Spatially augmented LP Boosting for AD classification with evaluations on the ADNI dataset

Chris Hinrichs[a,b] * **Vikas Singh**[b,a] **Lopamudra Mukherjee**[c] **Guofan Xu**[d,e]
Moo K. Chung[b] **Sterling C. Johnson**[d,e]
and the Alzheimers Disease Neuroimaging Initiative †

Summary

Structural and functional brain images are playing an important role in helping us understand the changes associated with neurological disorders such as Alzheimer's disease (AD). Recent efforts have now started investigating their utility for diagnosis purposes. This line of research has shown promising results where methods from machine learning (such as Support Vector Machines) have been used to identify AD-related patterns from images, for use in diagnosing new individual subjects. In this paper, we propose a new framework for AD classification which makes use of Linear Program (LP) boosting with novel additional regularization based on spatial "smoothness". The algorithm formalizes the expectation that since the examples for training the classifier are images, the voxels eventually selected for specifying the decision boundary must constitute spatially contiguous chunks, i.e., "regions" must be preferred over isolated voxels. This prior belief turns out to be useful for significantly reducing the space of possible classifiers and leads to substantial benefits in generalization. In our method, the requirement of spatial contiguity (of selected discriminating voxels) is incorporated within the optimization framework directly. Therefore, unlike some of the existing methods, post-processing of the optimized classifier to ensure spatial smoothness is not required. To our knowledge, our method is the first to directly include such a prior belief in a predictive classification framework for brain image analysis. We perform and report on extensive evaluations of our algorithm on MR

*[a]Department of Computer Sciences, University of Wisconsin-Madison, Madison, WI 53706.
[b]Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison Madison, WI 53705.
[c]Department of Mathematical and Computer Sciences, University of Wisconsin-Whitewater, Whitewater, WI 53190.
[d]William S. Middleton VA Medical Center, Madison, WI 53792.
[e]Department of Medicine, University of Wisconsin-Madison Madison, WI 53792.
Email addresses: hinrichs@cs.wisc.edu (Chris Hinrichs), vsingh@biostat.wisc.edu (Vikas Singh), mukherjl@uww.edu (Lopamudra Mukherjee), mkchung@wisc.edu (Moo Chung), gxu@medicine.wisc.edu (Guofan Xu), scj@medicine.wisc.edu (Sterling Johnson)

†Data used in the preparation of this article were obtained from the Alzheimers Disease Neuroimaging Initiative (ADNI) database http://www.loni.ucla.edu/ADNI. As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. ADNI investigators include (complete listing available at http://www.loni.ucla.edu/ADNI/Collaboration/ADNI_Manuscript_Citations.pdf).

and FDG-PET images from the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset, and discuss the relationship of the classification output with the clinical and cognitive biomarker data available within ADNI.

1. Introduction

Alzheimer's disease (AD) is an irreversible neurodegenerative disorder and the leading form of dementia worldwide. Significant ongoing research is devoted toward establishing clinical biomarkers of the disease and for the development of new drugs. A number of studies have indicated that AD-related neurodegenerative change begins decades in advance of symptomatic disease (Johnson et al. (2006), Reiman et al. (1996), Sager et al. (2005), Thompson and Apostolova (2007)). This suggests that advanced imaging techniques may be able to provide insights into the early phases of the disease, long before symptoms of dementia are observable. Studies have shown that AD characteristics such as structural atrophy Jack Jr. et al. (2005), deToledo-Morrell et al. (2004), Thompson et al. (2001) and impaired metabolism Hoffman et al. (2000), Matsuda (2001), Minoshima et al. (1994) can be identified (in structural and functional images) in Mild Cognitive Impaired (MCI) and AD patients as well as at-risk individuals Small et al. (2000). In an effort to utilize such images in the diagnostic process, a number of groups are focusing on the development of better diagnostic tools using ideas from machine learning. Typically, available scans of a cohort of confirmed (or highly likely) AD cases and control subjects, are exploited as *training examples* for a machine learning algorithm. The algorithm seeks to optimize some statistical discrimination measure corresponding to the image data (such as specific brain regions) that is *most* indicative of whether the subject image is from the AD or control group. The optimized classifier may then be used to automatically classify (or give a confidence score for) images of individual subjects where the diagnosis is unknown.

The classification of structural/functional brain images using machine learning techniques has been applied in the context of specific diseases such as schizophrenia Shen et al. (2003), Demirci et al. (2008), Alzheimer's disease Davatzikos et al. (2008), Klöppel et al. (2008), Vemuri et al. (2008), Duchesne et al. (2008), Arimura et al. (2008), and obsessive-compulsive disorders Soriano-Mas et al. (2007). In the remainder of this section, we briefly review several interesting AD classification papers, and lay the groundwork for introducing our contributions. In Fan et al. (2008), Fan et al. (2008), Davatzikos et al. (2008), Davatzikos et al. (2008), Davatzikos and colleagues proposed a pattern recognition technique for classification using structural Magnetic Resonance (sMR) scans from the Baltimore Longitudinal Study of Aging (BLSA) dataset Shock et al. (1984). Their method uses a segmentation of the images into different tissue types such as gray matter (GM), white matter (WM) and cerebrospinal fluid (CSF) regions, followed by a warping

that preserves a measure of specific tissue types. This is followed by a feature selection step[1] where voxels are discarded (or selected) based on their statistical relevance for classification Sahiner et al. (2000). The processed data is then used to train a linear Support Vector Machine (SVM) Bishop (2006), which gives good accuracy on their dataset. Recently, Klöppel *et al.* Klöppel et al. (2008) also used linear SVMs to classify AD subjects from controls. In addition, they were also successful in separating AD cases from other types of dementia (Frontal Temporal Lobar Degeneration or FTLD) using whole-brain images. The authors reported a high level of accuracy ($> 90\%$) on confirmed AD patients, and less where post-mortem diagnosis was unavailable. Independently, Vemuri *et al.* Vemuri et al. (2008) showed promising evaluations on another dataset obtaining $88-90\%$ classification accuracy (also using linear SVMs). The authors observed that using *all* image voxels as features within their framework was counter-productive, as many of these voxels were in fact misleading their method into choosing inferior classifiers. To address these difficulties, the authors employed demographic and Apolipoprotein E genotype (APOE) data as auxiliary features in their model and adopted significant pre- (and post-) processing on the images. For instance, the authors down-sampled the data to $22 \times 27 \times 22$ voxels, effectively aggregating many voxels' outputs into a single voxel at lower resolution. Then, they discarded voxels with less than 10% tissue densities in half or more of the images, and finally used an ROI to remove the cerebellum. Feature selection was performed by training a linear SVM, and discarding zero-weight voxels, and then training a second linear SVM on the remaining voxels as the core learning algorithm. In order to compensate for SVMs' inability to directly consider spatial relationships between voxels, they pruned the weights from the second SVM by only retaining non-zero weights in a spatially contiguous $3 \times 3 \times 3$ neighborhood around top-ranked voxels.

A feature of some of the studies discussed above is the observation that exploiting the spatial structure of the data can lead to improvements in accuracy. The spatial structure refers to the fact that neighboring pixels are related, and the feature vector representation of the image volumes also inherits this dependency (between its coordinate values). The techniques in Davatzikos et al. (2008), Klöppel et al. (2008), Vemuri et al. (2008) employ a *stand-alone* classification model which does not directly permit a natural incorporation of such spatial information. As a result, such properties can only be utilized via an extensive set of pre- (or post-) processing steps. This suggests (as also noted in Vemuri et al. (2008)), that improvements may be possible by designing a classification model that leverages the spatial information *explicitly*. Motivated by this observation, we pursue a unified learning framework better suited to exploit inter-pixel dependency, a typical characteristic of learning problems where the input is in the form of images. Our new model uses this

[1] If each voxel is considered a "feature", feature selection involves the estimation of which features are useful for the problem at hand, and which subset of features can be safely discarded. Note that this procedure almost always involves *loss* of information, and the extent of this loss varies as a function of the specific problem and dataset being studied.

additional information as constraints or priors *during* the optimization. The calculated classifier, therefore, does not require post-processing, (such as pruning or redistributing weights) as it is intrinsically *aware* of the spatial information. By directly incorporating this prior, our model allows a more nuanced balance between the need to address accuracy, and the need to enforce spatial regularity on the learned classifier than is possible when such priors are applied as pre- or post-processing steps. We consider the issue of efficacy in detail in Section §4 by an extensive set of experimental results on baseline image scans from the Alzheimers Disease Neuroimaging Initiative (ADNI) dataset, (`http://www.loni.ucla.edu/ADNI/Data/`) consisting of a large set of Magnetic Resonance (MR) and (18fluorodeoxyglucose Positron Emission Tomography) FDG-PET images. We also report on analysis relating the classifier confidence to approximately twenty different cognitive biomarker data made available as part of the ADNI study.

The main contributions of this paper are: (1) we present a new predictive classification framework based solely on imaging data, which incorporates spatial regularity priors, which until now have been utilized in other frameworks by pre- or post-processing steps, and not included in the learning model. We present this new model in Section §2; (2) We have conducted exhaustive experiments on the ADNI dataset which we hope will allow objective comparisons between classification methods, in a way which closely matches real-world conditions. We present these results in Section §4; (3) We have analyzed anomalous subjects in the hope of identifying examples of heterogeneous AD pathology in the interest of better characterizing them, that we may improve future iterations of classification methods, and perhaps even to discover subjects who are not properly identified as AD or controls by the study. These results are presented in Section §5. We conclude the paper in Section §6.

2. Algorithm

We briefly discuss some characteristics of the problem in the following section before outlining our proposed algorithm in §2.2 - §2.4.

2.1 *Problem setting*

Consider a learning problem in a computer assisted diagnosis setting. The learning task is to utilize "training data" (where confirmed (or highly likely) diagnosis of the patients into diseased or healthy *classes* is available) to learn a classifier to be used for disease diagnosis. Now, if the data is in the form of images, the first step is to encode the image as a feature vector. Notice that an image volume of size $100^2 \times 100$ in the training set yields a 10^6-dimensional vectorial representation. However, the image datasets are in general relatively small (with at most several hundred images) due to practical difficulties in volunteer recruitment and associated cost issues. As a result, our feature space is sparse, and the classification model may very easily overfit and give poor generalization Bradley and Mangasarian (1998). One effect of this may be that

the "learnt" classifier performs well on training data but poorly on "test" images that we want to classify. This happens because the model learns the examples in the (relatively small) training set, without learning the underlying distribution. One way to address this issue is to utilize a larger training set but this may be infeasible in a variety of settings. On the other hand, if sufficient information about the data is given (e.g., distribution is Gaussian), we may be able to effectively employ such knowledge in datasets where such assumptions are valid. Another common strategy to address the high dimensionality is to explicitly utilize dimensionality reduction tools such as principal components analysis (PCA) Jolliffe (2002). This works well in some cases but PCA also makes a linearity and Gaussian assumption Jolliffe (2002), and consequently the 'signal' may be attenuated for non-Gaussian datasets. These ideas and well known results from learning theory Bishop (2006), Mitchell (1997) suggest that inclusion of effective priors (introducing bias) to regularize the classification model is a promising means of improving performance. We will investigate such priors in the form of the spatial structure of our data, i.e., the fact that feature vectors in the training set are encodings of images.

Our classification method utilizes the idea of boosting. Boosting seeks to "boost" the accuracy of weak (or base) classifiers – the general idea is to assign each classifier a weight and evaluate the goodness of their aggregate response Freund (1995), Mitchell (1997), Schapire (1990), Demiriz et al. (2002). The weak classifiers, when considered individually may have low predictive power. However, the premise is that if the weak classifiers' errors are uncorrelated, their combination gives a better approximation of the underlying "signal". Linear Programming boosting (LPboosting) is a boosting approach Demiriz et al. (2002), Grove and Schuurmans (1998) where the final classifier is learnt within a linear optimization framework but with a soft margin bias. That is, emphasis is placed on separating the feature space into two regions (where each region contains either positively or negatively labeled examples), such that the *margin* between the positive and negative regions is maximized. The model places a 1-norm penalty on the weights, which also has the effect of reducing many of the weights to zero[2]. Our model in §2.4 will build upon the LPboosting model with a set of additional priors. Weak classifiers in our case correspond to individual voxels (or features), which we discuss in more detail in the next section.

2.2 *Boosting Approach and Weak Classifiers*

Let us denote the set of images in the training set as $I = \{I_1, I_2, \cdots, I_n\}$ with known class labels $\mathbf{y} = \{y_1, y_2, \cdots, y_n\}$, $y_l \in \{+1, -1\}$. Without loss of generality, AD-positive patients (and controls) are denoted as -1 (and $+1$) respectively, and $I = I_{\text{AD}} \cup I_{\text{CN}}$ where I_{AD} (and I_{CN}) are the image sets of the AD (and control) groups respectively. The set of image volumes in I are spatially normalized to a common

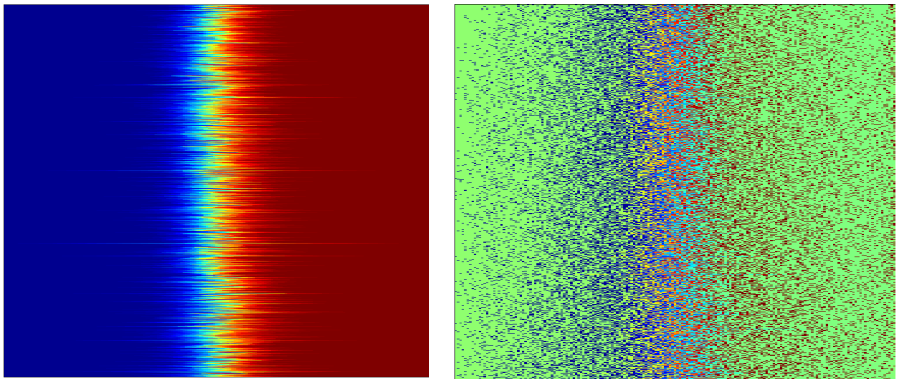
[2]In linear SVMs, the penalty is on the 2-norm of the weights, which places more emphasis on the *width* of the margin, in a *Euclidean* sense.

template space, as a first step. Therefore, a voxel located at (x, y, z) in one image "corresponds" to the voxel located at (x, y, z) in other images in I.

The proposed method makes no assumptions on a specific imaging modality. For instance, when utilizing T1-weighted MR scans, the images are segmented into gray matter (GM), white matter (WM), and cerebrospinal fluid (CSF), and probability maps of different tissue types are generated using standard techniques Ashburner and Friston (2000), Ashburner (2007), Friston et al. (1996). Either one of these quantities (voxel intensities) are then used to construct weak classifiers. Our weak classifier construction is partly motivated by voxel-wise group analysis methods. Each weak classifier at a voxel (x, y, z) tries to correlate variation at that voxel with the likelihood of AD diagnosis. Since AD is characterized by atrophy in specific brain regions, we should expect some weak classifiers to be more discriminative than others. Our algorithm will seek to automatically select and boost such classifiers. For notational convenience, in the remainder of this paper we will refer to voxels using a single index such as i, rather than (x, y, z).

Let us consider a list of the intensities of voxel i of all images in the training set, I. Now, given the class labels of individual images in the set, what is a good "thumb rule" if we were to use *only* this voxel for classification? Clearly, if this voxel is highly discriminative, the distribution is likely to be well separated (bi-modal). A *threshold* separating the two modes will work well for classifying any yet unseen test image (and also images in I), if the training set were sampled i.i.d. from the unknown but fixed underlying distribution. In general, however, the information from only one voxel will be far from the ideal setting above. Nonetheless, the labels on the training data can be used to determine a threshold. The classification induced by the threshold is the response of the weak classifier. Note that such a threshold may misclassify all examples in the region where the modes overlap. Fig. 1(b) shows that the weak classifiers give more incorrect outputs near the threshold, where there is more overlap between the modes, though they are also prone to errors even in the "safer" regions where their outputs have greater confidence. While such a predictor may be rather poor for many voxels, fortunately, we only require better accuracy ($> 50\%$ when there are two groups), and only on a subset of voxels.

The responses of the weak classifiers will populate a matrix, H of size $m \times n$, where m is the number of images and n is the number of classifiers (or voxels). We adopt a "soft" thresholding approach, i.e., the response of the weak classifier assigns a confidence score to the classification for each image rather than explicitly classifying it in either group. We use a logistic sigmoid function with a variable 'steepness' parameter ρ, and adjust the range to be $[-1, +1]$. We first choose a voxel specific threshold, τ_i, so that the response is negative (or positive) if less than (or greater than) the threshold. The τ_i value is calculated as the midpoint between the gray matter probabilities (or voxel intensities) means at voxel i for the I_{AD} and I_{CN} groups. Because a decline in GMP represents gray matter atrophy, a clinically consistent assumption here is



(a) (b)

Figure 1: Classifying/sorting image volumes in two classes using a single voxel. (a) Output of the 2000 most significant (voxel specific) weak classifiers using the given GMP data. Each row corresponds to a single weak classifier's output, individually sorted in non-decreasing order so that each column corresponds to an image volume in the training set. The image volumes are ordered differently in each row. Note that this is only the weak classifier output and does not correspond to ground truth. (b) If given access to ground truth labels, we can calculate the prediction error. Green regions denote entries where the sign of the weak classifier was correct, red and blue indicate false positive and false negative respectively. The prominent regions of misclassification suggest that individual weak classifiers are not very accurate.

that the control group mean, $\mu_{\text{CN}}(i)$ is greater than the AD group mean $\mu_{\text{AD}}(i)$ Fox and Schott (2004). Our choice of an adjusted logistic sigmoid curve is based on the fact that its first derivative closely approximates the Gaussian distribution, and conversely the value of the sigmoid (before adjustment) corresponds to the area under the Gaussian density function up to that point. This means that while the weak classifiers do not output actual probabilities, the level of confidence is related to the probability of class membership.

Let H_{ij} be the output of a weak classifier i (a certain voxel or feature) on image j.

$$H_{ij} = \frac{2}{1 + \exp\left(\tau_i - \rho \cdot I_j(i)\right)} - 1$$

where ρ is the "steepness" parameter, $I_j(i)$ is the GMP at voxel i in image $I_j \in I$, and the threshold is given as $\tau_i = (\mu_{\text{CN}}(i) - \mu_{\text{AD}}(i))/2$. We illustrate the observed steepness as a function of ρ in Fig. 2.

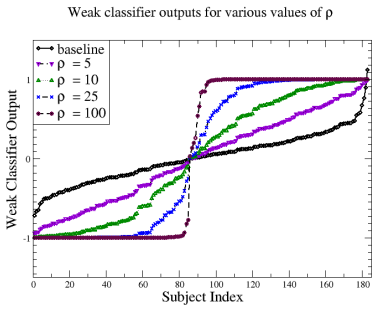


Figure 2: Weak classifier outputs as a function of ρ values.

2.3 *Spatial Constraints*

A characteristic of the problem, as discussed in §2.1 is that the feature vectors are representations of image data. This results in a certain dependency between the feature vector coordinates, and also the weak classifiers, see Fig. 3. This property of the data can be leveraged to introduce a bias (or prior) in the classification which has an advantage of constraining the complexity (expressiveness or degree of freedom) of possible classifiers, encouraging better generalization. The classifier consists of a set of weights on weak classifier outputs to define a separating hyperplane. We enforce spatial regularity by requiring that the weights assigned to neighboring weak classifiers should be similar. Such a spatial regularizer also has the benefit that it avoids selecting individual spatially isolated voxels. Rather, it prefers spatially localized 'regions' – a desirable characteristic since isolated voxels are seldom clinically relevant, and markers of AD, if observable in the image, must be spatially localized.

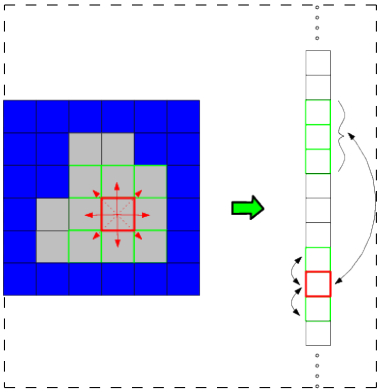


Figure 3: Spatial relationships (neighborhood) in the original image space are inherited as pair-wise relations in the feature-vector.

2.4 *Classification model*

Our final optimization model is given as

$$\begin{aligned} \min_{\mathbf{w},\xi_i} \quad & \mathbf{w}^T\tilde{\mathbf{p}} + C\sum_i \xi_i + D\sum_{j\sim k} t_{jk} \\ s.t. \quad & y_i\mathbf{w}^T H_i + \xi_i \geq 1 \quad \forall i \\ & w_j - w_k - t_{jk} \leq 0 \quad \forall j \sim k \\ & w_k - w_j - t_{jk} \leq 0 \quad \forall j \sim k. \end{aligned} \tag{1}$$

The vector $\mathbf{w}$ defines a separating hyperplane, and the term $\mathbf{w}^T H_i$ is the projection of example i onto the vector normal to the hyperplane. The sign of this quantity determines the side of the hyperplane onto which example i falls (this is zero for points *on* the decision boundary). By specifying that it must be at least ± 1, we place an emphasis on the *margin*. The term given by the product of $\mathbf{w}^T H_i$ and by y_i (the given class label of example i: $\{+1, -1\}$) imposes a lower bound of $+1$ for both positive and negative examples. For cases where the data are not *linearly separable*, a set of "slack variables" ξ_i are used to compensate for examples which cannot be placed on the correct side of the decision boundary. The penalty on the slack variables (second term in objective) ensures that the hyperplane will be chosen so that it correctly classifies as many examples as possible. The 1-norm penalty on weights $\mathbf{w}$ used here has the effect of selecting a *sparse* set of the most discriminative voxels, i.e., a large percentage of weak classifiers are suppressed and assigned a

zero weight. Sparsity of relevant features allows for an easier clinical interpretation as the output consists of only a few but highly discriminative (highly weighted) localized regions. This is preferable to diffuse outputs where many voxels are assigned non-zero weights, since it is difficult to analyze where significant structural variations are present, and which regions are most discriminative. The 1-norm penalty also serves a feature selection purpose Fung and Mangasarian (2004), Gaul and Ritter (2000), in many applications. The vector, $\tilde{\mathbf{p}}$, represents the training set error rate of every weak classifier (first term in objective). By adjusting the penalty on each weight w_j relative to its training set error rate, we allow weak classifiers with greater accuracy to be given slightly greater weight. The auxiliary variables t_{jk} represent the absolute difference between weights on neighboring voxels j and k (indicated as $i \sim j$). These variables are similarly penalized, which leads the optimizer to choose a separating hyperplane whose weights correspond to a set of spatially coherent voxels. We note that if $t = |w|$ then $t \geq w$ and $t \geq -w$ must both hold simultaneously. Thus, $t_{jk} = |w_j - w_k|$. The parameter C controls the amount of emphasis placed on *training set accuracy* relative to *margin width*. The emphasis on *spatial regularity* is similarly controlled by D. The model benefits from a good choice of regularizers, C and D. In Model (1) above, we observed that in practice $D > 10 \cdot C$ is a reasonable choice to sufficiently enforce the neighborhood constraints. Finally, we note that the linear program in (1) can be *optimally* solved efficiently in polynomial time. Once the solution is obtained, the weights $\mathbf{w}$ can be interpreted as the coefficients of a separating hyperplane in the feature space. We use this hyperplane *directly* as our classifier, *and no additional post-processing is required.*

3. Materials and Methods

3.1 *Data set*

The evaluations of our algorithm focus exclusively on the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (`www.loni.ucla.edu/ADNI`). The ADNI was launched in 2003 by the National Institute on Aging (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the Food and Drug Administration (FDA), private pharmaceutical companies and non-profit organizations, as a $60 million, 5-year public-private partnership. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimers disease (AD). Determination of sensitive and specific markers of very early AD progression is intended to aid researchers and clinicians to develop new treatments and monitor their effectiveness, as well as lessen the time and cost of clinical trials. The Principle Investigator of this initiative is Michael W. Weiner, M.D., VA Medical Center and University of California San Francisco. ADNI is the result of efforts of many co-investigators from a broad range of academic institutions and private corporations, and

	controls (mean)	controls (s.d.)	AD (mean)	AD (s.d.)
Age	75.81	4.46	76.11	6.99
Gender(M/F)	59/35		53/36	
MMSE	29.01	0.78	21.71	3.04
ADAS	10.14	4.26	32.32	9.10

Table 1
Demographic and neuropsychological characteristics of the study population. The FDG-PET population is a subset of this population.

subjects have been recruited from over 50 sites across the U.S. and Canada. The initial goal of ADNI was to recruit 800 adults, ages 55 to 90, to participate in the research – approximately 200 cognitively normal older individuals to be followed for 3 years, 400 people with MCI to be followed for 3 years, and 200 people with early AD to be followed for 2 years.

The baseline data used here includes:

1. T1-weighted Magnetic Resonance(MR) images: using both gray matter and white matter probability maps (for classification).

2. 18fluorodeoxyglucose-Positron Emission Tomography (FDG-PET) images (for classification).

3. Cognitive and neuropsychological biomarker data (only used to demonstrate that the classification confidence is correlated with known relevant biomarkers, and is *not* used in classification.)

Our experimental evaluations utilized a portion of the ADNI database. The subjects included 183 individuals (112 males, 71 females) in the T1-weighted MR image set, and 149 individuals (88 males, 61 females) in the FDG-PET image set, and some subjects were common to both populations. Of the 183 individuals in the MR population, neuropsychological test scores were available for 182 subjects, and semi-automatically derived brain region volumes from the Anders Dale lab at UCSD were available for 126 subjects. We will refer to this as UCSD. Similarly derived hippocampus volumes from Colin Studholme at UCSF were also available for 135 subjects. We will refer to this as UCSF. A summary of demographic and neuropsychological data are presented in Table 1.

3.2 *Preliminary Data Processing*

Image processing of the T1weighted images was performed using voxel-based morphometry (VBM) toolbox in Statistical Parametric Mapping software (SPM, `http://www.fil.ion.ucl.ac.uk/spm`). Segmentation in SPM employs a unified approach, combining: segmentation of the original anatomical images into gray matter (GM), white matter, and cerebrospinal fluid images; normalization (12-parameter affine transformation and non-linear deformation with a warp frequency cutoff of 25) of the segmented images to the

Montreal Neurological Institute template (MNI); and bias correction, in one iterative process. A modulation step was also employed, which scales the final GM images by the amount of contraction required to warp the images to the template. The final result is GM volume maps for each participant, where the total amount of GM remains the same as in the original images. Finally, the normalized maps were smoothed using an 8-mm isotropic Gaussian kernel to optimize signal to noise and facilitate comparison across participants. Analysis of gray matter volume employed an absolute threshold masking of 0.1 to minimize the inclusion of the white matter in analysis.

4. Experiments and Results

We first cover our results on the T1-weighted MR images, before moving to accuracy evaluations with FDG-PET image data in §4.2. We then discuss the relation between the classification confidence and various biomarkers in §4.3. Finally, we describe our solution to several implementation issues in our experiments.

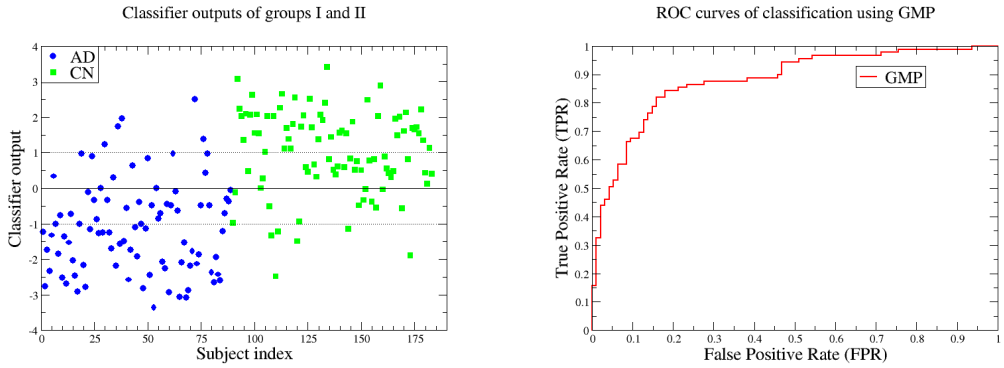
4.1 *MR image data*

Our evaluations with the ADNI MR image data were performed using leave-two-out cross-validation. In these experiments we used only the gray matter probability maps (GMPs). We also used GMPs together with the white matter probability maps (WMPs) for training and classification, however this did not yield any significant improvements. The classification accuracy was determined by calculating the number of 'test' images on which the classifier's class prediction (AD or CN) was incorrect; we report on the mean of these errors for both the above mentioned cases. The classification accuracy of the model using GMPs was 82%, and the sensitivity (and specificity) was 85% (and 80%). The results are summarized in Table 2, and suggest that the proposed technique works well for the AD classification task using MR image data.

Recall that in addition to a class label for the test images, the algorithm may be asked to report a classification confidence for each case (i.e., prediction), the summary of these results are shown in Fig. 4. In Fig. 4(a) we see that the classifier output on AD cases is concentrated between 0 (closest to the classification boundary) and -3 (farthest from the classification boundary), but the model incorrectly classifies some cases (which account for the misclassifications in the accuracy reported in Table 2 below). Fig. 4(b) shows the Receiver Operating Characteristic (ROC) plot where the area under the curve (AUC) of 0.8789 suggests a high predictive accuracy.

Data set	Accuracy	Sensitivity	Specificity	Area under ROC
GMP	82%	85%	80%	0.8789
FDG-PET	80%	78%	78%	0.8781

Table 2: Results of classification experiments on ADNI image data. One set of experiments were conducted with Gray Matter Probabilities (GMP) derived from T1 weighted MR images as input. The other set of experiments were conducted with FDG-PET images.



(a) (b)

Figure 4: (a) Classifier's output for test images on the MR population. (b) ROC curves on the MR population.

An important component of our experiments was to evaluate the relative importance of various brain regions in terms of specifying a good classifier, and whether these regions are consistent with clinically accepted distribution of AD-specific pathology. Figure 5 shows our results for the entire MR population. We see that the selected voxels (or weak classifiers) are concentrated in the hippocampus and parahippocampal gyri, but that there are also some voxels in the medial temporal lobe bilaterally, and scattered in other regions. We find these results encouraging because the selected regions are all known to be affected in AD patients.

4.2 *FDG-PET image data*

We applied our algorithm to the FDG-PET scans from the ADNI dataset as well. In all, there were 149 subjects in the MR population who also had FDG-PET scans. We call this group the FDG-PET population. Our method obtained 80% classification accuracy on the FDG-PET population, The specificity was 78%, and the sensitivity was 78%, while the area under the ROC curve was 0.8781 as shown in Table 2.

Figure 6(a) shows the output of our classifier on the 149 subjects of the FDG-PET population. Similarly to the MR population, most of the AD subjects are concentrated between -1 and -2, (and similarly the CN subjects are concentrated between 1 and 2, while some subjects were misclassified. Again, the area under the ROC curve in Fig. 6(b) is an indication of the high accuracy of this method.

We also evaluated the brain regions selected by our algorithm in the experiments utilizing FDG-PET scans in terms of their relevance to AD-specific pathology. From Fig. 7 we can see that the posterior

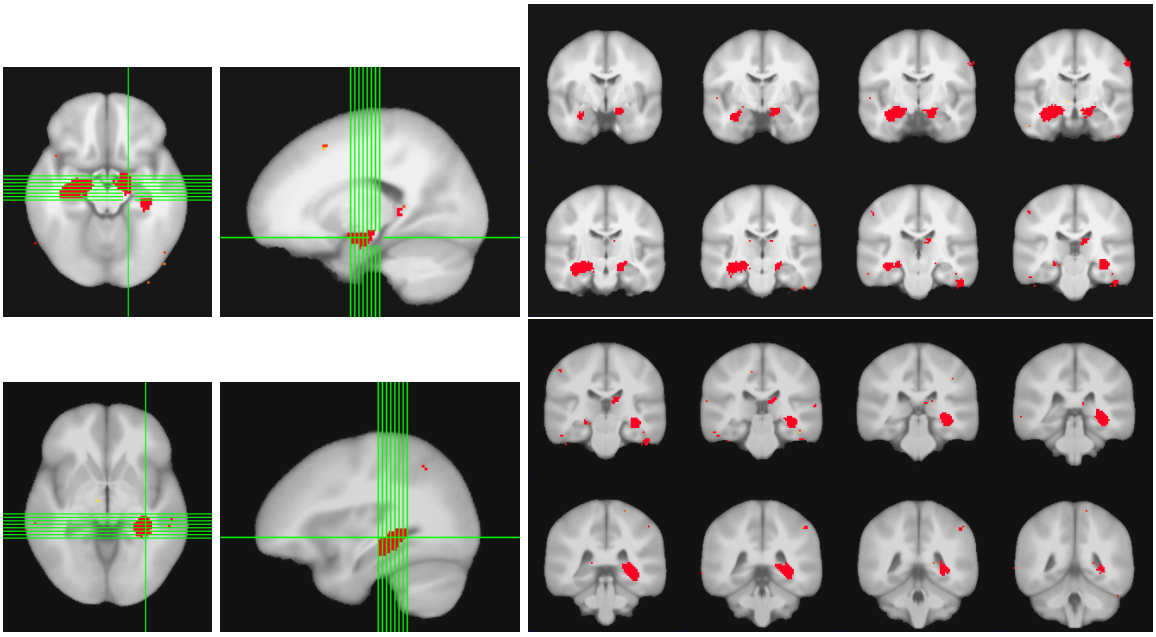
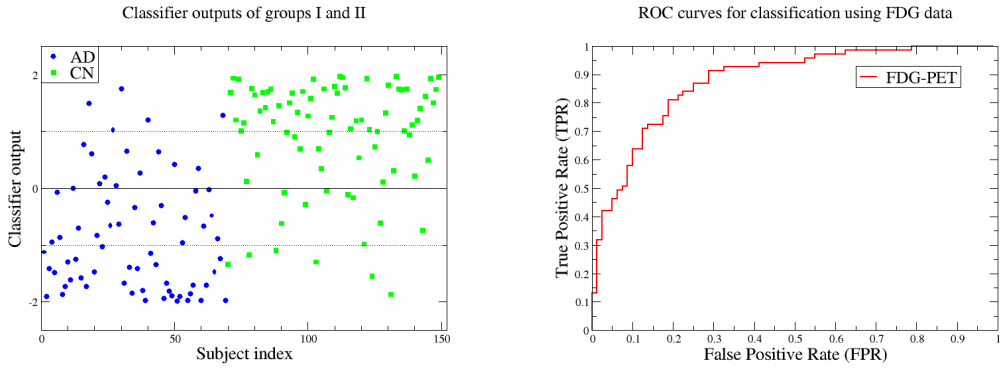


Figure 5: Brain regions selected when using GMPs derived from MR scans as input. Numerical scale corresponds to each voxel's weight in the classifier, and has no applicable units.



(a) (b)

Figure 6: (a) Classifier's output for test images in the FDG-PET population (149 subjects overall). (b) ROC curves on the FDG-PET population.

cingulate cortex and bilateral parietal lobules are well represented, as well as the left inferior temporal lobe. These regions are known to have well established associations with AD-related neurophysiological changes. These results illustrate that the algorithm is able to reliably determine clinically relevant regions in different scanning modalities.

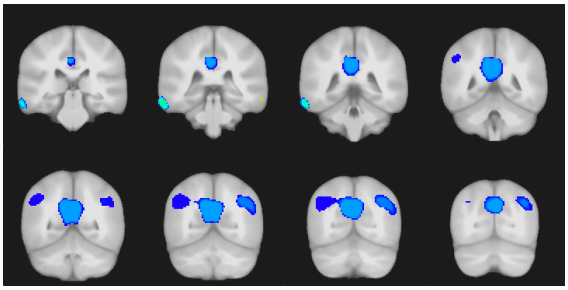


Figure 7: Brain regions selected when using FDG-PET scans as input. Numerical scale corresponds to each voxel's weight in the classifier, and has no applicable units.

4.3 *Relationship with cognitive biomarkers and semi-automatically traced brain region volumes*

Clinical diagnosis of AD depends on various cognitive test results, such as the Mini-Mental State Exam (MMSE). It is reasonable to expect that the output of an effective classification algorithm will agree with these cognitive and clinical measures. We present results showing that our algorithm exhibits these desirable characteristics. The biomarkers available are divided into two broad categories: neuropsychological battery scores and hand-traced brain region volumes. As expected, the classification confidence of the algorithm on the MR population displays a strong statistical correlation with many of these biomarkers, as shown in Fig. 8. Most of the image-based correlation indices are above 0.5 (in absolute value). In Fig. 8(b) we see that the MMSE scores (a measure of global cognitive status) are tightly correlated with the classification confidence of our algorithm.

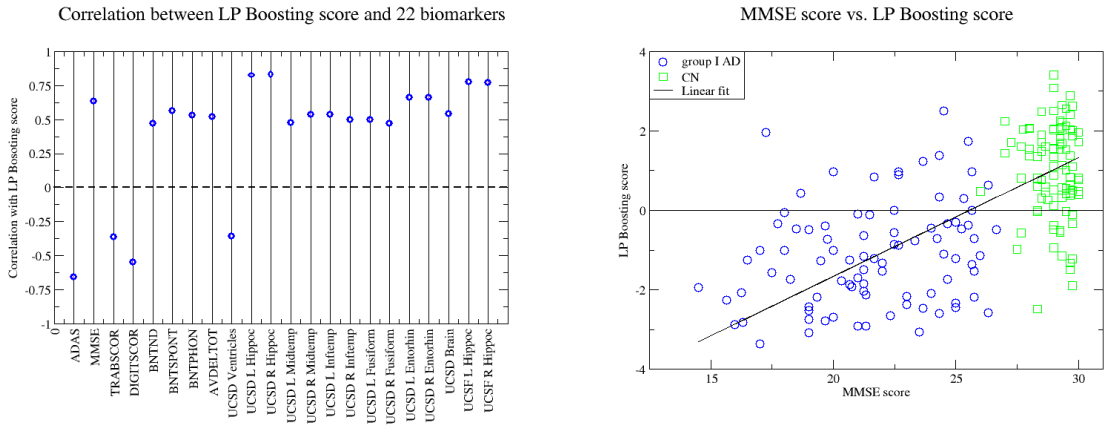


Figure 8: (a) Statistical correlation between each of the biomarker outputs and the algorithm's output in the MR population. Note that both the UCSD and UCSF hippocampal volumes are in close agreement with our method. (b) Classifier's output as a function of MMSE score for each subject in the MR population. A linear best-fit approximation is shown. Note that MMSE scores alone are nearly sufficient to decide the clinical diagnosis for the ADNI cohort, and in fact is a major criterion for the diagnosis of AD.

4.4 *Implementation issues*

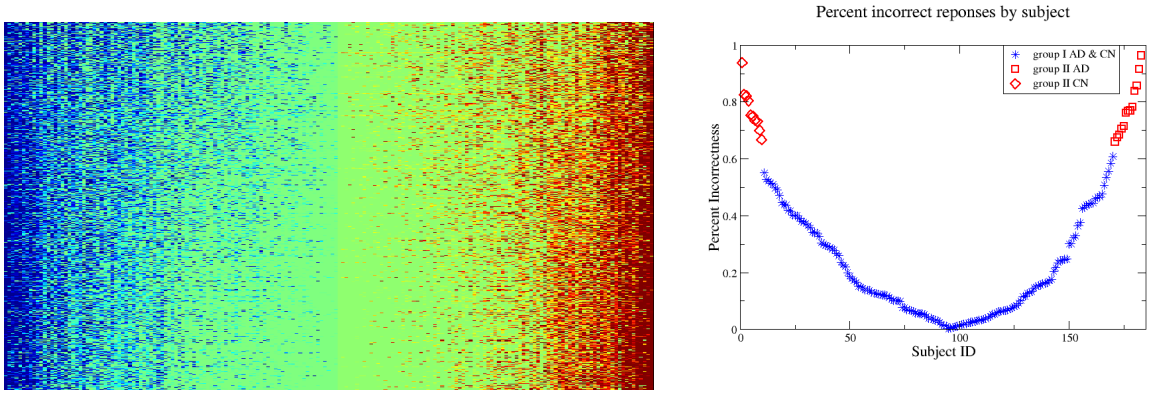
Our proposed algorithm was implemented in Matlab, using CPLEX as the linear program solver for (1). The parameters C and D in (1) were chosen to be 100 and 1000 respectively, and ρ was set to 20, see Fig. 3(a). In practice, we observed that when the parameter D, in (1) is set to 0 (removing the neighbor constraints), then irrespective of the variations introduced in the other parameters, the algorithm always chooses between 25 and 50 voxels (non-zero weights), and gives inferior accuracy. However, when D is set to a reasonable non-zero value depending on the smoothness of the data (e.g. FDG-PET data is far more regular than unsmoothed GMPs) the number of voxels selected varies between 150 to a few thousand. In most cases, a choice of D as described above leads to a 4% increase in classification accuracy. The 1-norm penalty in (1) was scaled (adjusted) using the p-values for each corresponding voxel. Neighboring constraints were not introduced between neighboring weak classifiers where their training set accuracy varied significantly, leaving several 'orphan' weak classifiers (i.e., those which do not participate in any neighboring constraints); such orphan voxels were discarded. For computational reasons, we limited the number of weak classifiers by calculating t-test p-values for each voxel, and discarding all but the most significant ones. We found that using about 2000 weak classifiers worked well in practice. The running time of the algorithm was 15s to 60s for each fold on a modern workstation (2.33 GHz quad-core Xeon). While the implementation is not optimized for speed or memory usage, the computation utilizes no more than 3 GB RAM on our dataset of about 180 volumes of size $91 \times 109 \times 91$. No resampling was needed. The paper has a companion website (`http://erdos.biostat.wisc.edu/~hinrichs/SA_LPBoost2008`) where the code and other supplemental information will be made freely available ¡after publication¿.

5. Analysis of anomalous cases

In addition to the classification experiments described above, we performed a post-hoc analysis on the images, in an effort to control or identify possible outliers. This analysis revealed that *a subset of the images strongly resembled the opposite class*, i.e., some AD subjects resembled controls, while some controls resembled AD subjects. We briefly discuss these results next. For convenience, we refer to this smaller subset of anomalous images as group II, while group I refers to the remaining images not included in group II. That is, group I represents the more homogeneous cases, while group II is comprised of anomalous cases.

Rationale. It is well known that AD-related neurodegenerative pathology is heterogeneous. Thompson et al. (2001) In addition, while the ADNI dataset is based on the most rigorous quality control protocol possible barring access to gold standard diagnostics such as biopsy or post-mortem analysis, there is some expectation that subjects will be misclassified. This may be because of the difficulty in distinguishing AD from other types of dementia such as Frontotemporal Lobar Dementia (FTLD) or Lewy bodies Klöppel et al.

(2008). Further confounding the situation is the possibility of comorbidity of AD with other neurodegenerative and neurovascular diseases such as stroke or multi-infarcts.



(a) (b)

Figure 9: (a) Weak classifier outputs for the 183 members of the MR population, ordered by the number of weak classifiers giving incorrect outputs. Color indicates type and degree of incorrectness; blue corresponds to false negative, red to false positive, and green indicates correct response. Note the sharp boundaries between the red and blue bands at either end – these are the members of group II. (b) Percent of weak classifiers giving incorrect responses for the same subjects.

Identification of possible outlying data. The criterion we used in order to find this group was based on the extent to which the gray matter levels over the whole brain seemed to contradict the label given each subject, i.e., AD or CN. In order to do this, we chose the 2000 most significant voxels in terms of p-values derived from a t-test, and examined the weak classifier predictive outputs on those voxels. These outputs are shown in Figure 9 (a). Each column corresponds to a single example, and each row to a single weak classifier. The columns, i.e. subjects, are ordered from those having the most false negatives at the left, to those having the most false positives at the right. The color indicates the degree of incorrectness, as in Figure 1 (b), with blue indicating false negative, green correct response, and red false positive, respectively. We can clearly see that there are two "bars" at either end, consisting of subjects which are given the wrong label by nearly the entire set of weak classifiers. Subjects for which more than 65% of the weak classifiers gave incorrect outputs were placed in group II (Note that this closely matched the "bars" in Figure 9 (a)). This gave 10 controls, and 13 AD subjects. Figure 9 (b) shows the percentage of weak classifiers giving incorrect outputs on each subject. Our labeling of anomalous subjects in this manner is not simply an artifact of our weak classifiers, but reveals a systemic pattern of deviation from the mean in each group. Evidence from hippocampus volume measures yields a similar labeling. That is, the set of subjects more than one standard deviation away from the group mean, (of hippocampal volume), is almost identical to the set of examples placed in group II as above.

Effect on classification. As may be expected, when we re-ran our experiments on group I alone, the accuracy of our algorithm improved dramatically - accuracy increased to 92.5%, with a sensitivity and specificity of 88% and 96%, respectively, and an area under the ROC curve of 0.9815. This contrasts significantly with the results reported in Table 2. Perhaps the best way to interpret these results is that the accuracy reported for the entire population is a *lower bound* on our method's accuracy, and the results we report for group I alone represent a likely *upper bound* on our algorithm's accuracy, while the true accuracy when a gold standard is available will probably lie somewhere in between.

Characteristics of group II controls. We found that in several respects the group II controls were very similar to group I AD subjects.

- Our first observation was that the group II controls had *significantly less* total brain volume, even relative to group I AD subjects: 8.8×10^5 (group II CN)[3] compared to 1.02×10^6 (group I CN) and 9.48×10^5 (group I AD) with p-values $< 10^{-9}$.

- *All* regions (where manual tracings are provided in the ADNI dataset) were significantly smaller in group II controls compared to group I controls (p-values $< 10^{-3}$). Regional volumes for group II controls were closer to the respective measures from group I AD subjects.

- The ventricles in group II controls were *not* significantly smaller than controls in group I, which indicates that the above variations cannot be attributed to smaller brain sizes alone (and suggests possible atrophy).

- The hippocampal volume measures showed even larger variations in controls between groups I and II.

- Our VBM analysis between group II controls and group I AD subjects gave *no* discriminating regions and only isolated voxels.

- VBM analysis also revealed a significant gray matter density deterioration (p-values $< 10^{-6}$) in the hippocampus and parahippocampal gyri for group II controls, when compared to controls in group I.

Characteristics of group II AD subjects.

AD subjects in group II similarly resembled group I controls.

- The mean total brain volume of group II AD subjects was almost identical to that of group I controls ($\approx 1.02 \times 10^6$ in both groups). By comparison, the mean total brain volume of group I AD subjects was 9.48×10^5.

[3]Units are mm^3.

- In the hippocampus and entorhinal cortex the mean volume among group II AD subjects was nearly the same as that of group I controls: 7159.93 (UCSD) in group II AD subjects versus 7390.93 (UCSD) in group I controls for the hippocampus. By comparison, the same measures were 5520.07 (UCSD) in group I AD subjects. The mean entorhinal cortex volumes had a similar proportion.

- Our VBM analysis showed greater gray matter densities in the hippocampus for group II AD subjects compared to group I AD and hypertrophy in the thalamus relative to group I controls.

Cognitive status. While the image based biomarkers showed significant variations between groups I and II, the associated cognitive status and neuropsychological scores (e.g., MMSE) were relatively consistent. This is not surprising because cognitive status, especially the MMSE score, is highly relevant to clinical diagnosis. However, Group II AD subjects did show significant group differences in tests measuring logical memory – both immediate and delayed recall, number of spontaneous correct responses given on the Boston Naming Test, and audio visual tests. In all of these, group II AD subjects scored higher indicating slightly healthier cognitive status (consistent with lower observed atrophy in the preceding discussion). Of these, the delayed recall was the most significantly different (p-value ≈ 0). There was no significant difference between the performance of group I and group II controls on any measure of cognitive status. Summaries of biomarkers significantly differing between both groups I and II are presented in Tables 3 and 4 in the Appendix.

Summary. We note that confirmed diagnosis of AD is only possible post-mortem. Given the clinical nature of the ADNI data set, it is possible that some AD subjects in the cohort may have another form of dementia or possibly depression, while some controls may have AD in the early stages, and have not yet begun showing signs of cognitive decline. The classification algorithm, however, assumes that every label in the training data is correct, and therefore tries to correctly classify every training example. In the presence of incorrectly labeled examples, however, it is difficult for a method to have a lower expected error rate than the fraction of mislabeled examples in the training set. Clearly, if our data set contains mislabeled examples Wade et al. (1987), Schofield et al. (1995), Burns et al. (1990), an automated method may not be able to to outperform this limitation. An interesting question then is, can we detect subjects with signs of abnormality? Characterizing this set will be useful for not only improving the accuracy of classification systems evaluated on this dataset, but may also suggest ways that the classifier can be modified to automatically handle them. Our analysis above, and evaluations of classifier's performance with/without group II may be a useful first step in potentially discovering mislabeled subjects that may not have been identified by the study's strict quality control protocols.

6. Conclusions and future directions

We have demonstrated a new algorithm for automated AD classification of the level of single subjects using either structural or functional image scans. Our technique directly incorporates spatial relationships between voxels into the learning framework, and requires *no* extra modality-dependent pre- or post-processing. We have shown extensive evaluations on the ADNI dataset. Since results from several other existing techniques were reported on different datasets with different sample sizes, we believe that our results and software will enable objective comparisons of different methods to evaluate their advantages and disadvantages in context of this large and well characterized image data. Such comprehensive evaluations will likely lead to standardization and development of improved classification systems for AD diagnosis.

Acknowledgments

This research was supported in part by the Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, UW ICTR through an NIH Clinical and Translational Science Award (CTSA) 1UL1RR025011, a Merit Review Grant from the Department of Veterans Affairs, the Wisconsin Comprehensive Memory Program, and NIH grant AG011915. The authors also acknowledge the facilities and resources at the William S. Middleton Memorial Veterans Hospital.

Data collection and sharing for this project was funded by the Alzheimers Disease Neuroimaging Initiative (ADNI; Principal Investigator: Michael Weiner; NIH grant U01 AG024904). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering (NIBIB), and through generous contributions from the following: Pfizer Inc., Wyeth Research, Bristol-Myers Squibb, Eli Lilly and Company, GlaxoSmithKline, Merck & Co. Inc., AstraZeneca AB, Novartis Pharmaceuticals Corporation, Alzheimers Association, Eisai Global Clinical Development, Elan Corporation plc, Forest Laboratories, and the Institute for the Study of Aging, with participation from the U.S. Food and Drug Administration. Industry partnerships are coordinated through the Foundation for the National Institutes of Health. The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimers Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory of Neuro Imaging at the University of California, Los Angeles.

References

Arimura, H., Yoshiura, T., Kumazawa, S., Tanaka, K., Koga, H., Mihara, F., Honda, H., Sakai, S., Toyofuku, F. and Higashida, Y. (2008). Automated method for identification of patients with Alzheimer's disease based on three-dimensional MR images. *Academic Radiology* **15**, 274–284.

Ashburner, J. (2007). A fast diffeomorphic image registration algorithm. *Neuroimage* **38**, 95.

Ashburner, J. and Friston, K. J. (2000). Voxel-Based Morphometry - the methods. *Neuroimage* **11**, 805–821.

Bishop, C. (2006). *Pattern Recognition and Machine Learning*. Springer New York.

Bradley, P. S. and Mangasarian, O. L. (1998). Feature selection via concave minimization and support vector machines. *Proc. 15th International Conf. on Machine Learning* pages 82–90.

Burns, A., Luthert, P., Levy, R., Jacoby, R. and Lantos, P. (1990). Accuracy of clinical diagnosis of Alzheimer's disease. *British Medical Journal* **301**, 47.

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

Davatzikos, C., Resnick, S., Wu, X., Parmpi, P. and Clark, C. (2008). Individual patient diagnosis of AD and FTD via high-dimensional pattern classification of MRI. *Neuroimage* **41**, 1220–1227.

Demirci, O., Clark, V. P. and Calhoun, V. D. (2008). A projection pursuit algorithm to classify individuals using fMRI data: Application to schizophrenia. *Neuroimage* **39**, 1774–1782.

Demiriz, A., Bennett, K. P. and Shawe-Taylor, J. (2002). Linear programming boosting via column generation. *Machine Learning* **46**, 225–254.

deToledo-Morrell, L., Stoub, T. R., Bulgakova, M., Wilson, R., Bennett, D., Leurgans, S., Wuu, J. and Turner, D. (2004). MRI-derived entorhinal volume is a good predictor of conversion from MCI to AD. *Neurobiology of Aging* **25**, 1197–1203.

Duchesne, S., Caroli, A., Geroldi, C., Barillot, C., Frisoni, G. B. and Collins, D. L. (2008). MRI-based automated computer classification of probable AD versus normal controls. *IEEE Transactions on Medical Imaging* **27**, 509–520.

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

Fan, Y., Resnick, S. M., Wu, X. and Davatzikos, C. (2008). Structural and functional biomarkers of prodromal Alzheimer's disease: a high-dimensional pattern classification study. *Neuroimage* **41**, 277–285.

Fox, M. C. and Schott, J. M. (2004). Imaging cerebral atrophy: normal ageing to Alzheimer's disease. *The Lancet* **363**, 392–394.

Freund, Y. (1995). Boosting a weak learning algorithm by majority. *Inf. Comput.* **121**, 256–285.

Friston, K. J., Holmes, A., Price, C. J. and Frith, C. D. (1996). Detecting activations in PET and fMRI: Levels of inference and power. *Neuroimage* **4**, 223–235.

Fung, G. M. and Mangasarian, O. L. (2004). A feature selection newton method for support vector machine classiacation. *Computational Optimization and Applications* **28**, 185–202.

Gaul, W. and Ritter, G., editors (2000). *Classification, Automation, and New Media*, chapter Sparse Kernel Feature Analysis. Springer.

Grove, A. and Schuurmans, D. (1998). Boosting in the limit: Maximizing the margin of learning ensembles. *Proceesings of the Fifteenth National Conference on Artificial Intelligence* .

Hoffman, J. M., Welsh-Bohmer, K. A., Hanson, M., Crain, B., Hulette, C., Earl, N. and Coleman, R. (2000). FDG PET imaging in patients with pathologically verified dementia. *Journal of Nuclear Medicine* **41**, 1920–1928.

Jack Jr., C. R., Shiung, M. M., Weigand, S. D., O'Brien, P., Gunter, J., Boeve, B., Knopman, D., Smith, G., Ivnik, R., Tangalos, E. et al. (2005). Brain atrophy rates predict subsequent clinical conversion in normal elderly and amnestic MCI. *Neurology* **65**, 1227–1231.

Johnson, S. C., Schmitz, T. W., Trivedi, M. A., Ries, M. L., Torgerson, B. M., Carlsson, C. M., Asthana, S., Hermann, B. P. and Sager, M. A. (2006). The influence of Alzheimer disease family history and apolipoprotein E varepsilon4 on mesial temporal lobe activation. *The Journal of Neuroscience* **26**, 6069–6076.

Jolliffe, I. T. (2002). *Principal Component Analysis*. Springer New York, second edition.

Klöppel, S., Stonnington, C., Chu, C., Draganski, B., Scahill, R., Rohrer, J., Fox, N., Jack, C., Ashburner, J. and Frackowiak, R. (2008). Automatic classification of MR scans in Alzheimer's disease. *Brain* **131**, 681–689.

Matsuda, H. (2001). Cerebral blood flow and metabolic abnormalities in Alzheimer's disease. *Annals of Nuclear Medicine* **15**, 85–92.

Minoshima, S., Foster, N. L. and Kuhl, D. E. (1994). Posterior cingulate cortex in Alzheimer's disease. *Lancet* **344**, 895.

Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill.

Reiman, E. M., Caselli, R. J., Yun, L. S., Chen, K., Bandy, D., Minoshima, S., Thibodeau, S. and Osborne, D. (1996). Preclinical evidence of Alzheimer's Disease in persons homozygous for the $\varepsilon 4$ allele for apolipoprotein e. *New England Journal of Medicine* **334**, 752–758.

Sager, M. A., Hermann, B. and Rue, A. L. (2005). Middle-aged children of persons with Alzheimer's Disease: APOE genotypes and cognitive function in the Wisconsin Registry for Alzheimer's Prevention. *Journal of Geriatric Psychiatry Neurology* **18**, 245–249.

Sahiner, B., Chan, H., Petrick, N., Wagner, R. and Hadjiiski, L. (2000). Feature selection and classifier performance in computer-aided diagnosis: The effect of finite sample size. *Medical Physics* .

Schapire, R. (1990). The strength of weak learnability. *Machine Learning* **5**, 197–227.

Schofield, P., Tang, M., Marder, K., Bell, K., Dooneief, G., Lantigua, R., Wilder, D., Gurland, B. and andR. Mayeux, Y. S. (1995). Consistency of clinical diagnosis in a community-based longitudinal study of dementia and Alzheimer's disease. *Neurology* **45**, 2159–2164.

Shen, L., Ford, J., Makedon, F. and Saykin, A. (2003). Hippocampal shape analysis: surface-based representation and classification. In *Proceedings of SPIE*, volume 5032, pages 253–264.

Shock, N., Greulich, R. and et al., R. A. (1984). Normal human aging: the baltimore longitudinal study of aging. Washington, DC: US Government Printing Office.

Small, G., Ercoli, L. M., Silverman, D. H., Huang, S., Komo, S., Bookheimer, S., Lavretsky, H., Miller, K., Siddarth, P., Rasgon, N. et al. (2000). Cerebral metabolic and cognitive decline in persons at genetic risk for Alzheimer's disease. *Proceedings of the National Aceademies of Science USA* **97**, 6037–6042.

Soriano-Mas, C., Pujol, J., Alonso, P., Cardoner, N., Menchn, J. M., Harrison, B. J., Deus, J., Vallejo, J. and Gaser,

C. (2007). Identifying patients with obsessive-compulsive disorder using whole-brain anatomy. *Neuroimage* **35**.

Thompson, P. M. and Apostolova, L. (2007). Computational anatomical methods as applied to ageing and dementia. *British Journal of Radiology* **80**, 78–91.

Thompson, P. M., Mega, M. S., Woods, R. P., Zoumalan, C. I., Lindshield, C. J., Blanton, R. E., Moussai1, J., Holmes, C. J., Cummings, J. L. and Toga, A. W. (2001). Cortical change in Alzheimer's disease detected with a disease-specific population-based brain atlas. *Cerebral Cortex* **11**, 1–16.

Vemuri, P., Gunter, J., Senjem, M. L., Whitwell, J. L., Kantarci, K., Knopman, D. S., Boeve, B. F., Petersen, R. C. and Jr., C. R. J. (2008). Alzheimer's disease diagnosis in individual subjects using structural MR images: validation studies. *Neuroimage* **39**, 1186–1197.

Wade, J. P., Mirsen, T. R., Hachinski, V. C., Fisman, M., Lau, C. and Merskey, H. (1987). The clinical diagnosis of Alzheimer's Disease. *Neurology* **44**.

Appendix

Supplementary material related to the discussion in Section 5: comparison of various biomarkers between group I and group II. AD subjects are treated separately from CN subjects. Only biomarkers showing significant variation are included here.

Biomarker (AD subjects)	Group I	Group II	Z-test p-value
Mini-Mental State Exam (MMSE)	21.5 (3.04)	22.94 (2.84)	0.08
Tau-protein	111.94 (51.77)	151.88 (88.34)	0.0147
Logical Memory - Immediate Recall	3.13 (2.18)	4.91 (3.338)	$\sim 10^{-3}$
Logical Memory - Delayed Recall	0.48 (0.8)	3.13 (2.54)	$\sim 10^{-16}$
Boston Naming - Spontaneous Correct Responses	19.69 (6.95)	25.49 (4.70)	$\sim 10^{-3}$
Audio Visual	1.1 (1.08)	1.99 (2.15)	0.0374
Brain volume (UCSD)	948005.03 (84947.07)	1025001.3 (79868.99)	$\sim 10^{-3}$
L. Hippocampal volume (UCSD)	2706.69 (382.98)	3446.61 (573.23)	$\sim 10^{-10}$
R. Hippocampal volume (UCSD)	2813.38 (432.2)	3713.32 (368.21)	$\sim 10^{-12}$
L. Entorhinal cortex volume (UCSD)	2.44 (0.46)	3.03 (0.36)	$\sim 10^{-5}$
R. Entorhinal cortex volume (UCSD)	2.50 (0.46)	3.18 (0.42)	$\sim 10^{-7}$
L. Hippocampal volume (UCSF)	1518.45 (246.11)	1996.95 (426.44)	$\sim 10^{-10}$
R. Hippocampal volume (UCSF)	1498.39 (334.53)	2163.35 (341.04)	$\sim 10^{-14}$

Table 3: Comparison of relevant biomarkers in group I AD and group II AD. MMSE is included for reference; all other biomarkers listed are significantly different between groups at at least the 0.05 level.

Biomarker (CN subjects)	Group I	Group II	Z-test p-value
Mini-Mental State Exam (MMSE)	28.98 (0.8)	29.19 (0.69)	0.33
Ventricles volume (UCSD)	38788.18 (23264.37)	40085.85 (13514.94)	0.84
Brain volume (UCSD)	1023746.53 (86217.87)	880452.33 (75572.03)	$\sim 10^{-9}$
L. Hippocampal volume (UCSD)	3599.87 (383.32)	3116.90 (301.58)	$\sim 10^{-5}$
R. Hippocampal volume (UCSD)	3791.06 (422.58)	3159.28 (359.84)	$\sim 10^{-7}$
L. Mid temporal volume (UCSD)	2.58 (0.17)	2.45 (0.12)	$\sim 10^{-3}$
R. Mid temporal volume (UCSD)	2.6 (0.20)	2.48 (0.21)	0.0454
L. Inf. temporal volume (UCSD)	2.64 (0.15)	2.49 (0.14)	$\sim 10^{-4}$
R. Inf. temporal volume (UCSD)	2.60 (0.19)	2.47 (0.25)	$\sim 10^{-2}$
L. Fusiform volume (UCSD)	2.39 (0.17)	2.25 (0.16)	$\sim 10^{-3}$
R. Fusiform volume (UCSD)	2.36 (0.17)	2.25 (0.18)	$\sim 10^{-2}$
L. Entorhinal cortex volume (UCSD)	3.19 (0.30)	2.86 (0.36)	$\sim 10^{-4}$
R. Entorhinal cortex volume (UCSD)	3.34 (0.32)	3.02 (0.51)	$\sim 10^{-4}$
L. Hippocampal volume (UCSF)	2126.69 (267.67)	1795.54 (208.3)	$\sim 10^{-5}$
R. Hippocampal volume (UCSF)	2176.57 (275.65)	1781.65 (252.45)	$\sim 10^{-7}$

Table 4: Comparison of relevant biomarkers in group I CN and group II CN. MMSE is included for reference; all other biomarkers listed are significantly different between groups at at least the 0.05 level.