# Comparison of AdaBoost and Support Vector Machines for Detecting Alzheimer's Disease through Automated Hippocampal Segmentation

Jonathan H. Morra, Zhuowen Tu, Liana G. Apostolova, Amity E. Green, Arthur W. Toga, and Paul M. Thompson

*Abstract*— **We compared four automated methods for hippocampal segmentation using different machine learning algorithms (1) hierarchical AdaBoost, (2) Support Vector Machines (SVM) with manual feature selection, (3) hierarchical SVM with automated feature selection (Ada-SVM), and (4) a publicly available brain segmentation package (FreeSurfer). We trained our approaches using T1-weighted brain MRI's from 30 subjects (10 normal elderly, 10 mild cognitive impairment (MCI), and 10 Alzheimer's disease (AD)), and tested on an independent set of 40 subjects (20 normal, 20 AD). Manually segmented gold standard hippocampal tracings were available for all subjects (training and testing). We assessed each approach's accuracy relative to manual segmentations, and its power to map AD effects. We then converted the segmentations into parametric surfaces to map disease effects on anatomy. After surface reconstruction, we computed significance maps, and overall corrected *p*-values, for the 3D profile of shape differences between AD and normal subjects. Our AdaBoost and Ada-SVM segmentations compared favorably with the manual segmentations and detected disease effects as well as FreeSurfer on the data tested. Cumulative *p*-value plots, in conjunction with the False Discovery Rate method, were used to examine the power of each method to detect correlations with diagnosis and cognitive scores. We also evaluated how segmentation accuracy depended on the size of the training set, providing practical information for future users of this technique.**

*Index Terms*— **AdaBoost, Alzheimer's disease, hippocampal segmentation, support vector machines, surface modeling**

## I. Introduction

Hippocampal segmentation is a key step in many medical imaging studies for statistical comparison of anatomy across populations, and for tracking group differences or changes over time. Specifically in Alzheimer's disease, hippocampal volume and shape measures are commonly used to examine the 3D profile of early degeneration, and detect factors that predict imminent conversion to dementia [2]. Early detection of AD has grown in importance over the last decade because of the acknowledged benefits of treating patients before severe degeneration has occurred [12]. In epilepsy, hippocampal shape measures computed from a pre-operative scan, can also predict whether patients will be seizure-free following surgical treatment [32]. A broad range of ongoing neuroscientific studies have used hippocampal surface models to examine the trajectory of childhood development [21], childhood-onset schizophrenia [43], autism [42], Alzheimer's disease and mild cognitive impairment [6], [18], [68], drug-related degeneration in methamphetamine users [60], and hypertrophic effects of lithium treatment in bipolar illness [5], [4]. Hippocampal models are also used in genetic studies that seek anatomical shape signatures associated with increased liability for illness, providing measures to assist in the search for genes influencing hippocampal morphology [40]. There has also been work developing algorithms for 3D nonlinear registration or computational matching of hippocampal surfaces, based on elastic flows in the surface parameter space [72], [69], direct surface matching using exterior calculus approaches [66], spherical harmonic approaches [23], or level-set approaches and intrinsic shape context measures to constrain 3D harmonic mappings [54].

One of the first steps for all these methods is segmenting out the hippocampus from a 3D brain MRI scan. Despite much active work on the computational anatomy of the hippocampus, segmentation is still commonly performed manually by human experts. Manual tracing is difficult and time consuming, so automating this process is highly desirable. As a result, several partially or fully automated approaches have been proposed to segment the hippocampus, but none is currently in wide use.

Semi-automatic methods still require some user input and therefore some amount of expert knowledge. Hogan et al. [25] used a deformable template approach to elastically deform a hippocampal model to match its counterpart in a target scan. This method was successful, but required 10-15 minutes of user interaction to define both global and hippocampal specific landmarks. Another approach by Yushkevich et al. (ITK-SNAP) [71] used active surface methods implemented in a level-set framework. In ITK-SNAP, the user must first determine an approximate boundary for the structure of interest, and the final segmentation depends to some extent on the starting position of the active surface. Also, the deforming surface is driven by an intensity-based energy minimization functional. This makes it very difficult to segment a structure like the hippocampus as local intensity information is not sufficient to determine the hippocampal boundary, particularly its junction with the amygdala. Shen et al. [52] also used an active contour method augmented by *a priori* shape information. Nevertheless, they are still subject to some of the same limitations as ITK-SNAP, requiring some user initialization.

Fully automatic methods do not require any user input, and are usually based on extracting and combining some set of image features to determine the structure boundary. Some commonly used features include image intensity, gradients, curvatures, tissue classifications, local filters, or spectral decompositions (e.g., wavelet analysis). However, determining which features are informative for segmentation, and how to

combine those features is difficult without expert knowledge of the problem domain, and without proper features for each different problem, segmentation becomes very difficult. Lao et al. [30], used a multispectral approach to segment white matter lesions based on co-registered MRI scans with different T1- and T2-dependent contrasts. They used SVMs to combine the intensity profile of these different scans, and performed multivariate classification in the joint signal space. This will only work if segmentation is possible with only these specific MRI signals, which in general it is not. Powell [46] also used SVMs and artificial neural networks to segment out the hippocampus. Although they report very good segmentation performance for their data, their test size is small (5 brains) and they use 25 manually selected features, which means that generalization to other datasets is not guaranteed. Golland et al. [22] proposed using a large feature pool, and Principal Component Analysis (PCA) to reduce the size of the feature pool, followed by SVM for classification. PCA does not choose features that are necessarily well-suited for segmentation, it only chooses features with a large variance. Therefore, the features chosen by PCA are not guaranteed to give good classification results. Another common approach for fully automated segmentation is to nonlinearly transform an atlas, where the hippocampus is already segmented, onto a new brain scan, using deformable registration. Such an approach was proposed by Hammers et al. [24], but its accuracy depends on the image data used to construct the atlas, as well as the registration model (e.g., octree- or spline-based, elastic, or fluid) and may have difficulty in labeling new scans with image intensities or anatomical shapes that differ substantially from the atlas. A fully automatic extension of the level-set approach was suggested by Pohl et al. [44]. In this approach the traditional signed distance function applied in most level-set implementations is transformed into a probability using the LogOdds space. This can lead to a more natural formulation of the multi-class segmentation problem by incorporating statistical information into the level-set approach.

Another fully automated approach for subcortical segmentation is FreeSurfer by Fischl et al. [16]. FreeSurfer uses a Markov Random Field to approximate the posterior distribution for anatomic labelings at each voxel in the brain. However, in addition to this, they use a very strong prior based on the knowledge of where structures are in relation to each other. For instance, the amygdala is difficult to distinguish from the hippocampus based on intensity alone. However, they always have the same spatial relationship, with the amygdala immediately anterior to the hippocampus, and this is encoded by the statistical prior in FreeSurfer to separate them correctly. FreeSurfer also makes use of additional statistical priors on the likely location of structures after scans are aligned into a standard stereotaxic space, and their expected intensities based on spatially-adaptive fitting of Gaussian mixture models to classify tissues in a training dataset. As FreeSurfer is a freely available package over the internet, we compared its segmentation results to ours throughout this paper. This required us to develop some extensions of the freely available capabilities of FreeSurfer, such as converting its usual outputs – multi-class segmented volumes – into parametric surfaces, allowing us to compare surface-based statistical maps of disease effects, based on the outputs of all segmentation methods.

Recent developments in machine learning, such as AdaBoost [17], have automated the feature selection process for several imaging applications. Support Vector Machines (SVM) [67] can effectively combine features for classification. AdaBoost and SVM may be used to classify vector-valued examples, and both have been separately applied to medical image analysis before, but this paper evaluates the benefits of combining them sequentially.

Statistical classification is an active area of pattern recognition and computer vision research in which scalar- or vector-valued observations are automatically assigned to specific groups, often based on a training set of previously labeled examples. In medical imaging, different types of classification tasks are performed, e.g., classifying image voxels as belonging to a certain anatomical structure, or classifying an individual scanned into one of several diagnostic groups (disease versus normal, semantic dementia versus Alzheimer's disease, for example). For clarification, we note that this paper classifies voxels in a brain MRI scan as belonging to the hippocampus versus not, but in a second step we use these classified structures to create statistical maps of systematic differences in anatomy between Alzheimer's patients and controls. As such, although the main goal of the paper is to achieve segmentations of the hippocampus, we illustrate the use of the these segmentations in an application where differences between disease and normality are detected and mapped.

Among several algorithms proposed for statistical classification, AdaBoost is a meta-algorithm that sequentially selects weak classifiers (i.e., ones that do not perform perfectly when used on their own) from a candidate pool and weights each of them based on their error. A weak learner is any statistical classifier that performs better than pure chance. Each iteration of AdaBoost assigns an "importance weight" to each example; examples with a higher weight, classified incorrectly on previous iterations, will receive more attention on subsequent iterations, tuning the weak learners to the difficult examples. Testing examples with AdaBoost is therefore simply a weighted vote of the weak-learners.

SVMs, on the other hand, seek a hypersurface in the space of all features that both minimizes the error of training examples and maximizes the margin, defined as the distance between the hypersurface and the closest value in feature space, in the training data. SVMs can use any type of hypersurface by making use of the "kernel trick". [10].

SVMs have been used widely in medical imaging for brain tumor recognition and malignancy prediction [35], white matter lesion segmentation [47], for discriminating schizophrenia patients from controls based on morphological characteristics [71] and for analyzing functional MRI time-series [28].

Although SVMs have been widely used in medical imaging, AdaBoost has not. However, as AdaBoost can select informative features from a potentially very large feature pool, it is likely to offer advantages in automatically finding good features for classification. This can greatly reduce, or eliminate the need for experts to choose informative features based on

knowledge of every classification problem. Instead, one just needs to define a list of possibly informative features, and AdaBoost will choose those that are actually informative.

For our classification problem, we compared four different classification techniques, (1) FreeSurfer [16], (2) SVM with manually selected features (manual SVM), (3) AdaBoost, and (4) SVM with features automatically selected by AdaBoost (Ada-SVM). As AdaBoost can select features automatically, we improved the classification ability of AdaBoost and Ada-SVM by implementing them in a hierarchical decision tree framework.

As a testbed to examine segmentation performance, we trained and tested our methods on a dataset of 70 3D volumetric T1-weighted brain MRI scans. 30 of these subjects were reserved for training, and 40 for testing. The training subjects were composed of 10 subjects with Alzheimer's disease (AD), 10 with mild cognitive impairment (MCI), a state which carries an increased risk for conversion to AD, and 10 age-matched controls. The 40 testing subjects were composed of 20 AD and 20 controls. Due to the small number of MCI subjects available for this study, we choose to add them to the training group because it increased the variability on which to train. All subjects were scanned on a 1.5 Tesla Siemens scanner, with a standard high-resolution spoiled gradient echo (SPGR) pulse sequence with a TR (repetition time) of 28 ms, TE (echo time) of 6 ms, field of view of 220mm, 256x192 matrix, and slice thickness of 1.5mm. For application to drug trials, and neuroscientific studies of disease, we would require our algorithm to perform accurate segmentation for normal subjects and those affected by degenerative disease, which affects hippocampal shape and image contrast; therefore, we trained our classifier on manually segmented scans from both normal and diseased subjects.

Recently the authors have also proposed another segmentation method based on AdaBoost [38]. In this implementation, we use AdaBoost inside of a new classification scheme which incorporates context information. We call this the auto context model, as a spatial prior on the labels is successively refreshed and features based on the updated spatial prior (e.g. gradients, and filter outputs) are also included as extra features for AdaBoost to consider. In this paper, we wish to show how AdaBoost is less effective than a combination of AdaBoost and SVM. Although the problems are the same in both papers, here we are focusing on the learner itself (AdaBoost versus Ada-SVM) and in our other work we are focusing on incorporating contextual information into the classification problem. In fact we could use our new Ada-SVM inside of the auto context model, but we forgo that here to concentrate on the added benefits of Ada-SVM as compared to AdaBoost.

## II. Problem

A typical goal of image segmentation problems is to assign each image voxel to one of several classes e.g. background, hippocampus, amygdala, ventricles, etc. For hippocampal segmentation, we focus here on the case where there are only two classes, hippocampus and background. Therefore, our problem is reduced to taking in an input volume $\mathbf{V}$ and outputting a binary classification $\mathbf{V_b}$ where each voxel in $\mathbf{V_b}$ has either $+1$ or $-1$ denoting whether we estimate it to be inside the hippocampus $(+1)$, or outside $(-1)$. If we let each voxel in $\mathbf{V}$ be an example $x$ and the corresponding output in $\mathbf{V_b}$ be $y$, the solution to this problem may be formulated in a Bayesian framework as shown in eqn. 1.

$$\begin{aligned}\mathbf{V_b}^* = argmax_{\mathbf{V_b}} P(\mathbf{V_b}|\mathbf{V}) = \\ argmax_{\mathbf{V_b}} P(\mathbf{V}|\mathbf{V_b})P(\mathbf{V_b})\end{aligned} \quad (1)$$

However, this approach is not reasonable in practice because it requires full knowledge of all possible features. Instead, we approximate the posterior distribution $P(\mathbf{V}|\mathbf{V_b})$ with both AdaBoost and SVM techniques, and implicitly integrate $P(\mathbf{V_b})$ as a shape parameter.

For the remainder of this paper, each example is considered as a vector of features (e.g., gradient strength, mean filter response) derived from a single voxel, and for the first voxel this example can be written as $\vec{x}_1$. If $\vec{x}_i$ is the feature vector for the $i$-th voxel, then the set of all examples can be represented as an ordered set of examples, or as the vector $(\vec{x}_1, \cdots, \vec{x}_N)$, where each $\vec{x}_i$ is the same length as the number of features. Individual voxels are treated as independent examples, and all the voxels from the same subject are treated in the same way as voxels from other subjects in the training set. In other words, $(\vec{x}_1, \cdots, \vec{x}_N)$ is a long vector where the number of examples, $N$, equals the number of labeled voxels in the training set. For each voxel with feature vector $\vec{x}_i$ in the training set, a label $y_i$ is assigned, so the set of labels is $(y_1, \cdots, y_N)$, with a label corresponding to each voxel (or to the feature vector derived from it). Since we are only dealing with the two-class classification problem, $y_i$ can only take values of $-1$ or $+1$. For the image segmentation task, an example is a specific voxel from a given image. Each voxel is treated as a separate example. A feature is any property of the image, such as intensity, $x$, $y$, $z$ position, or an image filter, such as a mean filter, Haar filter, $x$, $y$, $z$ gradient filter, etc. The specific feature set we use for our experiments is described in the Experiments section.

## III. Methods

In this section, we first formally define AdaBoost and SVMs, and then show how they approximate the ideal Bayesian classifier. Next we give reasons for using one method versus the other, or both together. Then, we outline how we express AdaBoost and SVMs in a hierarchical format. Finally, we define our methodology for mapping the effects of AD on the hippocampus.

### A. Support Vector Machines

SVMs are very popular for discrimination tasks because they can accurately combine many features to find an optimal separating hyperplane. SVMs minimize the classification error based on two constraints simultaneously. They both seek a hyperplane with a large margin – i.e. the distance from the closest example to the separating hyperplane – and minimize the number of wrongly classified training examples, using

slack variables. If an example is perfectly classifiable in feature space then the second constraint is not necessary. However, this is not the case in our problem, so SVMs both minimize the error on the training set and maximize the margin, increasing their generalization ability. Eqn. 2 summarizes the SVM formulation [67].

$$\begin{aligned} \min \quad & \frac{1}{2}||\vec{\alpha}||_2 + C\sum_i z_i \\ \text{subject to} \quad & y_i(\vec{\alpha}\cdot\vec{x} - b) \geq 1 - z_i \end{aligned} \tag{2}$$

Here, $\vec{\alpha}$ is the vector corresponding to the separating hyperplane, $\frac{1}{||\vec{\alpha}||_2}$ is the margin of the hyperplane, according to the $l_2 - norm$, $\vec{x}$ is a vector consisting of the features, $b$ is a scalar bias term (so the hyperplane is not forced to go through the zero point), $z_i$ are slack variables (those classified on the wrong side of the margin of the separating hyperplane), and C is a user-defined parameter controlling the tradeoff between margin and the number of slack variables.

To minimize eqn. 2, one can formulate the problem in its dual form (eqn. 3) and maximize that problem.

$$\max(\sum_i \alpha_i - \sum_{i,j} \alpha_i\alpha_j y_i y_j \vec{x}_i^T \vec{x}_j) \quad \text{subject to} \quad \alpha_i \geq 0 \tag{3}$$

$$\vec{w} = \sum_i \alpha_i y_i \vec{x}_i \tag{4}$$

$$class(\vec{x}) = \vec{w}\cdot\vec{x} = \sum_i \alpha_i y_i \vec{x}_i \cdot \vec{x} + b \tag{5}$$

Once formulated in its dual form, quadratic programming is used to find the best $\alpha_i$ and $b$ from eqn. 3. This formulation allows the introduction of the "kernel trick" [39] and extends the classification ability of SVMs from generating classifications that are purely linear to a large variety of hypersurfaces in feature space.

SVMs may be viewed as an approach to find the $\vec{w}$ and $b$ that maximize $P(y = \pm 1|\vec{w}, b)$. When expressed in this form we can formulate the posterior distribution as in eqn. 6.

$$P(y = \pm 1|\vec{w}, b) = \frac{P(\vec{w}, b|y = \pm 1)P(y = \pm 1)}{P(\vec{w}, b)} \tag{6}$$

The denominator is a constant, and a shape model is needed to capture the $P(y = \pm 1)$ term. Expressed in this form, SVMs may be seen as approximating the posterior distribution using a given set of features to define $\vec{w}$ and $b$.

[70], [9], [15]

### B. AdaBoost

AdaBoost combines a set of weak learners in order to form a strong classifier in a "greedy fashion," i.e., it always chooses the weak classifier with the lowest error, ignoring all others.

We use a decision stump as a weak learner. A decision stump, based on a given feature, classifies all examples less than a threshold as belonging to one class and greater than a threshold as another class. Formally, a decision stump consists of a feature on which the decision will be made, a separating threshold, and a boolean saying whether positive examples are less than or greater than the threshold. A decision stump is advantageous over other weak learners because it can be calculated very quickly and there is a one-to-one correspondence between a weak learner and a feature when a decision stump is used.

Given: $N$ training examples $(\vec{x}_1, \cdots, \vec{x}_N)$ with $x \in \mathcal{X}$, corresponding labels $(y_1, \cdots y_N)$ with $y_i \in \{-1, 1\}$, and an initial distribution of weights $D_1(i)$ over the examples.
For $t = 1, ..., T$ :

- Train a weak classifier $h_t : \mathcal{X} \rightarrow \{-1, 1\}$ using distribution $D_t$.
- Calculate the error of $h_t : \epsilon_t = \sum_{i=1}^{N} D_t(i)\mathbf{1}(y_i \neq h_t(x_i))$.
- Set $\alpha_t = -\frac{1}{2}\log(\epsilon_t/(1-\epsilon_t))$.
- Set $D_{t+1}(i) = D_t(i)\exp(-\alpha_t y_i h_t(x_i))/Z_t$, where $Z_t = 2\sqrt{\epsilon_t(1-\epsilon_t)}$ is a normalization factor.

Output the strong classifier $H(x) = sign(f(x))$, where $f(x) = \frac{\sum_{t=1}^{T}\alpha_t h_t(x)}{\sum_{t=1}^{T}\alpha_t}$.

Fig. 1. Discrete AdaBoost algorithm. $\mathbf{1}$ is an indicator function.

AdaBoost explicitly seeks to minimize the error according to a distribution of weights, $D_t$, at each iteration. However, if we follow the logic of [50] and view $\{\alpha_t\}_{t=1}^{T}$ as a vector of coordinates, $\vec{\alpha}$, then we can rewrite $f(x)$ as eqn. 7.

$$f(x) = \frac{\vec{\alpha}\cdot\vec{h}(x)}{||\vec{\alpha}||_1} \tag{7}$$

Here we can view $\vec{\alpha}$ as a hyperplane and $\frac{1}{||\vec{\alpha}||_1}$ as the margin. We can then see that AdaBoost explicitly minimizes the error, and implicitly maximizes the margin according to the $l_1 - norm$ at each iteration, causing it to generalize well. Because AdaBoost greedily selects features, it can take a complicated problem, one composed of many features, and create a sparse classification rule, one composed of only a few features. However, this is also a drawback. Due to the greedy nature of AdaBoost it can only minimize the error, and maximize the margin with respect to features that have already been selected. AdaBoost is also limited by the fact that it can only combine weak learners by adding them together.

AdaBoost approximates the Bayesian posterior distribution by incrementally adding new weak learners $(h_i(x))$ at each iteration. This is equivalent to formulating the overall classifier at time $t$ as $H(x) = sign[P(y = \pm 1|h_1(x)\cdots h_t(x) > 0.5)]$ [53]. If we let $h_1(x)\cdots h_t(x) = h_t$, we can formulate the posterior distribution as eqn. 8.

$$P(y = \pm 1|h_t) = \frac{P(h_t|y = \pm 1)P(y = \pm 1)}{P(h_t)} \tag{8}$$

The denominator is again a constant and $P(y = \pm 1)$ is a shape model which must be integrated later. In this formulation, AdaBoost also approximates the ideal Bayesian distribution after a long enough $t$, drawing features from a very large feature pool.

We could stop here and just apply an ideal Bayesian classifier to the features selected by AdaBoost. For problems with a large number of i.i.d. examples that lie in a low-dimensional space, this would be ideal. However, our problem lies in a high-dimensional space, meaning that it would require a large number of i.i.d. examples for the Bayesian classifier to generalize well. Although we do have many examples, they are all correlated (non-i.i.d) and therefore the ideal Bayesian classifier would most likely be memorizing the posterior probability $P(x_1 \cdots x_t | y = \pm 1)$, resulting in poor generalization.

### C. SVM and AdaBoost Comparison

As one can see, SVMs globally and explicitly maximize the margin while minimizing the number of wrongly classified examples, using any desired linear or non-linear hypersurface. This is both an advantage and a disadvantage. The advantage is that SVMs take into account each example in the entire feature space when creating the separating hypersurface. The disadvantage is that this makes them computationally intractable as the number of features becomes large.

Because of this, one must either have prior knowledge of the features most suited for classification for the specific problem, or one must select them at runtime. Feature selection for classification is an area that has been explored before [70], [9], [15] specifically in the SVM domain. However, the goal of the previously published papers is slightly different than ours. In previous feature selection literature, the goal was to ascertain useless features over the space of all available features. To do this, all examples calculated at all features must be stored at once, and a convex minimization problem is performed over this matrix. However, we are allowing this matrix to grow to a size that is too large to be computationally tractable (the number of examples times the number of features exceeds the storage capacity of the computer). Since AdaBoost greedily selects features, it does not have this requirement. But, since it is a greedy feature selector, given the same set of features, we expect SVM to outperform AdaBoost. We exploit this fact to design our Ada-SVM classifier. We use AdaBoost to select the features that most accurately span the classification problem, and SVMs to fuse those features together to form the final classifier.

To make AdaBoost directly compatible with SVM, one small adjustment must be made to the AdaBoost algorithm. Traditionally, AdaBoost may choose features more than once when constructing weak learners; however, having the same feature appear twice in an SVM formulation does not make sense. To overcome this, when choosing features with AdaBoost for Ada-SVM, features are chosen without replacement. In all experiments involving just AdaBoost, however, traditional AdaBoost is implemented.

We implicitly take into account the Bayesian prior (shape information) necessary in both models by creating a shape prior based on the LogOdds formulation by Pohl et al. [45]. We create a signed distance map for each training subject, with negative values inside the ROI and positive values outside the ROI and then transform each of those values into the interval (0, 1) using eqn. 9, where $I(x)$ is the intensity of voxel $x$:

$$\forall\, x \in voxels \quad I'(x) = \frac{1}{1 + e^{-I(x)}} \tag{9}$$

After getting a signed distance map transformed into the interval (0, 1) for each subject, we then perform a voxel-by-voxel averaging in order to create one prior image that we store for both training and testing. We note that this map contains statistical information on the likely position of the target structure in the coordinate space to which all images have been aligned.

### D. Hierarchical Formulation

AdaBoost uses all image voxels as examples when choosing features to minimize the segmentation error. However, many voxels are easy to classify, and features that perform well on a lot of easy examples may perform poorly on examples that are more difficult to classify. To overcome this problem, we implement a decision tree framework.

Each node in the decision tree represents a new classifier using either AdaBoost or Ada-SVM with only those examples that reach that node. After classification two new child nodes are created, and examples are passed to the children. Using this approach, examples that are difficult to classify can be classified with different features than those that are easy to classify.

However, overfitting can be a problem when examples are only passed to one child or the other. Therefore, we employ a fuzziness factor based on the margin of both AdaBoost and SVM to control the overfitting problem. When a decision tree is based only on AdaBoost, if examples fall within the margin defined by $\frac{1}{||\alpha||_1}$ then those examples are passed to both children. When using a decision tree based on Ada-SVM, examples that fall within the SVM margin defined by $\frac{1}{||\vec{\alpha}||_2}$ are passed to both children.

An overview of the training process is given in Figure 2. To test the tree, an example, $x$, is given to the root node and its assignment is determined by the leaf classification.

Procedure for training an Ada-SVM tree of depth D

- Use AdaBoost to select important features and SVM to classify examples with those features over all examples
- Test all examples using SVM, and obtain two classes, positive and negative
- If an example falls within the margin, assign it to both classes
- If tree depth is greater than D, quit
- Recursively train the positive child on only the data which the previous node classified as positive
- Recursively train the negative child on only the data which the previous node classified as negative

Fig. 2. Procedure for Ada-SVM tree training. The AdaBoost tree is trained in the same way except that it does not use SVM for classification – it uses traditional AdaBoost for classification.

Although hierarchical AdaBoost [64] has already been applied to medical image segmentation [65], the Ada-SVM tree can be substituted anywhere that traditional hierarchical boosting is used to allow for a margin maximization based segmentation approach.

### E. Alzheimer's Disease Detection

In neuroscientific studies of disease, it is typical to compute average hippocampal maps for disease and control groups, visualizing regions with systematic anatomical differences in the form of 3D statistical maps. In one popular approach, 3D parametric surface models are fitted to each hippocampal segmentation and combined across subjects by geometrical averaging. These average shapes may be compared, and the effects of factors that may influence local hippocampal morphology can be tested statistically.

To examine the performance of our classifiers in constructing this type of map, the hippocampal surface points segmented by each approach were made uniform by modeling them as a 3D parametric surface mesh in each subject, as described in our prior work [59]. To create a measure of 'radial size' for each subject's hippocampus, first a medial curve was computed threading through the hippocampus, and the distance from each surface point to this curve was calculated, providing a measure that is sensitive to local atrophy. Rather than use the approach developed by Blum and colleagues for surface skeletonisation [8], which would in general yield a stratified set of surfaces, a medial curve was derived from the line traced out by the centroid of the boundary for each hippocampal surface model. The local radial size was defined for each boundary point as the radial distance between that boundary point and its associated medial curve, in that subject. As in prior work, regressions were performed to assign a *p*-value to each point on the surface in order to link radial size to different covariates of interest. Surface contractions and expansions were statistically compared between groups using Student's t tests, and were correlated with clinical characteristics (such as Mini-Mental State Exam (MMSE) scores [19]) to yield an associated significance value at each point. Finally the *p*-maps were presented as color coded average subcortical shapes.

This surface parametrization allows measurements to be made at corresponding surface locations in each subject. The procedure also allows the averaging of hippocampal surface morphological features across all individuals belonging to a group and records the amount of variation between corresponding surface points relative to the group averages. Several groups have used parametric surface meshes for hippocampal shape analysis based on sampled medial representations (*M*-reps) [56], conformal mappings, spherical harmonic or spherical wavelet analysis, or high-dimensional diffeomorphic metric mappings (LDDMM) [68], [37]. Some groups have also used parametric surface meshes for anatomical analyses using Gaussian random fields defined on surfaces [3] and for asymmetry quantification [34].

Here, for simplicity, we use a surface averaging approach used frequently in past studies [59], but we note that many methods to establish pointwise correspondence for hippocampal surfaces are under active development by our group and others [54], [69], [66], [57]. Some use automatically defined intrinsic geometric landmarks on the hippocampal surface to enforce higher-order correspondences across subjects when averaging anatomy across a group.

Given that independent statistical tests were made at many hippocampal surface points and statistics from adjacent data points are highly correlated, permutation testing was employed to control for multiple comparisons [59]. All our permutation tests are based on measuring the total area of the hippocampus with suprathreshold statistics, after setting the threshold at $p < 0.01$. To correct for multiple comparisons and assign an overall *p*-value to each *p*-map [41], [58], permutation tests were used to determine how likely the observed level of significant atrophy (proportion of suprathreshold statistics, with the threshold set at $p < 0.01$) within each *p*-map would occur by chance [58], [59]. The number of permutations $N$ was chosen to be 100,000, to control the standard error SEp of omnibus probability $p$, which follows a binomial distribution $B(N, p)$ with known standard error [14]. When $N = 8000$, the approximate margin of error (95% confidence interval) for $p$ is around 5% of $p$. We prefer to use the overall extent of the suprathreshold region as we know that atrophy is relatively distributed over the hippocampus, and a set-level inference is more appropriate for detecting diffuse effects with moderate effect sizes at many voxels, rather than focal effects with very high effect size (which would be better detected using a test for peak height in a statistical map).

When reporting permutation test results, one-sided hypothesis testing was used, i.e. we only considered statistics in which the AD group showed greater atrophy than the controls, in line with prior findings. Likewise, the correlations are reported as one-sided hypotheses, i.e. statistics are shown in the map where the correlations are in the expected direction, e.g. greater atrophy associated with lower MMSE scores. This type of map has revealed aspects of brain structure that predicts imminent onset of AD, but they have been time-consuming to compute in past studies, that have relied on hand segmentations [2], [1], [6], [18], [49].

### F. Surface Reconstruction Error

For the volumetric comparisons, the posterior probability map in each subject's scan was thresholded at the voxel level and supra-threshold voxels were counted without performing surface fitting. For the surface reconstructions, we followed the algorithm detailed for open parametric patch-like surfaces [61] and [62], which was modified to cope with closed tubular surfaces (logical cylinders) [59]. In test data, the polyline determined by the boundary contour in each section, sampled using 1 mm cubic voxels, is replaced by a uniformly parameterized curvilinear mesh of grid size 100x150 (these values were chosen empirically to give good reconstruction fidelity, given the resolution of MRI). The resulting network of sampled grid points always falls on the edges of the voxels in the classified bitmap, and implied geometric tiles on the surface are at most $1/\sqrt{2}$ or $\sim$0.7 mm away from the original bitmap in each section. Even so, the cross-group statistics are computed from the sampled grid points and not from the points interior to the surface tiles, and these are exactly on the boundary of the bitmap. As such, no additional reconstruction error is introduced in the surface relative to the classified bitmap. Needless to say, when the objects are replaced by binary objects with a resolution of 1 mm cubed, an upper

bound on the reconstruction error between the bitmap and the true object is $\sqrt{3}/2$ or less than one voxel. This may impact the maximum achievable overlap between different methods, and the reproducibility of segmentations in different scans.

## IV. Experiments

To facilitate fast development of our software, we used CImg [63] to do many basic image manipulations and an implementation of SVM called SVMPerf developed by Joachims [27] for SVM analysis. We also made use of the LONI Pipeline environment (http://pipeline.loni.ucla.edu), which was developed by the Laboratory of Neuro Imaging, for fast and easy parallel processing [48].

Before performing classification, we registered all of the brain images into the same stereotaxic space. Each subject's brain MRI was co-registered with scaling (9-parameter transformation) to the ICBM53 average brain template [13]. Since this registration involves scaling, global scaling is removed during this stage of pre-processing. Because of this pre-processing step, we do not have to restrict our attention to rotation, scaling, or translation invariant features. This also allows us to define a bounding box around the training hippocampi plus some neighborhood voxels. These neighborhood voxels might contain hippocampal voxels outside the bounding box of the training set and are also necessary for computing neighborhood based features. Any voxels outside of this bounding box are definitely not hippocampus, and can therefore be ignored by our classifier. For all our experiments, our bounding box is a rectangular region with corners at (-48, -54, -44) and (-1, 5, 17) for the left hippocampus and, a corresponding region in the opposite hemisphere for the right hippocampus in the standard ICBM53 space [36].

Next, we have to define our pool of candidate features from which AdaBoost will select. The important conditions that must be taken into account are robustness to noise, sensitivity to local differences in image intensity and structure shape, and most importantly calculation speed. Our feature pool consists of information from three different image "channels": (1) the T1-weighted image, (2) tissue classification maps of gray matter, white matter, and CSF (obtained by an unsupervised classifier, PVC [51]), and (3) our Bayesian shape prior (eqn. 9). From each one of these images, the following features are computed: intensity, gradients, curvatures, 1D, 2D, and 3D Haar filters, mean filters, and standard deviation filters, all computed using a neighborhood kernel of size 7x7x7. Because of the large number of examples and features, we use randomization to decrease these numbers to a computationally tractable size. During each run of AdaBoost, a new set of 200,000 examples and 2500 features is randomly chosen to learn the classification rule (for either AdaBoost or Ada-SVM). These numbers were determined empirically to give optimal results.

Additionally, when running SVM there are several parameters that need to be specified. We found that using a polynomial kernel of order 3, with a b value of 0 and a C value of 20 gave the best results (eqn. 2). Most of these parameters were the defaults for the SVM implementation we used [27], with the only exception being the kernel choice, which was also chosen empirically.

As a final step, after segmentations are computed by either AdaBoost, Ada-SVM, or manual SVM, the binary masks are convolved with a 3x3x3 averaging kernel. Partial volume effects are removed from the resulting mask by setting voxels with a value of less than 0.5 (those with fewer than 13 neighbors) to 0 and greater than 0.5 (those with more than 13 neighbors) to 1. This is done to smooth the boundary and fill any holes.

### A. Evaluation Metrics

To assess the accuracy of our methods, we report some standard error metrics. To define each error metric we define 2 sets $A$ and $B$, where $A$ is the set of hippocampal voxels as defined by the manual segmentation and $B$ is the set of hippocampal voxels as defined by automatic segmentation. Now, we define precision (eqn. 10), recall (eqn. 11), relative overlap (R.O.) (eqn. 12), and similarity index (S.I.) (eqn. 13).

$$Precision = \frac{volume(A \cap B)}{volume(B)} \tag{10}$$

$$Recall = \frac{volume(A \cap B)}{volume(A)} \tag{11}$$

$$RelativeOverlap = \frac{volume(A \cap B)}{volume(A \cup B)} \tag{12}$$

$$SimilarityIndex = \frac{volume(A \cap B)}{\frac{volume(A)+volume(B)}{2}} \tag{13}$$

Additionally, we compute two distance metrics, Hausdorff distance [33] and mean distance. Hausdorff distance and mean distance are defined by equations 14 and 15, where $A$ and $B$ are all points in the volumes and $d(a, b)$ is the Euclidean distance between points $a$ and $b$. Because the Hausdorff distance is not symmetric, we make it symmetric by formulating it as $\frac{H(A,B)+H(B,A)}{2}$.

$$H(A, B) = max_{a \in A}(min_{b \in B}(d(a, b))) \tag{14}$$

$$M(A, B) = mean_{a \in A}(min_{b \in B}(d(a, b))) \tag{15}$$

It would be of interest to determine what added advantage the many additional features provide over the basic prior term used for approximate specification of statistics on structure position. However, the prior is such a strong constraint on the final labeling that it is not clear that some of the algorithms could operate without it, so a fair comparison would be difficult. For instance, it is not clear that FreeSurfer can be run without a prior, as the intensity distributions and adjacency priors are the main features used for segmentation. As the first two features selected by AdaBoost are based on the mean and Haar filters derived from the prior, we know that the selected additional features provably show additional error reduction on the test set (via the AdaBoost rule).

### B. Manual SVM v. Ada-SVM

In order to show the importance of automatic feature selection, we compare manual SVM and Ada-SVM. As noted already, manual SVM feeds a set of features chosen by the user into SVM, while Ada-SVM decides which features to use via the automated learning rules that are part of the AdaBoost method. In what follows, for the manually-guided SVM, our feature vector was chosen to be the same length as that learned by Ada-SVM (100 features) and consisted of intensity, x, y, z positions, mean curvatures defined over small neighborhoods, x, y, z intensity gradients, standard deviation filters, and Haar filters in 3D.

| | **Ada-SVM** | | **Manual SVM** | |
|---|---|---|---|---|
| | Left | Right | Left | Right |
| Precision | 0.785 | 0.802 | 0.364 | 0.755 |
| Recall | 0.851 | 0.848 | 0.973 | 0.719 |
| R.O. | 0.691 | 0.701 | 0.360 | 0.582 |
| S.I. | 0.814 | 0.822 | 0.526 | 0.732 |
| Hausdorff | 4.34 | 4.63 | 6.05 | 6.83 |
| Mean | 0.029 | 0.034 | 0.384 | 0.047 |

TABLE I

PRECISION, RECALL, RELATIVE OVERLAP, SIMILARITY INDEX, HAUSDORFF DISTANCE, AND MEAN DISTANCE MEASURES ARE REPORTED FOR MANUAL-SVM AND ADA-SVM. DISTANCE MEASURES ARE EXPRESSED IN MILLIMETERS. R.O. DENOTES RELATIVE OVERLAP, AND S.I. DENOTES SIMILARITY INDEX, AS DEFINED IN THE TEXT. THESE RESULTS ARE MEASURED ON THE TESTING DATASET.

Table I shows the large discrepancy between manual SVM and Ada-SVM (especially on the left side). This illustrates the necessity for using informative features. This means that an expert must select features which are appropriate for the dataset at hand each time a new problem is proposed, or use an automatic feature selection method. Due to this fact, for the remainder of the paper, we will not consider manual SVM. In order to emphasize this table II gives the first ten features selected by AdaBoost. Notice the wide variety of types and shapes of features selected, making manually choosing these features very difficult.

### C. Comparison to Manual Segmentations

Fig. 3 shows some of our segmentation results. Compared with the manual gold standard, Ada-SVM gives a smoother boundary and is visually close to the tracings obtained by hand. Both AdaBoost and FreeSurfer give a more jagged but visually reasonable segmentation.

Our overlap and distance metrics compare well with segmentations from FreeSurfer [16], as shown by Table III. Note that for each error metric tested, the training results are slightly better than the testing results. This is to be expected; however it is important to note that the metrics are only slightly worse in the testing case. This suggests that both AdaBoost and Ada-SVM are not memorizing the data, but learning a generalizable model. Also note that for each metric in the testing case Ada-SVM gave the best results, AdaBoost the next best, and FreeSurfer the worst. FreeSurfer also had the most visually inconsistent segmentations (fig. 3). In fairness, FreeSurfer provides segmentations of many brain structures other than the hippocampus; future work with Ada-SVM will examine how it generalizes to other structures. Even so, the time efficiency of our approach (it takes about 3-5 minutes per brain), at least for the steps after the training phase, is advantageous given the large scale of AD morphometry studies now underway (e.g., $N$=3000 [26]).

Table IV gives some error metrics reported by other semi- and fully automated approaches. These numbers are presented only to show that our methods are close to theirs since an exact comparison is not possible without using the same data. This is evident by the fact that the numbers reported by Fischl [16] are different from the numbers we are achieving by their algorithm on the data tested here.

One more question that we want to answer is how many brains must be labeled by hand, in a given dataset, in order to get an acceptably low test error. While this may depend on the image contrast and the power required for the study, it is still possible to test how robust the segmentations are to deliberate reductions in the size of the training set. To measure this, we plot the error in the test set, against the number of brains used in the training set. We expect that performance would inevitably degrade with reductions in the training set size, but that extensive increases in the training set would give diminishing returns, with asymptotic convergence to a maximum obtainable accuracy. Each point in Fig. 4 represents randomly varying the number of training brains, and testing on all 40 test brains each time. Fig. 4 suggests that for each of both AdaBoost and Ada-SVM about 20 brains is the point of diminishing returns. One can note a slight increase in the error when using 25 brains for Ada-SVM on the left hippocampus. This is due to the randomization processes for both feature and example selection, and such small perturbations are ordinary.

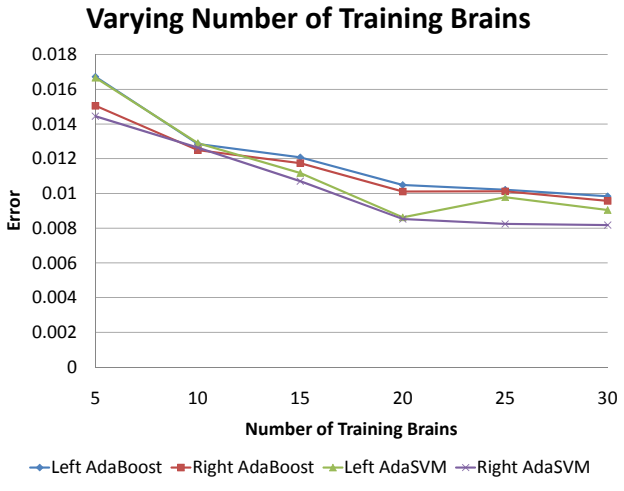


Fig. 4. The effect of varying the size of the training set versus the error between automated and gold standard manual segmentations. Error is defined on the number of incorrectly classified voxels inside the bounding box. Values are obtained for 5, 10, 15, 20, 25, and 30 brains. Note that the curves level off after 20 brains indicating diminishing returns by using more than 20 brain on which to train.

| Left | | | Right | | |
|---|---|---|---|---|---|
| **Channel** | **Name** | **Neighborhood** | **Channel** | **Name** | **Neighborhood** |
| Prior Image | Mean Filter | 3,6,3 | Prior Image | Mean Filter | 6,7,3 |
| Prior Image | Haar Filter | 6,3,3 (3D) | Prior Image | Haar Filter | 7,7,6 (3D) |
| Prior Image | Mean Filter | 6,5,1 | T1-weighted Image | Haar Filter | 7,7,3 (3D) |
| Tissue Classification Image | Mean Filter | 3,1,1 | Prior Image | Standard Deviation Filter | 7,6,6 |
| Prior Image | Standard Deviation Filter | 7,6,6 | Prior Image | Haar Filter | 5,4,5 (3D) |
| T1-weighted Image | Haar Filter | 4,5,7 (3D) | Tissue Classification Image | Intensity | n.a. |
| Prior Image | Haar Filter | 1,3,6 (3D) | Prior Image | Mean Filter | 6,5,1 |
| Prior Image | Haar Filter | 7,7,2 (2D) | Prior Image | Gradient Filter | 5,7,2 (y) |
| T1-weighted Image | Intensity | n.a. | T1-weighted Image | Haar Filter | 3,1,7 (3D) |
| Tissue Classification Image | Haar Filter | 3,2,4 (3D) | Prior Image | Haar Filter | 5,3,1 (2D) |

TABLE II

THESE ARE THE FIRST TEN FEATURES SELECTED BY THE ADABOOST ALGORITHM (USED DURING BOTH ADABOOST AND ADA-SVM). NOTE THE VARIETY OF SIZES, SHAPES, TYPES, AND CHANNELS SELECTED MAKING IT VERY DIFFICULT TO DISCERN A PATTERN FOR MANUAL FEATURE SELECTION. ALTHOUGH IT IS INTERESTING TO NOTE WHICH FEATURES ARE CHOSEN BY ADABOOST FOR A GENERAL OVERVIEW OF THE PROBLEM, DETAILED STUDY OF THESE FEATURES WILL NOT IN GENERAL GIVE A LOW-DIMENSIONAL SUBSET OF FEATURES FOR BUILDING A CLASSIFIER. ADABOOST IS DESIGNED TO SELECT WEAK LEARNERS (OR FEATURES) FROM A VERY LARGE POOL TO LEARN A POSTERIOR DISTRIBUTION SPECIFIC TO ONE DATASET, AND GENERALLY EACH OF THESE FEATURES MAY PERFORM ONLY SLIGHTLY BETTER THAN CHANCE. THERE IS NO EXPECTATION THAT THESE FEATURES WILL GENERALIZE WELL TO CREATING A MODEL FOR OTHER SUBCORTICAL STRUCTURES, OR EVEN HIPPOCAMPI FROM OTHER STUDIES.

| | **Ada-SVM** | | | | **AdaBoost** | | | | **FreeSurfer** | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Left | | Right | | Left | | Right | | Left | Right |
| | Training | Testing | Training | Testing | Training | Testing | Training | Testing | Testing | Testing |
| Precision | 0.821 | 0.785 | 0.844 | 0.802 | 0.792 | 0.771 | 0.777 | 0.760 | 0.716 | 0.737 |
| Recall | 0.868 | 0.851 | 0.848 | 0.848 | 0.841 | 0.828 | 0.827 | 0.839 | 0.743 | 0.732 |
| R.O. | 0.728 | 0.691 | 0.732 | 0.701 | 0.687 | 0.665 | 0.666 | 0.663 | 0.572 | 0.577 |
| S.I. | 0.841 | 0.814 | 0.845 | 0.822 | 0.813 | 0.795 | 0.797 | 0.795 | 0.726 | 0.729 |
| Hausdorff | 4.04 | 4.34 | 4.41 | 4.63 | 4.64 | 4.98 | 5.20 | 4.83 | 4.97 | 4.99 |
| Mean | 0.019 | 0.029 | 0.018 | 0.034 | 0.024 | 0.028 | 0.027 | 0.041 | 0.075 | 0.065 |

TABLE III

PRECISION, RECALL, RELATIVE OVERLAP, SIMILARITY INDEX HAUSDORFF DISTANCE, AND MEAN DISTANCE MEASURES ARE REPORTED FOR TRAINING AND TESTING DATA FROM EACH SEGMENTATION ALGORITHM: ADA-SVM, ADABOOST, AND FREESURFER [16]. DISTANCE MEASURES ARE EXPRESSED IN MILLIMETERS. R.O. DENOTES RELATIVE OVERLAP, AND S.I. DENOTES SIMILARITY INDEX, AS DEFINED IN THE TEXT. NOTE THAT THE ADA-SVM TESTING NUMBERS ARE THE SAME AS REPORTED IN TABLE I, AND ARE REPRODUCED HERE FOR CONSISTENCY.

| | Left | | | Right | | |
|---|---|---|---|---|---|---|
| | [46] (N = 5) | [16] (N = 134) | [25] (N = 5) | [46] (N = 5) | [16] (N = 134) | [25] (N = 5) |
| Recall | 0.82 | n.a. | n.a. | 0.83 | n.a. | n.a. |
| R.O. | 0.72 | 0.78 | 0.74 | 0.74 | 0.80 | 0.76 |
| S.I. | 0.84 | n.a. | n.a | 0.85 | n.a. | n.a. |

TABLE IV

RECALL, RELATIVE OVERLAP (R.O.), AND SIMILARITY INDEX (S.I.) MEASURES ARE REPORTED FOR THREE OTHER PUBLISHED METHODS. THESE ARE PRESENTED ONLY FOR A ROUGH COMPARISON AS AN EXACT COMPARISON IS NOT POSSIBLE UNLESS THE SAME DATASET IS USED. NOTE FOR INSTANCE THE VARIABILITY BETWEEN WHAT IS REPORTED BY FISCHL [16] AND THE METRICS REPORTED FOR THE SAME METHOD IN THIS PAPER.

### *D. Disease Detection*

In addition to segmentation accuracy, it is also important to assess how effectively each method can differentiate disease from normal. For instance, in a study aiming to map disease effects, increases in segmentation accuracy are beneficial if they provide additional power to differentiate groups. As the effect of AD on the brain is not uniform, such studies commonly rely on mapping of group differences to identify regions that are especially susceptible to early changes, or where changes predict imminent decline or help differentiate one type of dementia from another [49]. We note that in reporting classification accuracy and detection of disease effects on hippocampal anatomy in groups of subjects, both of these metrics evaluate desirable characteristics of a tissue segmentation approach, but they are not necessarily causally related or even correlated. That is, a method that produces relatively better segmentation is not necessarily more discriminative and vice versa, and it is misleading to suggest that one implies the other. From a logical standpoint, there could be a bad segmentation algorithm that exaggerates the difference between AD and controls, for example, and this could be a very good discriminator. In general, this depends on whether the voxels that are misclassified by a segmentation approach are also relevant for disease classification.

First, Table V shows the percent difference in agreement with manual tracings between all subjects, and subjects broken down by diagnosis. We do this by taking the difference between an error metric broken down by disease and the same error metric on all subjects and dividing by the error metric broken down by disease. Positive percentages indicate

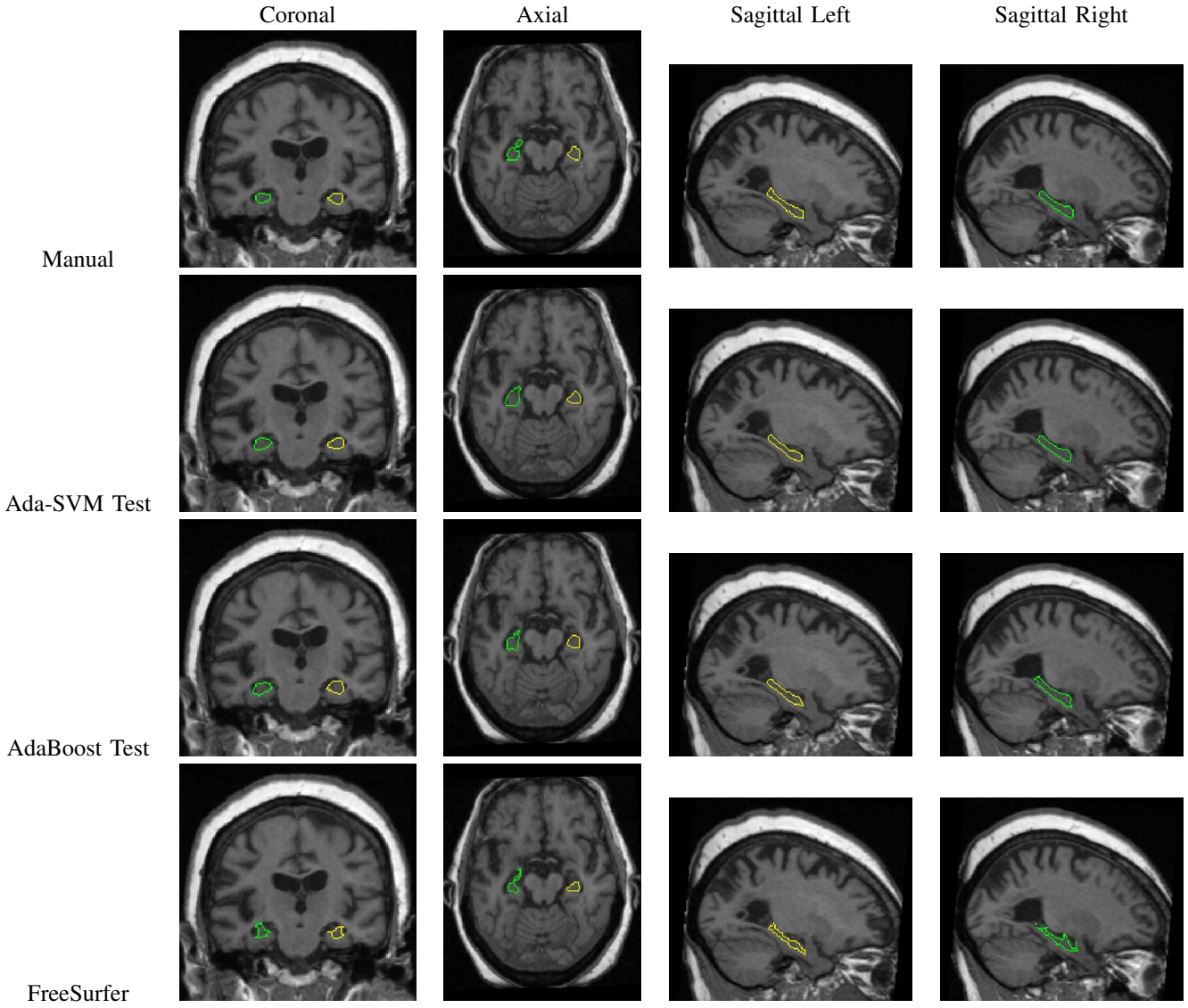


Fig. 3. Example hippocampal segmentations from each of the methods being compared: manual tracing, Ada-SVM, AdaBoost, and FreeSurfer [16]. The left hippocampus is shown in yellow, and the right hippocampus is shown in green. Ada-SVM gives smoother, more spatially coherent result than any of the other methods, and even appears slightly less noisy than the manual traces, which are typically created in coronal sections and may appear jagged when resliced in other planes. All methods give anatomically reasonable segmentations, but some give highly irregular or noisy boundaries. Here we show only the test cases for Ada-SVM and AdaBoost, because we wish to evaluate their performance on unseen images, not on the same manually segmented images that were used for training. The brain MRI quality here is typical of those used in AD morphometric studies, showing widespread atrophy and moderate to poor gray/white matter contrast.

that a given metric shows better performance on a specific diagnostic group (e.g. the controls) relative to the performance on all subjects combined, while negative percentages indicate a worsening in a given metric in a specific diagnostic group, relative to the performance on all subjects combined. For almost all error metrics, the normal group was segmented more accurately than the AD group, which is to be expected because there is less variance in the normal group, and disease-related atrophy can greatly distort the geometry of the structure. Secondly, for three out of the four volumetric measurements (with the exception of precision), Ada-SVM gives a more consistent segmentation for both normal and AD subjects (the distance metrics are too prone to outliers to be very useful in this table, and many of them show a better segmentation for AD than normal). This can be identified by the smaller absolute value of most error metrics when comparing methods.

Fig. 5 shows our results for mapping disease effects on the hippocampus, and for detecting associations between hippocampal atrophy and cognitive performance on the MMSE, a widely-used test in studies of AD. Strictly speaking, we do not have ground truth regarding the extent of anatomical atrophy, but it is reasonable that an approach that detects atrophy, while controlling for false positives at the accepted rate (by permutation testing) is more valuable than one that fails to detect atrophy (see below for more discussion of this premise). The overall pattern of atrophy in the maps based on the manual traces is also in strong agreement with past studies of hippocampal atrophy in independent samples of

| | **Ada-SVM** | | | | **AdaBoost** | | | | **FreeSurfer** | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Left | | Right | | Left | | Right | | Left | | Right | |
| | Normal | AD | Normal | AD | Normal | AD | Normal | AD | Normal | AD | Normal | AD |
| Precision | 5.13 | -5.71 | 0.90 | -0.92 | 6.13 | -6.98 | 2.29 | -2.40 | 1.21 | -1.24 | -1.56 | 1.51 |
| Recall | 1.10 | -1.12 | 0.17 | -0.17 | 1.66 | -1.72 | 1.22 | -1.25 | 5.86 | -6.63 | 6.29 | -7.19 |
| R.O. | 5.10 | -5.68 | 0.94 | -0.96 | 6.25 | -7.15 | 2.90 | -3.08 | 5.57 | -6.27 | 4.23 | -4.62 |
| S.I. | 3.31 | -3.54 | 0.64 | -0.65 | 4.16 | -4.54 | 1.94 | -2.02 | 3.77 | -4.07 | 2.89 | -3.06 |
| Hausdorff | -12.42 | 9.95 | 4.38 | -4.80 | -15.50 | 11.83 | 2.28 | -2.39 | -8.67 | 7.39 | -0.68 | 0.67 |
| Mean | -79.09 | 15.04 | -70.44 | 0.38 | -92.32 | 15.07 | -42.73 | -8.52 | -46.22 | -32.45 | -33.73 | -43.54 |

TABLE V

THESE ARE THE PERCENT DIFFERENCES IN PRECISION, RECALL, RELATIVE OVERLAP, SIMILARITY INDEX, HAUSDORFF DISTANCE, AND MEAN DISTANCE, MEASURES FROM ALL SUBJECTS (TABLE III) TO SUBJECT GROUPS BROKEN DOWN BY DIAGNOSIS. WE TOOK THE DIFFERENCE BETWEEN AN ERROR METRIC ON JUST A DIAGNOSTIC GROUP AND THE SAME ERROR METRIC ON ALL SUBJECTS AND THEN DIVIDED THAT BY THE ERROR METRIC FOR A DIAGNOSTIC GROUP. POSITIVE NUMBERS REPRESENT AN INCREASE IN A METRIC, AND NEGATIVE NUMBERS A DECREASE. THE MAIN THING TO NOTE FROM THIS TABLE IS HOW ALMOST ALL ERROR METRICS SHOW A BETTER SEGMENTATION FOR THE NORMAL GROUP THAN FOR THE AD GROUP. THIS IS NATURAL BECAUSE DEGENERATIVE DISEASES ARE OFTEN ACCOMPANIED BY A REDUCTION IN THE CONTRAST BETWEEN GRAY AND WHITE MATTER, AND DISEASE-RELATED ATROPHY CAN GREATLY DISTORT THE GEOMETRY OF THE STRUCTURE.

subjects with AD, showing widespread volume reductions in both the hippocampal head and tail [1], [18], [49]. All methods tested show widespread areas of significance. This shows that each method is correlating both diagnosis and MMSE well with radial atrophy. These observations are confirmed by the permutation tests of VI. Each entry is table VI is well below the significance level of 0.05. In a morphometric study of AD, these corrected significance values would be used to determine whether a disease effect had been detected. Based on several prior papers [59], [1], it is known that hippocampal atrophy correlates with MMSE scores in AD, and it is important in a morphometric study to establish that the atrophy detected is correlated with a meaningful behavioral measure or outcome measure for the patient, rather than just correlating with diagnosis [11].

Perhaps surprisingly, in the discriminative pattern shown in Fig. 5 (e.g., in the left column comparing AD with normal controls), AdaBoost methods find significant discriminative effects in the regions where manual segmentations do not. This is quite possible because the inter-rater reliability for manual segmentation is not spatially homogeneous, and there are some regions where it is more difficult for a human rater to segment the hippocampus accurately (the easiest region is typically the posterior hippocampus, and the hardest region is typically the anterior junction with the amygdala, where there is poor contrast between the two boundaries). If the image based criteria are more consistent than humans in identifying a boundary in the image in certain regions, they will tend to offer more statistical power in detecting systematic alterations in these regions.

To emphasize the differences between segmentation methods, we plotted the cumulative distribution function of the *p*-values in the maps, against the corresponding *p*-values that would be expected under the null hypothesis of no group difference (Fig. 6). For a null distribution, this cumulative plot falls along the line $y = x$, as represented by the black line. Larger upward inflections of the CDF curve near the origin are associated with significant signal, and greater effect sizes are represented by larger deviations (the theory of false discovery rates gives formulae for thresholds that control false positives at a known rate). For the association of diagnosis with

| | Diagnosis | | MMSE | |
|---|---|---|---|---|
| | Left | Right | Left | Right |
| Manual | **0.00173** | **0.00011** | **0.0112** | **0.00011** |
| Ada-SVM | **0.00011** | **0.00013** | **0.00061** | **0.0001** |
| AdaBoost | **0.00028** | **0.0002** | **0.0003** | **0.00011** |
| FreeSurfer | **0.00011** | **0.0001** | **0.001** | **0.00012** |

TABLE VI

PERMUTATION VALUES FOR DISEASE EFFECTS AND ASSOCIATIONS WITH COGNITIVE PERFORMANCE, BASED ON CREATING GROUP DIFFERENCE MAPS FROM 100,000 DIFFERENT RANDOM ASSIGNMENTS OF SUBJECTS TO GROUPS. IF AN OMNIBUS PROBABILITY (I.E., CORRECTED FOR MULTIPLE COMPARISONS) IS DETERMINED BY COMPARING THE NUMBER OF SUPRA-THRESHOLD VOXELS IN THE TRUE LABELING TO THE PERMUTATION DISTRIBUTION, THE NUMBER OF PERMUTATIONS $N$ MUST BE CHOSEN TO CONTROL THE STANDARD ERROR $SEp$ OF OMNIBUS PROBABILITY P, WHICH FOLLOWS A BINOMIAL DISTRIBUTION $B(N, p)$ WITH $SEp = \sqrt{p(1-p)/N}$ [14].

radial atrophy both manual tracings and FreeSurfer appear to perform best, followed closely by Ada-SVM and finally by AdaBoost. For the MMSE associations AdaBoost appears the best followed by Ada-SVM, manual tracings and finally FreeSurfer. The main point to note from these graphs are that all methods show a large significant value (very different from the $y = x$ line), and no clear winner can be determined. In order to show one method is clearly better than another a more sensitive correlation must be looked for (such as the correlation between normals and MCI with atrophy or MCI and AD with atrophy), however due to data limitations such experiments were not possible at this time. This CDF approach has been used in Leporé et al. [31] to compare effect sizes in TBM, and is based on the False Discovery Rate concept used in imaging statistics for multiple comparisons correction [55].

### *E. Inter Dataset Comparison*

As a final comparison, we used our models that were trained on the data used throughout this paper, and tested on data drawn from ADNI [26]. We performed this test as a way to show how our method generalizes between datasets. Table VII

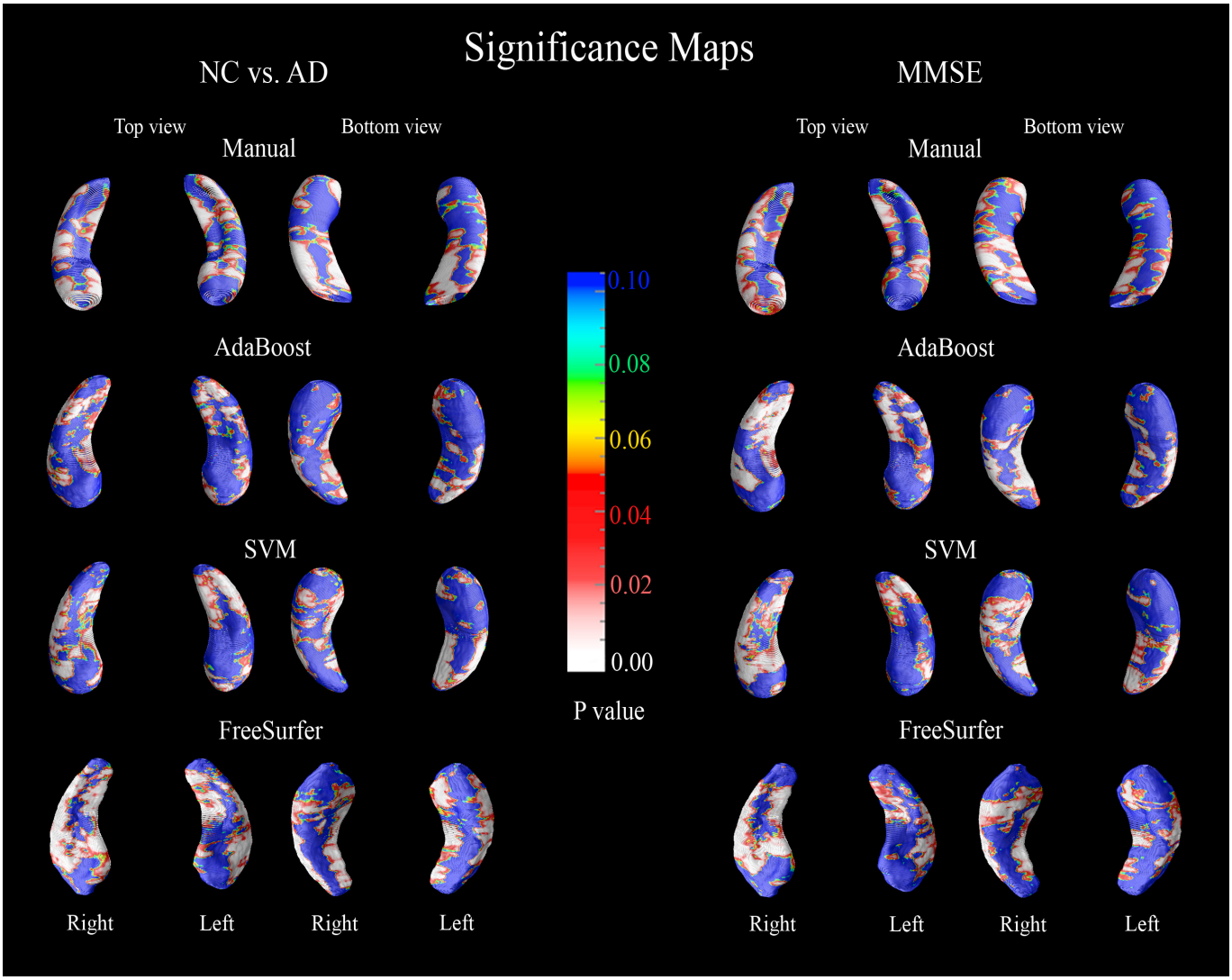


Fig. 5. Significance maps ($p$-maps) based on manual, Ada-SVM, AdaBoost, and FreeSurfer segmentations

shows our results.

| | **Ada-SVM** | | **AdaBoost** | |
|---|---|---|---|---|
| | Left | Right | Left | Right |
| Precision | 0.195 | 0.353 | 0.237 | 0.375 |
| Recall | 0.315 | 0.557 | 0.320 | 0.471 |
| R.O. | 0.144 | 0.299 | 0.164 | 0.279 |
| S.I. | 0.234 | 0.428 | 0.267 | 0.414 |
| Hausdorff | 13.86 | 13.13 | 12.88 | 12.00 |
| Mean | 0.435 | 0.346 | 0.322 | 0.292 |

TABLE VII

PRECISION, RECALL, RELATIVE OVERLAP, SIMILARITY INDEX HAUSDORFF DISTANCE, AND MEAN DISTANCE MEASURES ARE REPORTED FOR MODELS THAT WERE TRAINED ON THE DATA USED THROUGHOUT THIS PAPER AND TESTED ON THE ADNI DATA. THESE RESULTS ARE WELL BELOW THE OTHER REPORTED RESULTS SHOWING THAT THIS METHOD IS ONLY USEFUL WHEN TRAINED AND TESTED ON DATA DRAWN FROM THE SAME STUDY.

As shown in Table VII, the metrics are well below other error metrics reported for both our own method and FreeSurfer. This gives two important conclusions. First, our method is not scan parameter independent. Since we only use a linear registration and our features are not independent on the MRI parameters used to acquire the scans, our model does poorly on and inter-dataset basis. This shows that our model is most useful when a large cohort of subjects need to be analyzed, where hand segmenting 20 subjects will allow the rest of the study to be automatically segmented. Secondly, this shows that FreeSurfer is robust to inter-dataset variability. This is due to the fact that FreeSurfer uses a very accurate non-linear registration algorithm and a strong prior. FreeSurfer, therefore, lends itself better to segmenting hippocampi from smaller studies.

## V. CONCLUSIONS AND FUTURE WORK

While manual segmentation detects differences with greatest effect size, it can become prohibitively difficult if the number of MRI's in a study is very large. We have shown some evidence that Ada-SVM may perform better than AdaBoost and FreeSurfer in finding the approximate boundary of the hippocampus. We have also shown that all methods are capable of

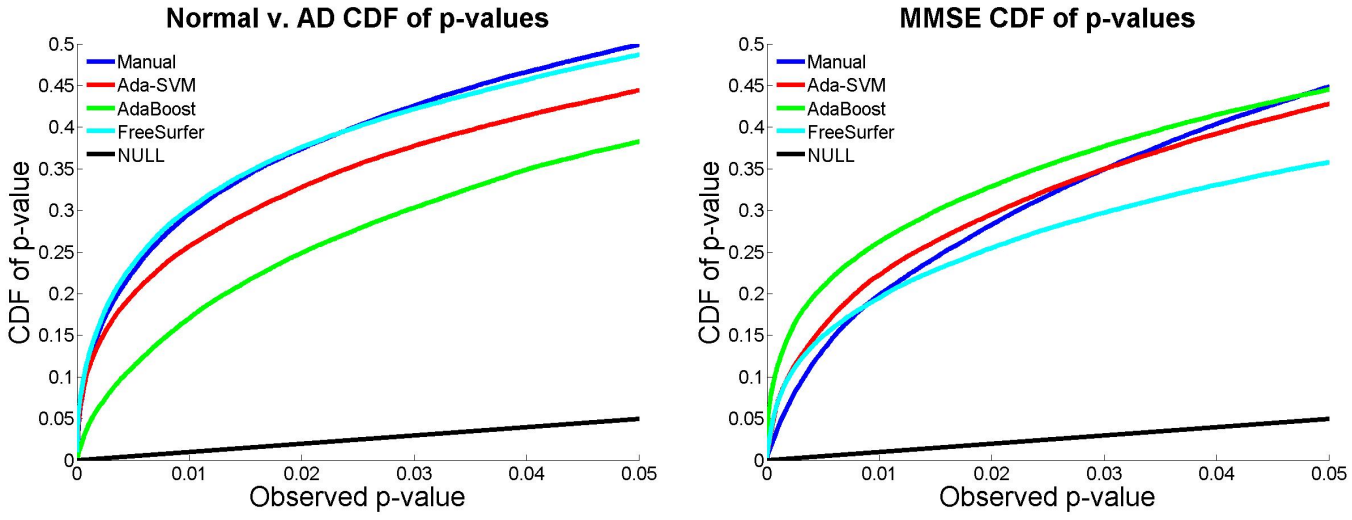


Fig. 6. Cumulative distribution of $p$-values for different methods. (a) shows the $p$-values when the covariate is the Alzheimer's disease diagnosis. (b) shows the $p$-values when the covariate is the MMSE [19] score. These CDF plots are commonly generated when using false discovery rate methods to assign overall significance values to statistical maps [7], [20], [55]; they may also be used to compare effect sizes of different methods, subject to certain caveats [31], as they show the proportion of supra-threshold voxels in a statistical map, for a range of thresholds. A cumulative plot of $p$-values in a statistical map, after the $p$-values have been sorted into numerical order, can compare the proportion of supra-threshold statistics with null data, or between one method and another, to assess their power to detect statistical differences that survive thresholding at both weak and strict thresholds (in fact at any threshold in the range [0,1]). In the examples shown here, the cumulative distribution function of the $p$-values observed for the statistical comparison of patients versus controls is plotted against the corresponding $p$-value that would be expected, under the null hypothesis of no group difference (shown here in black).

capturing both disease related effects and correlations between cognition and structure for these well known, widespread effects.

Although for the experiments shown in this paper, both AdaBoost and Ada-SVM outperform FreeSurfer, this is not a completely fair test. FreeSurfer was trained with different raters on different images, and although we test on different subjects than we train on, all subjects are still from the same study. An important subject of future study is the generalizability of the methods proposed here as a function of the MRI acquisition parameters. Even though this is interesting, this method is already useful in large scale studies for which the scanning parameters remain stable (such as ADNI [26]), where one could segment 20 brains manually for training purposes, and then segment all the rest automatically.

In the future, we will apply both of these techniques to new datasets to examine different diseases and to rank segmentation methods for power and accuracy. It will be interesting to note if Ada-SVM more powerfully detects disease effects or segments other subcortical structures better than AdaBoost does.

Although the ability to map disease effects automatically is encouraging and likely to benefit many ongoing studies, one caveat is necessary regarding the use of $p$-value plots to compare the effect sizes of different methods. These plots provide a clear comparison of the distribution of effect sizes in a statistical map when methodological parameters are varied, strictly speaking, many repeated large and independent samples would be required to prove that one cumulative $p$-value distribution differs from another on the interval [0,1]. Without confirmation on multiple samples, it may not reflect a reproducible difference between methods. FDR and its variants [29], [55] declare that a CDF shows evidence of a signal if it rises more than 20 times more sharply than a null distribution, so a related criterion could be developed to compare two empirical mean CDFs after multiple experiments. As simple numeric summaries sacrifice much of the power of maps, and provide a rather limited view of the differences in sensitivity among voxel-based mapping methods, additional work on CDF-based comparisons of methods seems warranted.

In addition, although the results presented here are anatomically congruent with hippocampal mapping studies in Alzheimer's disease, strictly speaking, we do not have ground truth regarding the extent and degree of hippocampal atrophy in AD. So, although an approach that finds greater effect sizes in disease is likely to be more accurate and valuable than one that fails to detect disease, it would be better to compare these models in a predictive design where ground truth regarding the dependent measure is known (i.e., morphometry predicting future atrophic change, future cognitive deterioration, or drug response). We are collecting this data at present. Any association between the segmentation method employed and the resulting power for a predictive model may allow a stronger statement regarding the relative power of AdaBoost variants for hippocampal mapping versus manual or FreeSurfer segmentations.

## VI. Acknowledgments

Grant support for this work was provided by the National Institute for Biomedical Imaging and Bioengineering, the National Center for Research Resources, National Institute on Aging, the National Library of Medicine, and the National Institute for Child Health and Development (EB01651, RR019771, HD050735, AG016570, LM05639 to P.M.T.) and by the National Institute of Health Grant U54 RR021813

(UCLA Center for Computational Biology). L.G.A. was also supported by NIA K23 AG026803 (jointly sponsored by NIA, AFAR, The John A. Hartford Foundation, the Atlantic Philanthropies, the Starr Foundation and an anonymous donor) and NIA P50 AG16570.

## REFERENCES


[1] L. Apostolova, I. Dinov, R. Dutton, K. Hayashi, A. Toga, J. Cummings, and P. Thompson, "3D comparison of hippocampal atrophy in amnestic mild cognitive impairment and Alzheimer's disease," *Brain*, Oct. 2006 [Epub ahead of print].

[2] L. Apostolova, R. Dutton, I. Dinov, K. Hayashi, A. Toga, J. Cummings, and P. Thompson, "Conversion of mild cognitive impairment to Alzheimer's disease is predicted by hippocampal atrophy maps," *Archives of Neurology*, vol. 63, pp. 693–699, May 2006.

[3] R. Bansal, L. Staib, D. Xu, H. Zhu, and B. Petersen, "Statistical analyses of brain surfaces using Gaussian random fields on 2-D manifolds," *IEEE TMI*, vol. 26, no. 1, pp. 46–57, Jan. 2007.

[4] C. Bearden, J. Soares, A. Klunder, M. Nicoletti, N. Diershke, K. Hayashi, P. Brambilla, R. Sassi, D. Axelson, N. Ryan, B. Birmaher, and P. Thompson, "Three-dimensional mapping of hippocampal anatomy in early-onset bipolar disorder," June 2007 submitted.

[5] C. Bearden, P. Thompson, R. Dutton, B. Frey, M. Peluso, M. Nicoletti, N. Diershke, K. Hayashi, A. Klunder, D. Glahn, P. Brambilla, R. Sassi, A. Mallinger, and J. Soares, "Three-dimensional mapping of hippocampal anatomy in unmedicated and lithium-treated patients with bipolar disorder," *Neuropsychopharmacology*, in press: 2007.

[6] J. Becker *et al.*, "3D patterns of hippocampal atrophy in mild cognitive impairment," *Archives of Neurology*, vol. 63, no. 1, pp. 97–101, Jan. 2006.

[7] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *J.R. Statist. Soc. B*, vol. 57, no. 1, pp. 289–300, 1995.

[8] H. Blum, "A Transformation for Extracting New Descriptors of Shape," in *Models for the Perception of Speech and Visual Form*, W. Wathen-Dunn, Ed. Cambridge: MIT Press, 1967, pp. 362–380.

[9] P. Bradley and O. Mangasarian, "Feature selection via concave minimization and support vector machines," in *Proc. 15th International Conf. on Machine Learning*, 1998, pp. 82–90.

[10] C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery*, vol. 2, pp. 121–167, 1998.

[11] Y. Chou, N. Leporé, G. de Zubicaray, O. Carmichael, J. Becker, A. Toga, and P. Thompson, "Automated ventricular mapping with multi-atlas fluid image alignment reveals genetic effects in Alzheimer's disease," *NeuroImage*, July 2007 submitted.

[12] L. Clare, R. Woods, E. M. Cook, M. Orrell, and A. Spector, "Cognitive rehabilitation and cognitive training for early-stage Alzheimer's disease and vascular dementia," *Cochrane Database of Systematic Reviews*, vol. 4, 2003.

[13] D. Collins, P. Neelin, T. M. Peters, and A. C. Evans, "Automatic 3D intersubject registration of MR volumetric data in standardized Talairach space," *J Comput Assist Tomogr*, vol. 18, pp. 192–205, 1994.

[14] E. Edgington and P. Onghena, *Randomization tests*. New York: Marcel Dekker.

[15] Y. Fan, D. Shen, and C. Davatzikos, "Classification of structural images via high-dimensional image warping, robust feature extraction, and svm," in *MICCAI*, 2005.

[16] B. Fischl *et al.*, "Whole brain segmentation: Automated labeling of neuroanatomical structures in the human brain," *Neurotechnique*, vol. 33, pp. 341–355, 2002.

[17] Y. Freund and R. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of Computer Sys. Sci.*, vol. 55, pp. 119–139, 1997.

[18] G. Frisoni, F. Sabattoli, A. Lee, R. Dutton, A. Toga, and P. Thompson, "In vivo neuropathology of the hippocampal formation in AD: A radial mapping MR-based study," *Neuroimage*, vol. 32, no. 1, pp. 104–110, Aug. 2006.

[19] D. Galasko, M. Klauber, C. Hofstetter, D. Salmon, B. Lasker, and L. Thal, "The mini-mental state examination in the early diagnosis of Alzheimer's diesase," *Archives of Neurology*, vol. 47, no. 1, pp. 49–52, Jan. 1990.

[20] C. Genovese, N. Lazar, and T. Nichols, "Thresholding of statistical maps in functional neuroimaging using the false discovery rate," *NeuroImage*, vol. 15, no. 4, pp. 870–878, 2002.

[21] N. Gogtay, T. Nugent, D. Herman, A. Ordonez, D. Greenstein, K. Hayashi, L. Classen, A. Toga, J. Giedd, J. Rapoport, and P. Thompson, "Dynamic mapping of normal human hippocampal development," *Hippocampus*, 2006.

[22] P. Golland, W. Grimson, M. Shenton, and R. Kikinis, "Detection and analysis of statistical differences in anatomical shape," *Medical Image Analysis*, vol. 9, no. 1, pp. 69–86, 2005.

[23] B. Gutman, Y. Wang, L. Lui, T. Chang, and P. Thompson, "Hippocampal surface analysis using spherical harmonic functions applied to surface conformal mapping," in *International Conference on Pattern Recognition*, 2006, pp. 964–967.

[24] A. Hammers, R. Heckemann, M. Koepp, J. Duncan, J. Hajnal, D. Rueckert, and P. Aljabar, "Automatic detection and quantification of hippocampal atrophy on MRI in temporal lobe epilepsy: A proof of principle study," *Neuroimage*, 2007.

[25] R. Hogan, K. Mark, L. Wang, S. Joshi, M. Miller, and R. Bucholz, "Mesial temporal sclerosis and temporal lobe epilepsy: MR imaging deformation-based segmentation of the hippocampus in five patients," *Radiology*, vol. 216, pp. 291–297, 2000.

[26] C. Jack *et al.*, "The Alzheimer's Disease Neuroimaging Initiative (ADNI): The MR imaging protocol," *Journal of MRI*, vol. 27, no. 4, pp. 685–691, 2008.

[27] T. Joachims, "Training linear SVMs in linear time," in *Proceedings of the ACM Conference of Knowledge Discovery and Data Mining (KDD)*, 2006, pp. 217–226.

[28] S. LaConte, S. Strother, V. Cherkassky, J. Anderson, and X. Hu, "Support vector machines for temporal classification of block design fMRI data," *Neuroiamge*, vol. 26, no. 2, pp. 317–329, June 2005.

[29] D. Langers, J. Jansen, and W. Backes, "Enhanced signal detection in neuroimaging by means of regional control of the global false discovery rate," *NeuroImage*, Aug. accepted manuscript 2007.

[30] Z. Lao, D. Shen, A. Jawad, B. Karacali, D. Liu, E. Melhem, N. Bryan, and C. Davatzikos, "Automated segmentation of white matter lesions in 3D brain MR images, using multivariate pattern classification," in *Proc. of 3rd IEEE In'l Symp. on Biomedical Imaging (ISBI)*, Arlington, Apr. 2006, pp. 307–310.

[31] N. Leporé, C. Brun, Y. Chou, M. Chiang, R. Dutton, K. Hayashi, A. Lu, O. Lopez, H. Aizenstein, A. Toga, J. Becker, and P. Thompson, "Generalized tensor-based morphometry of HIV/AIDS using multivariate statistics on strain matrices and their application to HIV/AIDS," *IEEE Transactions on Medical Imaging*, in press 2007.

[32] J. Lin, N. Salamon, A. Lee, R. Dutton, J. Geaga, K. Hayashi, A. Toga, J. Engel, and P. Thompson, "3D pre-operative maps of hippocampal atophy predict surgical outcomes in temporal lobe epilepsy," *Neurology*, vol. 65, pp. 1094–1097, Oct. 2005.

[33] S. Loncaric, "A survery of shape analysis techniques – from automata to hardware," *Pattern Recognition*, vol. 31, no. 8, pp. 983–1001, Aug. 1998.

[34] N. Lord, J. Ho, B. Vemuri, and S. Eisenschenk, "Simultaneous registration and parcellation of bilateral hippocampal surface pairs for local asymmetry quantification," *IEEE Trans Med Imaging*, vol. 26, no. 4, pp. 471–178, Apr. 2007.

[35] J. Luts, A. Heerschap, J. Suykens, and S. V. Huffel, "A combined MRI and MRSI based multiclass system for brain tumor recognition using LS-SVMs with class probabilities and feature selection," *Artificial Intelligence in Medicine*, vol. 40, no. 2, pp. 87–102, June 2007.

[36] J. Mazziotta, A. Toga, A. Evans, P. Fox, J. Lancaster, K. Zilles, R. Woods, T. Paus, G. Simpson, B. Pike, C. Holmes, D. Collins, P. Thompson, D. MacDonald, T. Schormann, K. Amunts, N. Palomero-Gallagher, L. Parsons, K. Narr, N. Kabani, G. L. Goualher, D. Boomsma, T. Cannon, R. Kawashima, and B. Mazoyer, "A probabilistic atlas and reference system for the human brain [invited paper]," *Journal of the Royal Society*, vol. 356, no. 1412, pp. 1293–1322, Jan. 2001.

[37] M. Miller, "Computational anatomy: shape, growth and atrophy comparison via diffeomorphisms," *Neuroimage*, vol. 23, no. 1, pp. 19–33, 2004.

[38] J. Morra, Z. Tu, L. Apostolova, A. Green, C. Avedissian, S. Madsen, X. Hua, A. T. C. Jack, M. Weiner, and P. Thompson, "Validation of a fully automated 3d hippocampal segmentation method using subjects with alzheimer's disease mild cognitive impairment, and elderly controls," *NeuroImage*, vol. 43, no. 1, pp. 59–68, Oct. 2008.

[39] K. Müller, M. Sebastian, G. Rätsch, T. Koji, and B. Schölkopf, "An introduction to kernel-based learning," *IEEE Transactions on Neural Networks*, vol. 12, no. 2, pp. 181–201, Mar. 2001.

[40] K. Narr, T. van Erp, T. Cannon, R. Woods, P. Thompson, S. Jang, R. Blanton, V. Poutanen, M. Huttunen, J. Lonnqvist, C. Standerksjold-Nordenstam, J. Kaprio, J. Mazziotta, and A. Toga, "A twin study of genetic contributions to hippocampal morphology in schizophrenia," *Neurobiology of Disease*, vol. 11, no. 1, pp. 83–95, Oct. 2002.

[41] T. Nichols and A. Holmes, "Nonparametric permutation tests for functional neuroimaging: a primer with examples," *Human Brain Mapping*, vol. 15, no. 1, pp. 1–25, 2002.

[42] R. Nicolson, T. DeVito, C. Vidal, Y. Sui, K. Hayashi, D. Drost, P. Williamson, N. Rajakumar, A. Toga, and P. Thompson, "Detection and mapping of hippocampal abnormalities in autism," *Psychiatry Neuroimaging Research*, vol. 148, no. 1, pp. 11–21, Nov. 2006.

[43] T. Nugent, D. Herman, A. Ordonez, D. Greenstein, K. Hayashi, M. Lenane, L. Clasen, D. Jung, A. Toga, J. Giedd, J. Rapoport, P. Thompson, and N. Gogtay, "Dynamic mapping of hippocampal development in childhood onset schizophrenia," *Schizophrenia Research*, vol. 90, pp. 62–70, Feb. 2007.

[44] K. Pohl, J. Fisher, R. Kikinis, W. Grimson, and W. Wells, "A Bayesian model for joint segmentation and registration," *NeuroImage*, vol. 31, no. 1, pp. 228–239, 2006.

[45] K. Pohl, J. Fisher, M. Shenton, R. McCarley, W. Grimson, R. Kikinis, and W. Wells, "Logarithm odds maps for shape representation," in *MICCAI*, vol. 6, 2006, pp. 955–963.

[46] S. Powell, V. Magnotta, H. Johnson, V. Jammalamadaka, R. Pierson, and N. Andreasen, "Registration and machine learning based automated segmentation of subcortical and cerebellar brain structures," *NeuroImage*, vol. 39, no. 1, pp. 238–247, Jan. 2008.

[47] A. Quddus, P. Fieguth, and O. Basir, "Adaboost and support vector machines for white matter lesion segmentation in MR images," in *Proc. of 27th IEEE-EMBS*, 2005, pp. 463–466.

[48] D. Rex, J. Ma, and A. Toga, "The LONI pipeline processing environment," *Neuroimage*, vol. 19, no. 3, pp. 1033–1048, July 2003.

[49] F. Sabattoli, M. Boccardi, S. Galluzzi, A. Treves, P. Thompson, and G. Frisoni, "Hippocampal shape changes in dementia with Lewy bodies," Sept. 2007 submitted.

[50] R. Schapire, Y. Freund, P. Bartlett, and W. Lee, "Boosting the margin: A new explanation for the effectiveness of voting methods," *Annals of Statistics*, vol. 26, no. 5, pp. 1651–1686, 1998.

[51] D. Shattuck, S. Sandor-Leahy, K. Schaper, D. Rottenberg, and R. Leahy, "Magnetic resonance image tissue classification using a partial volume model," *Neuroimage*, vol. 13, pp. 856–876, 2001.

[52] D. Shen, S. Moffat, S. Resnick, and C. Davatzikos, "Measuring size and shape of the hippocampus in MR images using a deformable shape model," *Neuroimage*, vol. 15, no. 2, pp. 422–434, Feb. 2002.

[53] Y. Shi, A. Bobick, and I. Essa, "A Bayesian view of boosting and its extensions," ser. GVU Technical Report; GIT-GVU-05-22. Georgia Institute of Technology, 2005.

[54] Y. Shi, P. Thompson, G. de Zubicaray, S. Rose, Z. Tu, I. Dinov, and A. Toga, "Direct mapping of hippocampal surfaces with intrinsic shape contex," *Neuroimage*, in press 2007.

[55] J. Storey, "A direct approach to false discovery rates," *J.R. Statist. Soc. B*, vol. 64, no. 3, pp. 479–498, 2002.

[56] M. Styner, J. Lieberman, and G. Gerig, "Boundary and medial shape analysis of the hippocampus in schizophrenia," in *MICCAI*, Feb. 2003, pp. 464–471.

[57] T. Terriberry, J. Damon, S. Pizer, S. Joshi, and G. Gerig, "Population-based fitting of medial shape models with correspondence optimization," *Inf Process Med Imaging*, vol. 20, pp. 700–712, 2007.

[58] P. Thompson, K. Hayashi, G. de Zubicaray, A. Janke, S. Rose, J. Semple, D. Herman, M. Hong, S. Dittmer, D. Doddrell, and A. Toga, "Dynamics of gray matter loss in Alzheimer's disease," *Journal of Neuroscience*, vol. 23, no. 3, pp. 994–1005, Feb. 2003.

[59] P. Thompson, K. Hayashi, G. de Zubicaray, A. Janke, S. Rose, J. Semple, M. Hong, D. Herman, D. Gravano, D. Doddrell, and A. Toga, "Mapping hippocampal and ventricular change in Alzheimer's disease," *NeuroImage*, vol. 22, no. 4, pp. 1754–1766, Aug. 2004.

[60] P. Thompson, K. Hayashi, S. Simon, J. Geaga, M. Hong, Y. Sui, J. Lee, A. Toga, W. Ling, and E. London, "Structural abnormalities in the brains of human subjects who use methamphetamine," *Journal of Neuroscience*, vol. 24, no. 26, pp. 6028–6036, June 2004.

[61] P. Thompson, C. Schwartz, R. Lin, A. Khan, and A. Toga, "3D statistical analysis of sulcal variability in the human brain," *Journal of Neuroscience*, vol. 16, no. 13, pp. 4261–4274, July 1996.

[62] P. Thompson, C. Schwartz, and A. Toga, "High-resolution random mesh algorithms for creating a probabilistic 3D surface atlas of the human brain," *NeuroImage*, vol. 3, no. 1, pp. 19–34, Mar. 1996.

[63] D. Tschumperlé. (2007) Cimg library. [Online]. Available: http://cimg.sourceforge.net/index.shtml

[64] Z. Tu, "Probabilistic boosting tree: Learning discriminative models for classification, recognition, and clustering," in *Proceedings of ICCV*, 2005.

[65] Z. Tu *et al.*, "Brain anatomical structure parsing by hybrid discriminative/generative models," *IEEE TMI*, 2008.

[66] M. Valliant and J. Glaunès, "Surface matching via currents," *Inf Process Med Imaging*, vol. 19, pp. 381–392, 2005.

[67] V. Vapnik, *Statistical Learning Theory*. New York: Wiley-Interscience, 1998.

[68] L. Wang, F. Beg, T. Ratnanather, C. Ceritoglu, L. Younse, J. Morris, J. Csernansky, and M. Miller, "Large deformation diffeomorphism and momentum based hippocampal shape discrimination in dementia of the Alzheimer type," *IEEE TMI*, vol. 26, no. 4, pp. 462–470, 2007.

[69] Y. Wang, M. Chiang, and P. Thompson, "Mutual information-based 3D surface matching with applications to face recognition and brain mapping," in *International Conference on Computer Vision 2005*, Beijing, China, Oct. 2005, pp. 527–534.

[70] P. Yushkevich, S. Joshi, S. Pizer, J. Csernansky, and L. Wang, "Feature selection for shape-based classification of biological objects," in *IPMI*, Aug. 2003, pp. 114–125.

[71] P. A. Yushkevich, J. Piven, C. Hazlett, H. Smith, G. Smith, R. Ho, S. Ho, J. Gee, and G. Gerig, "User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability," *Neuroimage*, vol. 31, pp. 1116–1128, 2006.

[72] M. Zeineh, S. Engel, P. Thompson, and S. Bookheimer, "Dynamics of the hippocampus during encoding and retrieval of face-name pairs," *Science*, vol. 299, no. 5606, pp. 577–580, Jan. 2003.