

Sparse Bayesian Learning for Identifying Imaging Biomarkers in AD Prediction

Li Shen^{1,2,*}, Yuan Qi^{3,*}, Sungeun Kim^{1,2}, Kwangsik Nho^{1,2}, Jing Wan^{1,2}, Shannon L. Risacher¹, Andrew J. Saykin^{1,*}, and ADNI

¹Center for Neuroimaging, Department of Radiology and Imaging Sciences,

²Center for Computational Biology and Bioinformatics,

Indiana University School of Medicine, 950 W Walnut St, Indianapolis, IN 46202

³Departments of Computer Science, Statistics and Biology,

Purdue University, 305 N. University Street, West Lafayette, IN 47907

Abstract. We apply sparse Bayesian learning methods, automatic relevance determination (ARD) and predictive ARD (PARD), to Alzheimer’s disease (AD) classification to make accurate prediction and identify critical imaging markers relevant to AD at the same time. ARD is one of the most successful Bayesian feature selection methods. PARD is a powerful Bayesian feature selection method, and provides sparse models that is easy to interpret. PARD selects the model with the best estimate of the predictive performance instead of choosing the one with the largest marginal model likelihood. Comparative study with support vector machine (SVM) shows that ARD/PARD in general outperform SVM in terms of prediction accuracy. Additional comparison with surface-based general linear model (GLM) analysis shows that regions with strongest signals are identified by both GLM and ARD/PARD. While GLM P-map returns significant regions all over the cortex, ARD/PARD provide a small number of relevant and meaningful imaging markers with predictive power, including both cortical and subcortical measures.

1 Introduction

Neuroimaging is a powerful tool for characterizing neurodegenerative process in the progression of Alzheimer’s disease (AD) and can provide potential surrogate biomarkers for therapeutic trials. This paper is focused on identifying relevant imaging biomarkers from structural magnetic resonance imaging (MRI) data for AD classification. Machine learning methods have been applied to many problems in computational neuroscience, including computer-aided diagnosis for

* Correspondence to Li Shen (shenli@iupui.edu), Yuan Qi (alanqi@cs.purdue.edu), or Andrew J. Saykin (asaykin@iupui.edu). Data collection and sharing for this project was funded by the Alzheimer’s Disease Neuroimaging Initiative (ADNI) (U01 AG024904). This project was also supported in part by Indiana CTSI IUSM/CTR(RR025761), 1RC 2AG036535, NIA R01 AG19771, Foundation for the NIH, IEDC #87884, NIBIB R03 EB008674, U01 AG032984, and P30 AG10133. The original publication is available at www.springerlink.com.

Table 1. Participant characteristics

Category	HC	AD	<i>p</i> -value
Number of Subjects	203	175	-
Gender (M/F)	111/92	97/78	0.8840
Baseline Age (years; Mean±STD)	76.09±5.00	75.53±7.58	0.3884
Education (years; Mean±STD)	16.13±2.73	14.93±3.00	< 0.0001
Handedness (R/L)	188/15	163/12	0.8413

AD [1, 3, 4, 6, 7, 9]. While popular methods like support vector machines (SVMs) [15] can achieve decent prediction accuracy, most of them are not optimized for selecting sensitive features.

This paper presented the results of applying novel sparse Bayesian learning methods, automatic relevance determination (ARD) and predictive ARD (PARD) [13], to MRI-based AD classification for achieving two goals at the same time: (1) accurate prediction rate and (2) selection of relevant imaging biomarkers. Linear SVM and general linear model (GLM) based cortical thickness analyses were also performed on the same data for comparison to ARD/PARD. Our overarching goal is to learn from these data sparse Bayesian models so that they are easy to interpret while maintaining high predictive power.

2 Materials and Methods

MRI Data used in this study were obtained from the Alzheimer’s Disease Neuroimaging Initiative (ADNI) database (www.loni.ucla.edu/ADNI). ADNI is a landmark investigation sponsored by the NIH and industrial partners designed to collect longitudinal neuroimaging, biological and clinical information from 800 participants that will track the neural correlates of memory loss from an early stage. Further information can be found in [11] and at www.adni-info.org. Following a previous imaging genetics study [14], 378 non-Hispanic Caucasian participants (203 healthy control (HC) and 175 AD participants) were selected for this work. For one baseline scan of each participant, FreeSurfer V4 was employed to automatically label cortical and subcortical tissue classes [2, 5] and to extract target region volume and cortical thickness, as well as to extract total intracranial volume (ICV), as previously described [14]. For each hemisphere, thickness measures of 34 cortical regions of interest (ROIs) (Fig. 1(a-f)) and volume measures of 15 cortical and subcortical ROIs (Fig. 1(c-f)) were included in this study. All these measures were adjusted for the baseline age, gender, education, handedness, and baseline ICV using the regression weights derived from the HC participants. Participant characteristics are summarized in Table 1.

ARD and Predictive ARD

We apply ARD and predictive ARD (PARD) [13] to classify the imaging features. ARD is one of the most successful Bayesian feature selection methods [8, 12]. It is a hierarchical Bayesian approach where there are hyperparameters which explicitly represent the relevance of different input features. These

relevance hyperparameters determine the range of variation for the parameters relating to a particular input, usually by modeling the width of a zero-mean Gaussian prior on those parameters. If the width of that Gaussian is zero, then those parameters are constrained to be zero, and the corresponding input cannot have any effect on the predictions, therefore making it irrelevant. ARD optimizes these hyperparameters to discover which inputs are relevant.

Predictive ARD improves upon ARD in the following aspects. First, the Laplace approximation used in ARD [8] is replaced by the more accurate expectation propagation (EP) [10]. Second, EP computes an estimate of leave-one-out predictive performance without requiring expensive cross-validation experiments. This estimate of predictive performance can be used as an important criterion for ARD to avoid the overfitting problem associated with evidence maximization. Last, predictive ARD uses a fast *sequential* optimization method such that we can efficiently prune and add new features without updating a full covariance matrix for the classifier.

Now we describe ARD for linear classification. A linear classifier classifies a point \mathbf{x} according to $t = \text{sign}(\mathbf{w}^T \mathbf{x})$ for some parameter vector \mathbf{w} (the two classes are $t = \pm 1$). Given a training set $D = \{(\mathbf{x}_1, t_1), \dots, (\mathbf{x}_N, t_N)\}$, the likelihood for \mathbf{w} can be written as

$$p(\mathbf{t}|\mathbf{w}, X) = \prod_i p(t_i|\mathbf{x}_i, \mathbf{w}) = \prod_i \Psi(t_i \mathbf{w}^T \phi(\mathbf{x}_i)) \quad (1)$$

where $\mathbf{t} = \{t_i\}_{i=1}^N$, $X = \{\mathbf{x}_i\}_{i=1}^N$, $\Psi(\cdot)$ is the cumulative distribution function for a Gaussian. One can also use the step function or logistic function as $\Psi(\cdot)$. The basis function $\phi^T(\mathbf{x}_i)$ allows the classification boundary to be nonlinear in the original features. This is the same likelihood used in logistic regression and in Gaussian process classifiers. Given a new input \mathbf{x}_{N+1} , we approximate the predictive distribution:

$$p(t_{N+1}|\mathbf{x}_{N+1}, \mathbf{t}) = \int p(t_{N+1}|\mathbf{x}_{N+1}, \mathbf{w})p(\mathbf{w}|\mathbf{t})d\mathbf{w} \quad (2)$$

$$\approx P(t_{N+1}|\mathbf{x}_{N+1}, \langle \mathbf{w} \rangle) \quad (3)$$

where $\langle \mathbf{w} \rangle$ denotes the posterior mean of the weights, called the Bayes Point.

The basic idea in ARD is to give the feature weights independent Gaussian priors:

$$p(\mathbf{w}|\boldsymbol{\alpha}) = \prod_i \mathcal{N}(w_i|0, \alpha_i^{-1}),$$

where $\boldsymbol{\alpha} = \{\alpha_i\}$ is a hyperparameter vector that controls how far away from zero each weight is allowed to go. The hyperparameters $\boldsymbol{\alpha}$ are trained from the data by maximizing the Bayesian ‘evidence’ $p(\mathbf{t}|\boldsymbol{\alpha})$, which can be done using a fixed point algorithm or an expectation maximization (EM) algorithm treating \mathbf{w} as a hidden variable [8]. The outcome of this optimization is that many elements of $\boldsymbol{\alpha}$ go to infinity such that the classifier \mathbf{w} would have only a few nonzero weights w_j . This naturally prunes irrelevant features in the data.

Unlike previous approaches that use the EM algorithm and find a solution that maximizes the evidence, the **predictive-ARD** (PARD) algorithm trains

Table 2. Performance comparison. Training and testing error rates (mean±std) of 10-fold cross validation are shown for SVM, PARD (Predictive ARD) and ARD.

	SVM		PARD		ARD	
	training	testing	training	testing	training	testing
Left [†]	0.129 ± 0.011	0.156 ± 0.062	0.106 ± 0.009	0.147 ± 0.082	0.108 ± 0.010	0.154 ± 0.065
Right [†]	0.149 ± 0.009	0.185 ± 0.076	0.137 ± 0.009	0.168 ± 0.086	0.139 ± 0.006	0.175 ± 0.077
Left [‡]	0.112 ± 0.006	0.139 ± 0.051	0.078 ± 0.008	0.132 ± 0.056	0.078 ± 0.008	0.124 ± 0.056
Right [‡]	0.130 ± 0.006	0.142 ± 0.053	0.117 ± 0.005	0.160 ± 0.040	0.118 ± 0.006	0.162 ± 0.039

[†] Use cortical thickness measures only. [‡] Use both thickness and volume measures.

the sparse classifier as follows: (1) Initialize the model so that it only contains a small fraction of features. (2) Sequentially update the classifiers via a fast sequential optimization method and calculate the required statistics by EP until the algorithm converges. The sparsity level of the classifiers increases along the optimization iterations. (3) From all the classifiers, choose the classifier with minimum predictive error probability estimate.

SVM and GLM

A linear support vector machine (SVM) was applied in our study to provide a comparison to ARD/PARD in terms of prediction accuracy. SVMs represent a new generation of learning systems based on recent advances in statistical learning theory [15]. The aim in training a linear SVM is to find the separating hyperplane with the largest margin; the expectation is that the larger the margin, the better the generalization of the classifier. We employed the OSU SVM Matlab Toolbox (sourceforge.net/projects/svm/) in this work.

We also performed surface based analysis for identifying thickness changes on the brain cortex and comparing these regions with the imaging markers detected by ARD/PARD. We consider the following general linear model (GLM): $y = X\Psi + Z\Phi + \epsilon$, where the dependent variable y is cortical thickness; $X = (x_1, \dots, x_p)$ are the variables of interest (*diagnosis* in our case); $Z = (z_1, \dots, z_k)$ are the variables whose effects we want to exclude (*age, gender, education, handedness* and *ICV* in our case); $\Psi = (\psi_1, \dots, \psi_p)^T$ and $\Phi = (\phi_1, \dots, \phi_k)^T$ are the coefficients; and ϵ is the error term. The goal is to test if X is significant (i.e., $\Psi \neq 0$) for some $y \in \partial\Omega$, where $\partial\Omega$ is the cortical surface manifold. To test GLMs, we used SurfStat [16], a Matlab toolbox for the statistical analysis of univariate and multivariate surface and volumetric data using linear mixed effects models and random field theory (RFT) [17].

3 Results

Classification was performed on each hemisphere separately, using two sets of imaging features: (1) 34 thickness measures (Fig. 1(a-b)), and (2) 34 thickness measures and 15 volume measures (Fig. 1(c-f)). 10-fold cross-validation was performed for accuracy estimation. Shown in Table 2 is the performance comparison among ARD, PARD and SVM. ARD and PARD outperformed SVM except for the case of using both thickness and volume measures from right hemisphere.

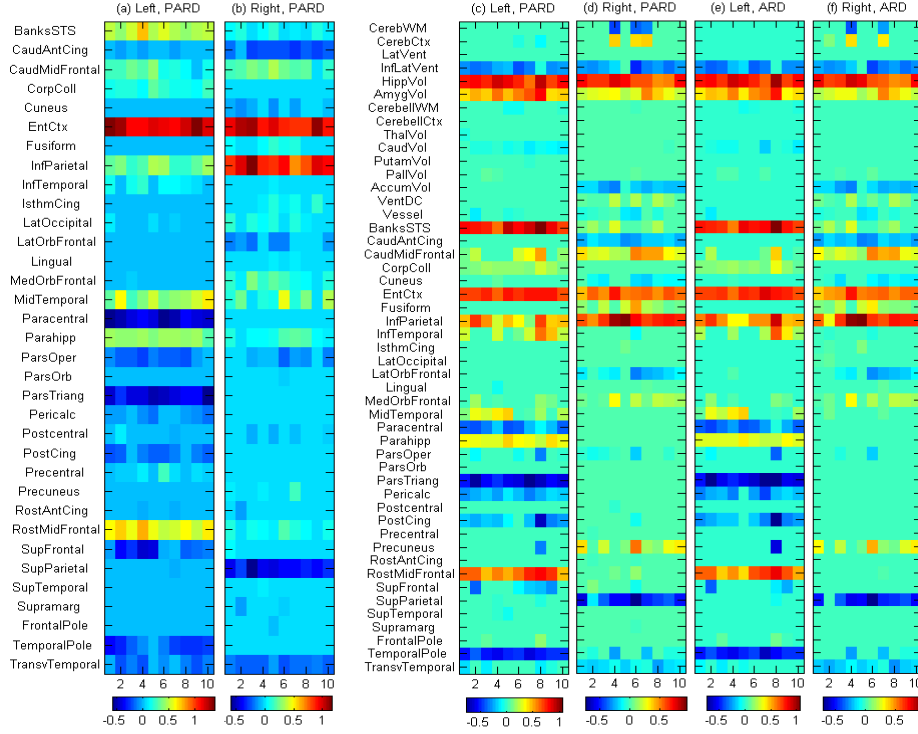


Fig. 1. (a-b) Heat maps of PARD weights $-w$ in cortical analyses using 34 thickness measures. (c-f) Heat maps of PARD (c,d) and ARD (e,f) weights $-w$ in analyses using 15 volume (top) and 34 thickness (bottom) measures. 10-fold cross-validation was performed for left (a,c,e) and right (b,d,f) hemisphere data. In each heat map, feature weights were plotted against 10 different trials in cross validation tests.

PARD outperformed ARD except for the case of using both thickness and volume measures from left hemisphere. PARD was designed for improving ARD predictive performance based on theoretical considerations, which empirically worked better for most cases but not all. Using thickness measures only, the best prediction rate was obtained at 85.3% by PARD for left hemisphere. Using both thickness and volume measures, the best prediction rate was improved to 87.6% by applying ARD to the left hemisphere data. In all cases, the prediction rates were improved after including 15 additional volume measures in the analyses, indicating both cortical and subcortical changes were related to AD.

A linear classifier is usually characterized by a weight vector w , which projects each individual data point (i.e., a feature vector) into a 1-D space for getting a discriminative value. Each weight measures the amount of the contribution of the corresponding feature to the final discriminative value. ARD and PARD aim to reduce the number of nonzero weights so that only relevant features

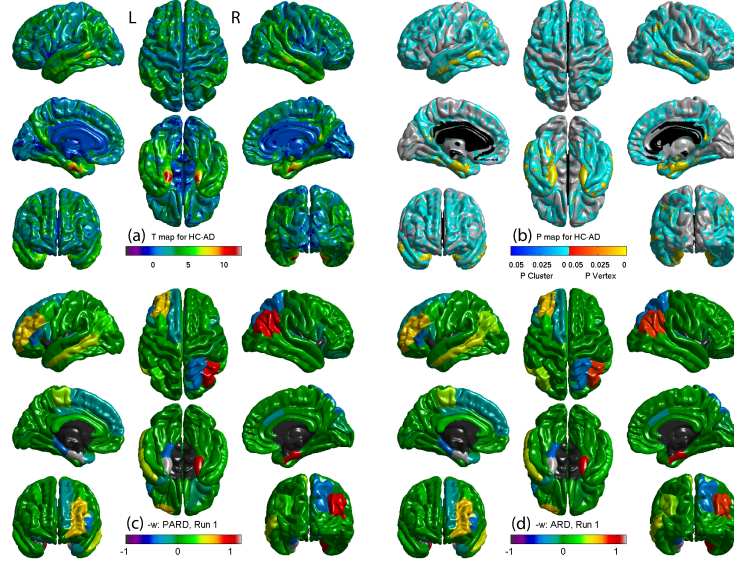


Fig. 2. (a-b) GLM results of diagnosis effect (HC-AD) on cortical thickness include (a) the map of the t statistics and (b) the map of corrected P values for peaks and clusters (only regions with corrected $p \leq 0.01$ are shown), where positive t values (red, yellow) indicate more grey matter in HC. (c-d) Back-projection of negative weights ($-w$) of the linear classifier for (c) PARD and (d) ARD, where positive values (gray, red, yellow) indicate more grey matter in HC.

are selected by examining these weights. For consistency, we always visualize negative weights $-w$ so that larger values (red) correspond to more grey matter in HC. Fig. 1(a-b) shows the heat maps of PARD weights $-w$ in cortical thickness analysis for one run of 10-fold cross validation for both hemispheres. The weight vectors (i.e., columns in the map) derived by different trials in cross validation are very similar. Most weights are close to zero, indicating a small number of relevant imaging markers. While entorhinal cortex (EntCtx) appears to be a strong predictor in both sides, rostral middle frontal gyri (RostMidFrontal) are strong only on the left and inferior temporal gyri (InfParietal) on the right.

These weights can be back-projected to the original image space for an intuitive visualization. Fig. 2(c-d) shows such a visualization for PARD and ARD results using thickness data. Since we only examine the mean thickness of each cortical subregion in our analysis, the entire region is painted with the same color defined by the corresponding weight. The patterns of imaging marker selection between PARD and ARD are very similar to each other. For comparison, surface-based GLM analysis using SurfStat is also performed to examine diagnosis effect (HC-AD) on cortical thickness and Fig. 2(a-b) shows the resulting T-map and P-map. Regions with strongest signals, such as entorhinal cortex on both sides

Table 3. Top imaging markers: “mean weight, rank” shown in each cell.

ID	Description	Thick. & Vol., ARD		Thickness, PARD	
		Left	Right	Left	Right
HippVol	hippocampus	-0.804, 1	-0.641, 2	N/A	N/A
BanksSTS	banks of the superior temporal sulcus	-0.790, 2	-0.077, 10	-0.503, 3	-0.043, 9
EntCtx	entorhinal cortex	-0.745, 3	-0.523, 3	-1.199, 1	-0.954, 1
RostMidFrontal	rostral middle frontal gyri	-0.621, 4	-0.001, 17	-0.607, 2	-0.063, 7
InfParietal	inferior parietal gyri	-0.569, 5	-0.656, 1	-0.332, 6	-0.901, 2
AmygVol	amygdala	-0.514, 6	-0.271, 5	N/A	N/A
Parahipp	parahippocampal gyri	-0.316, 7	0.000, 19	-0.405, 5	-0.086, 6
InfTemporal	inferior temporal gyri	-0.189, 8	0.009, 38	-0.097, 9	-0.001, 15
MidTemporal	middle temporal gyri	-0.179, 9	-0.011, 13	-0.446, 4	-0.167, 4
CorpColl	corpus collosum	-0.168, 10	0.000, 20	-0.195, 8	0.000, 17

and left middle temporal gyri are picked up by GLM and ARD/PARD. While GLM P-map returns significant regions across the entire cortex, PARD/ARD maps provide a small number of selective regions with predictive power.

Heat maps of ARD/PARD weights $-w$ in combined thickness and volume analyses are shown in Fig. 1(c-f). Again, the patterns are very similar between ARD and PARD. Shown in Table 3 are top imaging markers selected by ARD using thickness and volume measures (PARD data not shown but extremely similar to ARD) and by PARD using thickness measures (ARD data not shown but extremely similar to PARD). While most top markers are thickness measures from cortical regions, two markers are volume measures from subcortical structures including hippocampus and amygdala.

4 Discussion

We presented a novel application of sparse Bayesian learning methods, ARD and PARD, to AD classification. Our strategy was to minimize the complexity of both data and methods for deriving a simple model easy to interpret. For methods, we focused on linear classifiers and showed that ARD/PARD in general outperformed SVM. For data, we focused on summary statistics (i.e., thickness and volume) of anatomically meaningful grey matter regions across the whole brain, and showed that promising prediction accuracy (87.6%) could be achieved with a small number of relevant imaging measures. Most prior studies (e.g., [1, 3, 4]) performed feature selection/extraction before classification. Our method integrated feature selection into the learning process to form a simple and principled procedure. Prior research [6] also integrated feature selection into classification and reported lower prediction rates (77-82%) for analyzing a subset of the same ADNI MRI data. Comparison to other feature selection schemes merits further investigation. While some prior studies [3, 4, 7, 9] reported better prediction rates, they analyzed many more imaging variables in much smaller data sets. One interesting future topic is to apply our method to more detailed imaging

features to determine if better prediction rates and refined imaging marker maps can be achieved. It is unclear if disease duration of AD is comparable between ADNI cohort examined by us and [1, 4, 6] and others cohorts by [3, 7, 9], and this could have an effect on prediction rates. Incorporating disease duration in predictive models warrants further investigation. To sum up, contributions of this work include: (1) a simple and unified learning method that inherently does feature selection and enables biomarker discovery while maintaining high predictive power; (2) a much larger AD sample tested with much fewer variables, resulting in a better power; and (3) promising rates for predicting mild AD with identified biomarkers that are known to be related to AD.

References

1. Batmanghelich, N., Taskar, B., Davatzikos, C.: A general and unifying framework for feature construction, in image-based pattern classification. *Inf Process Med Imaging* 21, 423–34 (2009)
2. Dale, A., Fischl, B., Sereno, M.: Cortical surface-based analysis. I. Segmentation and surface reconstruction. *Neuroimage* 9(2), 179–94 (1999)
3. Duchesne, S., Caroli, A., Geroldi, C., Barillot, C., Frisoni, G.B., Collins, D.L.: MRI-based automated computer classification of probable ad versus normal controls. *IEEE Trans Med Imaging* 27(4), 509–20 (2008)
4. Fan, Y., Batmanghelich, N., Clark, C.M., Davatzikos, C.: Spatial patterns of brain atrophy in MCI patients, identified via high-dimensional pattern classification, predict subsequent cognitive decline. *Neuroimage* 39(4), 1731–43 (2008)
5. Fischl, B., Sereno, M., Dale, A.: Cortical surface-based analysis. II: Inflation, flattening, and a surface-based coordinate system. *Neuroimage* 9(2), 195–207 (1999)
6. Hinrichs, C., Singh, V., et al.: Spatially augmented LPboosting for AD classification with evaluations on the ADNI dataset. *Neuroimage* 48(1), 138–49 (2009)
7. Kloppel, S., Stonnington, C.M., et al.: Automatic classification of MR scans in Alzheimer’s disease. *Brain* 131(Pt 3), 681–9 (2008)
8. MacKay, D.J.: Bayesian interpolation. *Neural Computation* 4(3), 415–447 (1992)
9. Magnin, B., Mesrob, L., Kinkingnehun, S., et al.: Support vector machine-based classification of Alzheimer’s disease from whole-brain anatomical MRI. *Neuroradiology* 51(2), 73–83 (2009)
10. Minka, T.P.: Expectation propagation for approximate Bayesian inference. In: 17th Conf. in Uncertainty in Artificial Intelligence. pp. 362–369 (2001)
11. Mueller, S.G., Weiner, M.W., et al.: The Alzheimer’s disease neuroimaging initiative. *Neuroimaging Clin N Am* 15(4), 869–77, xi–xii (2005)
12. Neal, R.M.: Bayesian Learning for Neural Networks. No. 118 in *Lecture Notes in Statistics*, Springer, New York (1996)
13. Qi, Y., Minka, T., et al.: Predictive automatic relevance determination by expectation propagation. In: 21st Int. Conf. on Machine learning. pp. 671–678 (2004)
14. Shen, L., Kim, S., et al.: Whole genome association study of brain-wide imaging phenotypes for identifying quantitative trait loci in MCI and AD: A study of the ADNI cohort. *Neuroimage*, doi:10.1016/j.neuroimage.2010.01.042 (2010)
15. Vapnik, V.: *Statistical Learning Theory*. John Wiley and Sons (1998)
16. Worsley, K.J.: SurfStat: <http://www.math.mcgill.ca/keith/surfstat>
17. Worsley, K.J., Andermann, M., Koulis, M., et al.: Detecting changes in non-isotropic images. *Human Brain Mapping* 8, 98–101 (1999)