizing the training dataset are considered to be part of the machine, saved and applied to the test datasets as well to ensure that they are scaled consistently.

MODEL I: MRI based tissue densities

Model I can roughly be divided into four main steps: First, the features of each image (i.e. tissue densities) were extracted for each individual structural MR scan. Second, a linear SVM was used for feature reduction to choose a subset of the voxel locations or patterns which most significantly differentiate AD from CN. The tissue density values at the subset of voxel locations were subsequently input to the third step in which SVM parameters are optimized. Optimization is such that the two classes of subjects, AD and CN, are differentiated with maximum possible accuracy. In the fourth and final step, the final classification Model I was obtained using these optimized parameters and information from adjacent voxels. The steps in Model I are also shown in Fig. 2 and detail of the main steps are given below.

1. Feature Extraction: Tissue Segmentation—SPM5 was used for tissue segmentation and normalization (<http://www.fil.ion.ucl.ac.uk/spm>). The SPM5 implementation includes image normalization, segmentation as well as bias correction (Ashburner and Friston, 2005). In order to reduce any potential normalization and segmentation bias across the disease groups, customized tissue probability maps (TPMs) were created. To create the customized TPMs all 380 images were normalized to the MNI template and segmented into grey matter (GM), white matter (WM) and CSF using the unified segmentation model in SPM5. Average probability maps of GM, WM and CSF were created from all 380 GM, WM and CSF probability maps. The average maps were smoothed using 8 mm full-width at half-maximum (FWHM) smoothing kernel to create the custom priors or TPMs. The segmentation and normalization was then re-done using the TPMs. All the images were re-sampled to an isotropic resolution of 1 mm at the end of the normalization and segmentation process to keep the resolution constant across the subjects. All images were modulated to keep the total tissue density in each subject before and after normalization constant. The modulated normalized segmented images were then smoothed with an 8 mm FWHM smoothing kernel.

2. Feature Reduction: Input to SVM—The smoothed, modulated GM, WM and CSF tissue densities of isotropic voxel sizes 1 mm were down-sampled to an isotropic voxel size of 8 mm by simple averaging. The resulting tissue density maps were $22 \times 27 \times 22$ voxels. Feature vectors were created from the down-sampled GM, WM and CSF maps. The 8 mm voxels that contained less than 10 % tissue density values and CSF in half or more of the images were not considered for further analysis. Also, an ROI was drawn on the custom template to remove the cerebellum from all the datasets. Let $\mathbf{x}_{GM,k}$ be the vector of remaining GM voxels for the k -th subject ($k=1,2,3,\dots,m$). Similarly vectors for WM, $\mathbf{x}_{WM,k}$, and CSF, $\mathbf{x}_{CSF,k}$, were

created. The total number of retained tissue densities values for each patient was of the order 10^4 . Blindly using all these as input in the development of the SVM algorithm leads to inferior performance as many locations lack discriminative power. Therefore, a feature selection step is typically used to winnow the features input to the SVM isolating features important for the classification (Guyon and Elisseeff, 2003).

The feature reduction in this paper was done using an SVM-based criterion in two steps:

1) A linear SVM-based criteria was used for variable selection independently on the GM, WM and CSF densities. In the feature selection step, a linear SVM (with $\log_{10} C = -4$) was used to ensure large margins in separating the classes. Using the training samples for each of the tissue densities $\{x_{\alpha,k}, y_k\} (k=1,2,\dots,\text{number of training samples}; \alpha=\text{GM, WM, CSF})$, the weights corresponding to a linear SVM for each tissue density, w_{α} , were estimated. To be clear, for initial feature selection an independent linear SVM was used for each tissue density. The weights, w_{α} , obtained represent the weight that should be applied to each individual voxel in each tissue density vector $x_{\alpha,k}$ to make a prediction of the class (+1 for AD and -1 for controls). The first dimensional reduction was obtained by considering only weights consistent with increased neurodegeneration in AD (i.e. keeping negative weights for brain tissue classes and positive weights for CSF classes, or lesser GM and WM density and greater CSF density in AD than CN). This step reduces the feature vectors to roughly half their original length. By removing voxels with positive weights for GM and WM and negative weights for CSF, most of the voxels which have increased brain tissue and decreased CSF in AD compared to CN were removed. The maps of positive and negative weight values for GM are shown in Appendix.

2) The remaining tissue density values were concatenated into a single vector $\mathbf{X}_k^{MR} = \{x_{GM,k}(w_{GM}0), x_{WM,k}(w_{WM}0), x_{CSF,k}(w_{CSF}0)\}$ for each patient k . The vectors for all 280 subjects were then input to the linear SVM and the weight vector $\mathbf{W}$ at each location was obtained. Each element of the vector defines the weight that needs to be applied to the tissue density input of each subject to make a prediction of the class (again +1 for AD and -1 for CN). The length of the vector $\mathbf{X}_k^{MR}$ was of the order of (5×10^3) , much larger than the number of training datasets. To mitigate the curse of dimensionality and improve the performance, the input vector size needed to be further reduced.

A weight-based ranking was applied (Mladeni et al., 2004). Vector elements associated with small magnitude weights $|\hat{W}|$ are relatively unimportant for classification; where $|\hat{W}|$ the absolute value of is $|\hat{W}|$ normalized to the observed maximum. Therefore, $|\hat{W}|$ can be used to rank features. A threshold t was applied retaining only the top ranked voxels for classification. This method works well because it simultaneously takes into account all the voxel locations and then ranks them independently based on their importance for classification. Therefore if the disease affects independent anatomic points, these regions would be detected by this technique.

Incorporating information from spatially contiguous voxels: As explained above, a threshold t is applied to retain only the top ranked voxels. Given the regional nature of neurodegeneration, it is logical to assume that voxels adjacent to those with top ranked voxels may also be useful for classification. In this context, a major disadvantage of SVM is that it does not utilize any spatial information. Therefore, the thresholding procedure for retaining the weight vector elements was modified. For each top ranked voxel that is detected as important for classification, the non-zero weight voxels in the surrounding 26 voxels (in a $3 \times 3 \times 3$ cube) were also included for classification. The modified weight vector $\mathbf{W}$ represents

the maximum absolute weight in the neighborhood of $3 \times 3 \times 3$ cube of the voxel and this weight vector is now thresholded using t to obtain the top ranked voxels for classification.

3. SVM Optimization: Parameter Selection—Cross-validation (CV), a widely used approach, was used to tune the parameters of the SVM model (C). In addition, the feature ranking threshold t was also determined using CV. A grid search for C and t was performed and for each set of parameters the prediction accuracy for the 280 training datasets was estimated using four-fold CV. The four-fold CV involved dividing the 140 AD and 140 CN in the training set randomly into four sub-groups of 35 AD and 35 CN. The algorithm was then trained using three sub-groups and tested on the remaining sub-group (Duda et al., 2000). This process was repeated by leaving each sub-group out and estimating the accuracy of the algorithm on the left-out sub-group. The average accuracy was taken as an estimate of the true accuracy of the algorithm. The grid search was performed over the range $\log_{10}C = [-5, -4.5, -4, \dots, +2.5, +3]$ and $t = [0.06, 0.08, \dots, 0.88, 0.90]$. The optimized C and t were then used to create the optimized SVM model using 280 sMR training images.

The division of the patients and controls into sub-groups can affect the performance of the algorithm. A key assumption of CV methods is that both the training sets and measurement sets are drawn from and adequately represent the same parent distribution. To estimate the effects of sampling, the CV method was performed ten independent times and the average sensitivity and specificity was used for the selection of C and t . The average sensitivity and specificity for Model I for different parameters are plotted in Figs. 3(a) and 3(b) respectively. Sensitivity and specificity can be traded off depending on the expected application of the method. For example, if an algorithm is to be used for screening then high sensitivity should be used to choose the operating point. But, if an algorithm is to be used to select subjects for therapeutics with a high risk profile, then high specificity should be given priority when choosing the operating point. For our algorithm, we choose a point where sensitivity was approximately equal to the specificity and overall accuracy (defined as the mean of sensitivity and specificity) was high. Optimal values of $\log_{10}C$ and t were found to be -2 and 0.66 respectively. The 4-fold CV accuracy of Model I was 85.8% .

4. Optimized SVM model—Having determined the model parameters, all 280 training datasets were used to construct information from the selected GM, WM and CSF voxels to train the final Model I SVM. The output of any new dataset k into this model gives a STAND-score.

MODEL II: Reduced set of tissue densities and demographic variables

The input to Model II is $\mathbf{X}_k^{MRD} = \left\{ \mathbf{X}_k^{MR} \left(\tilde{W} 0.66 \right) s_1 x_{k,age}; s_k, gender \right\}$, where $x_{k,age}$ and $x_{k,gender}$ represents the normalized age and gender information ($x_{k,gender} = 1$ for male and $x_{k,gender} = -1$ for female) respectively for subject k and $\mathbf{X}_k^{MR} \left(\tilde{W} 0.66 \right)$ indicates the tissue densities at the voxels that were obtained after thresholding in Model I. When heterogeneous inputs are input into SVM, optimal relative weighting for each type of input features is likely to improve the prediction accuracy (Palvidis et al., 2002). Therefore, the relative scaling of age (s_1) and gender (s_2) with respect to the tissue densities were determined empirically using the 4-fold cross-validation on the training data as 5 and 0.6 respectively. The optimal value for $\log_{10}C$ was found to be -2 .

The 4-fold CV step for this model involves running the entire Model II SVM on three fourths of the training data for a given set of parameters and testing it using the remaining one fourth. The 4-fold CV accuracy of Model II was 88.5% .

MODEL III: Reduced set of tissue densities, demographic variables and APOE

The input to Model III is $\mathbf{X}_k^{ALL} = \{\mathbf{X}_k^{MRD}; s_3 x_{k,APOE}\}$, where $x_{k,APOE}$ represents the vector containing both the APOE allele information for each subject k i.e. (2,2); (2,3); (3,3); (2,4); (3,4); (4,4). The linear kernel reinforces the well established fact that the risk of AD increases with the numeric value of the allele i.e. 4 confers greater risk than 3 while 2 confers less risk than 3 (Bickeboller et al., 1997; Craft et al., 1998; Farrer et al., 1997). The relative scaling of APOE (s_3) with respect to the tissue densities, age and gender and $\log_{10}C$ were determined empirically using the 4-fold cross-validation on the training data as 10 and -3 respectively. The 4-fold CV accuracy for these parameters for Model III was found to be 89.3%.

RESULTS

Model I

The components of the neighborhood weight vector $\tilde{\mathbf{W}}$ in GM, WM and CSF in 280 training datasets are shown in Fig. 4. Warm colors in Fig. 4, indicate greater vector weights. Note that the weight vector images are quite sensible biologically. Large GM weight vectors (indicating GM loss) are concentrated on the medial temporal lobe, temporal-parietal association cortex, posterior cingulate/precuneus and insula. WM loss is concentrated in temporal lobe (particularly in the WM of the entorhinal-parahippocampal gyrus) and parietal lobe. CSF space expansion is concentrated in medial-basal temporal areas, lateral ventricles, temporal-parietal areas and insula.

The voxels with maximum discriminative power between AD and CN in each tissue density map are found using the threshold $t=0.66$. These voxel locations are overlaid on the corresponding custom T1 template in Fig. 5 at the actual down sampled resolution. The color scale of the voxel was used to indicate the frequency of occurrence in multiple tissue maps. Yellow: voxel location used in all three tissues (GM, WM and CSF); orange: voxel location used in at least two tissues and red: voxel location used in one tissue. The total number of voxel locations were GM 124, WM 45 and CSF 68, yielding a 237 dimensional feature vector. Most of these locations are in the medial-basal temporal lobe. The voxel locations chosen in each tissue density map were used to obtain the optimized SVM model as presented in Step 4 of Model I.

In order to rigorously test this machine learning system, entirely novel data were presented to the Model I SVM. The “test data” sets of 50 CN and 50 AD patients were input to the previously trained model which uses the voxel locations shown in Fig. 5 to classify each case as AD or CN. On the novel test data, the algorithm performed with a sensitivity of 86 % and specificity of 86 % (Table 2). The corresponding ROC curves are presented in Fig 6(b).

Model II

The reduced tissue density features from Model I, plus age and gender were then presented to the Model II SVM. Using the 100 test data sets, the Model II aSTAND-scores performed with a sensitivity of 88 % and specificity of 90 % (Table 2 and Fig 6 (a)). Age as well as gender improved the classification accuracy for the following reason: the modulation step used to obtain each subjects GM, WM and CSF tissue densities in the customized template space preserves the original tissue quantity. On average, women have lower tissue density than men in a given voxel in the customized template space owing to their smaller head size. In addition, tissue density in a given voxel of older age subjects is typically lower than younger subjects since brain volume decreases with age.

Model III

The reduced tissue density features, age, gender and APOE genotype information for the 100 subjects with APOE genotype information were presented to the Model III SVM. The Model III aSTAND-scores performed with a sensitivity of 86 % and specificity of 92% (Table 2 and Fig 6(a)). A plot before and after adding the genotype information for the 100 patients is shown in Fig 7. Comparing model II and model III in Fig 7, Model II aSTAND scores of APOE 4 non-carriers (both CN and AD) were adjusted to more negative values with the addition of APOE genotype to the model, while Model II STAND scores of APOE 4 carriers were adjusted to more positive values.

Correlation with indices of clinical disease severity in AD subjects

We evaluated the degree to which Model III aSTAND-scores (Fig. 6(b)) correlated with the measures of disease severity in AD and CN subjects. The correlation between Model III aSTAND-score and MMSE was -0.7 with CDR summary score was 0.71 and with DRS (dementia rating scale) was -0.65 . The signs of the correlations are as expected: the MMSE and DRS scores decrease and the CDR increases with worse cognitive performance.

DISCUSSION AND CONCLUSIONS

Classification algorithms for the diagnosis of AD in single subjects using structural MR images are presented in this paper. Model I generates a structural abnormality index (STAND)-score based on a linear SVM can be summarized as a weighted sum of volumes in voxel locations that are more affected in AD than in CN in a manner consistent with disease related brain atrophy – i.e. brain shrinkage and CSF expansion. Unlike traditional ROI based volumetric techniques where predefined volumes of structures are used, this technique automatically detects the voxel locations that show AD specific changes using training datasets where AD and CN are already labeled. The voxel selection method proposed has two important features: 1) it examines all voxels simultaneously and selects those that are important for classification; 2) once a voxel is selected, the GM, WM voxels in the $3 \times 3 \times 3$ neighborhood that show tissue loss in AD relative to CN and CSF voxels in the neighborhood that show increase are preserved. As shown in Fig. 4, the regions obtained through the voxel feature selection process accurately mirror the known anatomic distribution of neuronal pathology in AD (Braak and Braak, 1994). This serves as an additional qualitative validation of the algorithm – i.e. it produces biologically sensible results. Neurofibrillary pathology begins and ultimately is most severe in the medial temporal lobe, particularly the entorhinal cortex and hippocampus. From there it spreads to the basal temporal lobe and paralimbic cortical areas such as the posterior cingulate gyrus and precuneus. Basal forebrain and the dorsal ponto-mesencephalic areas are also involved, the latter being the source of diencephalic and cortical cholinergic innervation which is depleted in AD. The larger number of GM locations compared to WM locations selected is in agreement with the nature of the disease; the primary pathological processes begin in GM while WM atrophy presumably results from dying back of axonal processes. The data in these 237 selected voxels are weighted by the degree of change due to the disease e.g. hippocampus are weighted the most in all the tissue density maps because these areas are affected earliest and most severely (Figs 4 and 5). The weighted sum of the volumes gives a measure of normality or abnormality in the brain structures ($\geq +1$ for the most abnormal and ≤ -1 for the most normal brain), which is labeled the STAND-score.

In Models II and III, the reduced set of tissue features selected as important in Model I are re-weighted based on additional inputs to account for differences due to demographics alone and then due to both demographics and APOE status respectively producing aSTAND-scores. In Fig. 7, the SVM assigned weights at a voxel for each model is plotted as a function of the original weight assigned to the same voxel in Model I. Each additional feature in models II

and III re-scales the original weight according to how the added feature affects the tissue density at that particular voxel. Since there is additional information available from the small set of features, this increases the overall classification accuracy. Another interesting observation is that the geometric margin of the SVM increased by 1.02 times in Model II and 2.75 times in Model III in comparison to the margin in Model I. Thus, the small set of features add incremental information and increase the overall accuracy as well as aid in better separation of AD and CN.

The aSTAND-score takes into account the additional subject-specific information for each individual subject and gives a number that represents the severity of dementia/degree of atrophy in the brain of that individual subject in comparison to the AD and CN cohorts of the same gender and genotype. Since the performance of Model III is the best of the three models considered, including demographic and genotype based risk factors appears useful. The strongest known genetic association with risk of late onset AD is that of APOE. The risk of developing AD increases with the number of copies of the $\epsilon 4$ allele. Fig. 8 illustrates that adding genotype information tends to shift all $\epsilon 4$ non-carriers downwards towards more “CN-like” values; while $\epsilon 4$ carriers are shifted upward toward more “AD-like” values. This is in near perfect agreement with the genotype specific adjustment of the output scores that would be anticipated on the basis of the known biological effect of APOE $\epsilon 4$. This result validates the ability of the algorithm to recognize the specific risk profile of APOE alleles during training; and then appropriately adjust the output for a per subject basis.

Note that for obtaining true unbiased estimates of the error, it is necessary to run the feature reduction step inside the CV loop, so that the CV error is representative of the true error (Varma and Simon, 2006). If the CV for selection of parameters is done after the feature reduction using all 280 training datasets, the CV error rate will be close to zero or overall accuracy will be near 100 % since all 280 datasets are ultimately used to both train and cross validate. To further emphasize this point, when the feature reduction was done outside the CV loop the CV error was approximately zero but was very large on the independent test dataset. Whereas, if the feature reduction was done inside the CV loop the error values are close to the error as measured on the independent test dataset as in our case.

The main objective of this paper is validation- i.e. does the algorithm sort subjects based on the sMR scans into CN vs. AD using the clinical diagnosis (i.e. cognitive performance) as the gold standard. The real clinical value will come in applying the algorithm to subjects who are not clearly demented or normal, but rather in subjects with border line cognitive performance. This paper is therefore a first step, a validation using clinical category as ground truth. Sound clinical validation is a necessary first step. Subsequent studies will evaluate clinical utility where the clinical/ psychometric evaluation does not provide an obvious answer. This algorithm could be used as a stand alone application or more likely, in conjunction with evaluation methods employed in standard clinical practice. Presently the operating point for the algorithm is set where sensitivity is approximately equal to the specificity and overall accuracy is as high, but this could be changed to give higher sensitivity if the algorithm were used for screening or high specificity if it were used to identify candidates for high risk therapeutics. In addition, the algorithm can be extended by adding other sources of information such as additional imaging modalities or additional clinical information. It is important to note that APOE genotype information in Model III adds only 0.8 % improvement in accuracy to Model II i.e. to the information from sMR based information and demographics. Hence, in a clinical setting it might suffice to have only the sMR scans and demographics because sMR scans and demographics encode most of the information that is additionally available through APOE genotype information.

Prediction of AD pathology at autopsy is not possible with 100% accuracy antemortem using clinical criteria i.e. all clinically diagnosed AD patients will not have AD pathology at postmortem and up to 30% of CN subjects will have enough AD pathology at post mortem to qualify for a pathological diagnosis of AD (Crystal et al., 1993; Hulette et al., 1998; Jicha et al., 2006; Katzman et al., 1988; Knopman et al., 2003; Morris and Price, 2001; Riley et al., 2002; Schmitt et al., 2000). Based on the estimates in (Evans et al., 1989) an average of 18.7% and 47.2% of the general population at an age (75–85) yrs and >85 years respectively have AD. This implies that roughly 28 % of our CN cases might develop clinical AD within the next 10 years. Since pathological changes precede clinically evident impairment by up to a decade or more, it is almost certain that some of our CN (perhaps as many as 1/3) with an average age of 77 years currently have some degree of AD-related tissue atrophy. Given that a clinical diagnostic gold standard is unlikely to exactly match findings at autopsy in every subject, the algorithm presented in this paper performs with reasonably good sensitivity and specificity.

The fact that the aSTAND-scores correlate well with accepted measures of disease severity (MMSE, CDR, and DRS) indicates that the algorithm not only discriminates AD and CN, but also scales reasonably well with disease severity in AD. It is important to note that the accuracy reported in this paper is rigorously estimated from novel test data not data recycled from machine training. Moreover, the subjects used to train and test this algorithm were derived from a community based recruitment mechanism. The performance estimates cited therefore are more relevant to the general population than they would be had the subjects been recruited only from memory clinics. The patients were also typical late-onset AD subjects, again making the results reported more generalizable than they would be had the patients consisted of rare early-onset familial cases. It is important to note that the data was collected from multiple scanners over a long period of time (~11 yrs) and with different software upgrades. Given the large number of factors that cause variance in the acquired data, we believe that the algorithm performs reasonably well and the reported error (validated on an independent dataset) reflects the true performance when a new case is input to the system.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGEMENTS

This study was supported by grants P50 AG16574, R01 AG11378 and R01 AG15866 from the National Institute on Aging, Bethesda MD, RR24151 K12 CTSA Mentored Career Development Program, the support of the Robert H. and Clarice Smith and Abigail Van Buren Alzheimer's Disease Research Program of the Mayo Foundation, U.S.A. The authors would like to thank Stephen D. Weigand and Scott Przybelski; Department of Health Science Research at Mayo Clinic for their statistical assistance in identifying cohorts for this study and Dr. Vladimir Cherkassky, University of Minnesota for discussions on SVM. We would like to thank the reviewers for their suggestions that have led to improving the methodology presented in the paper.

APPENDIX

REDUCTION OF VOXELS USING WEIGHTS OF A LINEAR SVM

In this part of the feature reduction step, the aim is to eliminate the voxels that are noisy and have a large variance among different subjects. The pre-processing step that is followed in this paper was estimating the weight vectors of a linear SVM and then removing the tissue density voxels with positive weights where positive weights correspond to voxels with biologically implausible results-i.e. greater atrophy in AD compared to CN.

In the absence of noise, one would expect only negative weights in tissue densities consistent with neuro-degeneration. In real world data, there are positive weight values as well since some

regions/voxels values across populations have a large variance in the data due to noise. The major reasons for the noisy data are 1) there is a large variability of brain morphology as well as registration and segmentation across populations that makes it difficult to conclude with confidence that there is any atrophy in one group (AD) versus the other (CN) and 2) there is no significant atrophy across the populations in some regions of the brain since AD pathology is not randomly distributed throughout the brain, but follows established topological patterns, 3) partial volume effects at the edges of the GM, WM voxels.

Excluding such voxels in the first pass serves two purposes: 1) reduces the dimensionality of the dataset and makes the algorithm more stable and less prone to segmentation and registration errors across populations when a new dataset is input; 2) a visual inspection of the positive and negative weights shown below of a linear SVM on the GM densities. Note that the positive weights occur in areas of the brain not heavily affected by AD, whereas large negative weights occur in heavily affected areas-e.g. medial temporal lobe. A histogram of the weights and the mean GM density at each of these weights of both AD and CN are shown below. It can be seen that the mean GM density in CN is consistently higher than GM density in AD when the voxels have positive weights whereas GM density of CN is approximately equal or slightly lower than AD in most of the voxels with positive weights.

NEGATIVE WEIGHT VOXELS (red in color)

In these voxels, there is significant decrease in the GM density in AD when compared to CN.

POSITIVE WEIGHT VOXELS (blue in color)

Mostly noisy measurements, the means of the GM density at these voxels across the populations are not significantly different between AD and CN. And the variance across populations makes it difficult to conclude with confidence that there is atrophy in AD when compared to CN. Looking at the map, excluding these points which are mainly in WM as well as in regions that are relatively unaffected in AD (e.g. motor cortex as well primary visual cortex) makes a lot of sense biologically. Hence, these were excluded at this point.

REFERENCES

- Alexander GE, Moeller JR. Application of the scaled subprofile model to functional imaging in neuropsychiatric disorders: a principal component approach to modeling regional patterns of brain function in disease. *Human Brain Mapping* 1994;2:79–94.
- American Psychiatric Association. Diagnostic and statistical manual of mental disorders. 3rd rev edn. American Psychiatric Press; Washington, D.C.: 1987.
- Ashburner J, Friston KJ. Voxel-based morphometry--the methods. *Neuroimage* 2000;11:805–821. [PubMed: 10860804]
- Ashburner J, Friston KJ. Unified segmentation. *Neuroimage* 2005;26:839–851. [PubMed: 15955494]
- Barnes J, Scallan RI, Boyes RG, Frost C, Lewis EB, Rossor MN, Fox NC. Differentiating AD from aging using semiautomated measurement of hippocampal atrophy rates. *Neuroimage* 2004;23:574–581. [PubMed: 15488407]
- Bickeboller H, Campion D, Brice A, Amouyel P, Hannequin D, Didierjean O, Penet C, Martin C, Perez-Tur J, Michon A, et al. Apolipoprotein E and Alzheimer disease: genotype-specific risks by age and sex. *Am J Hum Genet* 1997;60:439–446. [PubMed: 9012418]
- Bigler ED, Lowry CM, Anderson CV, Johnson SC, Terry J, Steed M. Dementia, quantitative neuroimaging, and apolipoprotein E genotype. *AJNR Am J Neuroradiol* 2000;21:1857–1868. [PubMed: 11110538]
- Bozzali M, Filippi M, Magnani G, Cercignani M, Franceschi M, Schiatti E, Castiglioni S, Mossini R, Falautano M, Scotti G, et al. The contribution of voxel-based morphometry in staging patients with mild cognitive impairment. *Neurology* 2006;67:453–460. [PubMed: 16894107]

- Braak H, Braak E. Morphological criteria for the recognition of Alzheimer's disease and the distribution pattern of cortical changes related to this disorder. *Neurobiol Aging* 1994;15:355–356. [PubMed: 7936061]discussion 379–380
- Canu, S.; Grandvalet, Y.; Guigue, V.; Rakotomamonjy, A. *Perception Systèmes et Information*. INSA de Rouen; Rouen, France: 2005. SVM and Kernel Methods Matlab Toolbox..
- Craft S, Teri L, Edland SD, Kukull WA, Schellenberg G, McCormick WC, Bowen JD, Larson EB. Accelerated decline in apolipoprotein E-epsilon4 homozygotes with Alzheimer's disease. *Neurology* 1998;51:149–153. [PubMed: 9674794]
- Cristianini, N.; Shawe-Taylor, J. *An Introduction to Support Vector Machines*. Cambridge University Press; 2000.
- Crystal HA, Dickson DW, Sliwinski MJ, Lipton RB, Grober E, Marks-Nelson H, Antis P. Pathological markers associated with normal aging and dementia in the elderly. *Ann Neurol* 1993;34:566–573. [PubMed: 8215244]
- Csernansky JG, Wang L, Swank J, Miller JP, Gado M, McKeel D, Miller MI, Morris JC. Preclinical detection of Alzheimer's disease: hippocampal shape and volume predict dementia onset in the elderly. *Neuroimage* 2005;25:783–792. [PubMed: 15808979]
- Davatzikos C, Ruparel K, Fan Y, Shen DG, Acharyya M, Loughead JW, Gur RC, Langleben DD. Classifying spatial patterns of brain activity with machine learning methods: application to lie detection. *Neuroimage* 2005;28:663–668. [PubMed: 16169252]
- Duda, RO.; Hart, PE.; Stork, DG. *Pattern Classification*. 2 edn. Wiley Interscience; 2000.
- Evans DA, Funkenstein HH, Albert MS, Scherr PA, Cook NR, Chown MJ, Hebert LE, Hennekens CH, Taylor JO. Prevalence of Alzheimer's disease in a community population of older persons. Higher than previously reported. *Jama* 1989;262:2551–2556. [PubMed: 2810583]
- Fan Y, Shen D, Davatzikos C. Classification of structural images via high-dimensional image warping, robust feature extraction, and SVM. *Med Image Comput Comput Assist Interv Int Conf Med Image Comput Comput Assist Interv* 2005;8:1–8. [PubMed: 16685822]
- Farrer LA, Cupples LA, Haines JL, Hyman B, Kukull WA, Mayeux R, Myers RH, Pericak-Vance MA, Risch N, van Duijn CM. Effects of age, sex, and ethnicity on the association between apolipoprotein E genotype and Alzheimer disease. A meta-analysis. APOE and Alzheimer Disease Meta Analysis Consortium. *Jama* 1997;278:1349–1356. [PubMed: 9343467]
- Fleisher A, Grundman M, Jack CR, Jr. Petersen RC, Taylor C, Kim HT, Schiller DH, Bagwell V, Sencakova D, Weiner MF, et al. Sex, apolipoprotein E epsilon 4 status, and hippocampal volume in mild cognitive impairment. *Arch Neurol* 2005;62:953–957. [PubMed: 15956166]
- Folstein MF, Folstein SE, McHugh PR. “Mini-mental state”. A practical method for grading the cognitive state of patients for the clinician. *J Psychiatr Res* 1975;12:189–198. [PubMed: 1202204]
- Freeborough PA, Fox NC. Modeling brain deformations in Alzheimer disease by fluid registration of serial 3D MR images. *J Comput Assist Tomogr* 1998a;22:838–843. [PubMed: 9754126]
- Freeborough PA, Fox NC. MR image texture analysis applied to the diagnosis and tracking of Alzheimer's disease. *IEEE Trans Med Imaging* 1998b;17:475–479. [PubMed: 9735911]
- Gomez-Isla T, Price JL, McKeel DW, Jr. Morris JC, Growdon JH, Hyman BT. Profound loss of layer II entorhinal cortex neurons occurs in very mild Alzheimer's disease. *J Neurosci* 1996;16:4491–4500. [PubMed: 8699259]
- Guyon I, Elisseeff A. An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research* 2003;3:1157–1182.
- Hastie, T.; Tibshirani, R.; Friedman, JH. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer-Verlag; New York: 2001.
- Hirata Y, Matsuda H, Nemoto K, Ohnishi T, Hirao K, Yamashita F, Asada T, Iwabuchi S, Samejima H. Voxel-based morphometry to discriminate early Alzheimer's disease from controls. *Neurosci Lett* 2005;382:269–274. [PubMed: 15925102]
- Hughes CP, Berg L, Danziger WL, Coben LA, Martin RL. A new clinical scale for the staging of dementia. *Br J Psychiatry* 1982;140:566–572. [PubMed: 7104545]
- Hulette CM, Welsh-Bohmer KA, Murray MG, Saunders AM, Mash DC, McIntyre LM. Neuropathological and neuropsychological changes in “normal” aging: evidence for preclinical

- Alzheimer disease in cognitively normal individuals. *J Neuropathol Exp Neurol* 1998;57:1168–1174. [PubMed: 9862640]
- Jack CR, Jr. Petersen RC, Xu YC, O'Brien PC, Smith GE, Ivnik RJ, Boeve BF, Waring SC, Tangalos EG, Kokmen E. Prediction of AD with MRI-based hippocampal volume in mild cognitive impairment. *Neurology* 1999;52:1397–1403. [PubMed: 10227624]
- Jack CR, Jr. Petersen RC, Xu YC, Waring SC, O'Brien PC, Tangalos EG, Smith GE, Ivnik RJ, Kokmen E. Medial temporal atrophy on MRI in normal aging and very mild Alzheimer's disease. *Neurology* 1997;49:786–794. [PubMed: 9305341]
- Jicha GA, Parisi JE, Dickson DW, Johnson K, Cha R, Ivnik RJ, Tangalos EG, Boeve BF, Knopman DS, Braak H, Petersen RC. Neuropathologic outcome of mild cognitive impairment following progression to clinical dementia. *Arch Neurol* 2006;63:674–681. [PubMed: 16682537]
- Katzman R, Terry R, DeTeresa R, Brown T, Davies P, Fuld P, Renbing X, Peck A. Clinical, pathological, and neurochemical changes in dementia: a subgroup with preserved mental status and numerous neocortical plaques. *Ann Neurol* 1988;23:138–144. [PubMed: 2897823]
- Knopman DS, Parisi JE, Salvati A, Floriach-Robert M, Boeve BF, Ivnik RJ, Smith GE, Dickson DW, Johnson KA, Petersen LE, et al. Neuropathology of cognitively normal elderly. *J Neuropathol Exp Neurol* 2003;62:1087–1095. [PubMed: 14656067]
- Lao Z, Shen D, Xue Z, Karacali B, Resnick SM, Davatzikos C. Morphological classification of brains via high-dimensional shape transformations and machine learning methods. *Neuroimage* 2004;21:46–57. [PubMed: 14741641]
- Lemaitre H, Crivello F, Grassiot B, Alperovitch A, Tzourio C, Mazoyer B. Age- and sex-related effects on the neuroanatomy of healthy elderly. *NeuroImage* 2005;26:900–911. [PubMed: 15955500]
- McKhann G, Drachman D, Folstein M, Katzman R, Price D, Stadlan EM. 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* 1984;34:939–944. [PubMed: 6610841]
- Metz, CE. Evaluation of CAD methods. . In *Computer-Aided Diagnosis in Medical Imaging*. In: Doi, K.; MacMahon, H.; Giger, ML.; Hoffmann, KL., editors. Elsevier Science; Amsterdam: 1999. p. 543-554.
- Mladeni, D.; Brank, J.; Grobelnik, M.; Milic-Frayling, N. Feature selection using linear classifier weights: interaction with classification models; *Proceedings of the 27th annual international ACM SIGIR conference on Research and development.*; (Sheffield, United Kingdom). 2004.
- Morris JC, Price AL. Pathologic correlates of nondemented aging, mild cognitive impairment, and early-stage Alzheimer's disease. *J Mol Neurosci* 2001;17:101–118. [PubMed: 11816784]
- Palvidis P, Weston J, Cai J, Noble SN. Learning Gene Functional Classifications from Multiple Data Types. *Journal of Computational Biology* 2002;9:401–411. [PubMed: 12015889]
- Petersen RC, Smith GE, Ivnik RJ, Tangalos EG, Schaid DJ, Thibodeau SN, Kokmen E, Waring SC, Kurland LT. Apolipoprotein E status as a predictor of the development of Alzheimer's disease in memory-impaired individuals. *Jama* 1995;273:1274–1278. [PubMed: 7646655]
- Riley KP, Snowden DA, Markesbery WR. Alzheimer's neurofibrillary pathology and the spectrum of cognitive function: findings from the Nun Study. *Ann Neurol* 2002;51:567–577. [PubMed: 12112102]
- Schmitt FA, Davis DG, Wekstein DR, Smith CD, Ashford JW, Markesbery WR. "Preclinical" AD revisited: neuropathology of cognitively normal older adults. *Neurology* 2000;55:370–376. [PubMed: 10932270]
- Shiino A, Watanabe T, Maeda K, Kotani E, Akiguchi I, Matsuda M. Four subgroups of Alzheimer's disease based on patterns of atrophy using VBM and a unique pattern for early onset disease. *Neuroimage* 2006;33:17–26. [PubMed: 16904912]
- Smith CD, Chebrolu H, Wekstein DR, Schmitt FA, Markesbery WR. Age and gender effects on human brain anatomy: A voxel-based morphometric study in healthy elderly. *Neurobiol Aging*. 2006
- Testa C, Laakso MP, Sabattoli F, Rossi R, Beltramello A, Soininen H, Frisoni GB. A comparison between the accuracy of voxel-based morphometry and hippocampal volumetry in Alzheimer's disease. *J Magn Reson Imaging* 2004;19:274–282. [PubMed: 14994294]

- Thompson PM, Hayashi KM, Dutton RA, Chiang MC, Leow AD, Sowell ER, G DEZ, Becker JT, Lopez OL, Aizenstein HJ, Toga AW. Tracking Alzheimer's disease. *Ann N Y Acad Sci* 2007;1097:183–214. [PubMed: 17413023]
- Vapnik, V. *Statistical Learning Theory*. NY, Wiley: 1998.
- Varma S, Simon R. Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics* 2006;7:91. [PubMed: 16504092]

Fig. 1.

(a) Plot of Hippocampal and Ventricular volume of 20 AD and 20 Controls. (b) Linear SVM and (c) Third degree polynomial kernel SVM to separate both the classes shown in Fig 1(a). (d) SVM output for the polynomial kernel. The circle data points represent control and square data points represent AD patients. The filled symbols are the support vectors. The decision boundary is marked as 0.

Fig. 2.

(a): Flow chart of the SVM based classification algorithm of AD and CN. Hierarchical representation of the three models evaluated in this paper. **(b)** Division of data for the algorithm training and then testing on independent samples of subjects. The numbering in the flow diagram indicates the order of operations.

Fig. 3. Four-fold CV for Model I: (a) Sensitivity and (b) Specificity as a function of C and threshold t on the feature rank.

Fig 4.
Interpolated weight vectors for GM, WM and CSF (before threshold is applied) overlaid on the corresponding custom template. Color scale indicates the weight i.e. the importance of the voxel location for classification. For display purposes only the scale from 0.1 to 1 was used to display the non-zero voxels in the custom template.

Fig 5. Anatomic patterns with maximum discriminative power between AD and controls are overlaid on the corresponding custom T1 template. Color scale used to indicate the occurrence of the voxel in multiple tissue maps. Yellow: voxel location used in all three tissues (GM, WM and CSF); orange: voxel location used in at least two tissues and red: voxel location used in one tissue only.

Fig 6.
(a) Model III aSTAND-Score. The circle data points represent CN and square data points represent AD patients. **(b)** ROC curve for the test data for all the three models.

Fig 7.
SVM weights of tissue voxels of each model plotted as a function of SVM weights of Model I. The blue circles indicate Model I, red diamonds Model II and black plus signs Model III weights at the same tissue voxel.

Fig 8. Output of the SVM categorized according to the number of APOE alleles for (a) Model II (b) Model III. The blue circles indicate the test CN and red squares indicate the test AD patients. The squares below the line zero indicate AD misclassified as CN and circles above zero indicate CN misclassified as AD.

Table 1

Patient characteristics at the time of the MRI scan.

	Training AD (n=140)	Training CN (n=140)	Test AD (n=50)	Test CN (n=50)
No. of females	60	60	16	17
Median (range) age, yrs.	78 (51, 98)	77 (48, 92)	78.5 (53, 93)	79 (57, 91)
No. of ApoE ε4 carriers	88 (62.8%)	37 (26.4%)	31 (62%)	13 (26%)
Median MMSE score (range)	22(3, 29)	29 (23,30)	20 (8, 30)	29 (26, 30)
Median CDR score (range)	5.5 (0.5–1.8)	0 (0–0.5)	5.25 (1–17)	0(0–0.5)

MMSE = Mini-Mental Status Examination; CDR = Clinical Dementia Rating.

Table 2

Training data 4-fold CV sensitivity, specificity and test data sensitivity, specificity and accuracy for each of the three models.

MODEL	Training data CV average accuracy ($\pm$ std)		Test data	
	Sensitivity	Specificity	Sensitivity	Specificity
I: sMR	85.6 ($\pm$ 1.2)	86.0 ($\pm$ 2.0)	86	86
II: sMR +Demographics	88.4 ($\pm$ 1.6)	88.6 ($\pm$ 1.3)	88	90
III: sMR + Demographics + ApoE	88.1 ($\pm$ 1.0)	90.5 ($\pm$ 1.0)	86	92