Ensemble Learning Incorporating Uncertain Registration

Ivor J. A. Simpson*, Mark W. Woolrich, Jesper L. R. Andersson, Adrian R. Groves, Julia A. Schnabel and the Alzheimer's Disease Neuroimaging Initiative

***Abstract*—This paper proposes a novel approach for improving the accuracy of statistical prediction methods in spatially normalised analysis. This is achieved by incorporating registration uncertainty into an ensemble learning scheme. A probabilistic registration method is used to estimate a distribution of probable mappings between subject and atlas space. This allows the estimation of the distribution of spatially normalised feature data, e.g. grey matter probability maps. From this distribution, samples are drawn for use as training examples. This allows the creation of multiple predictors, which are subsequently combined using an ensemble learning approach. Furthermore, extra testing samples can be generated to measure the uncertainty of prediction. This is applied to separating subjects with Alzheimer's disease from normal controls using a linear support vector machine on a region of interest in magnetic resonance images of the brain. We show that our proposed method leads to an improvement in discrimination using voxel based morphometry and deformation tensor based morphometry over bootstrap aggregating, a common ensemble learning framework. The proposed approach also generates more reasonable soft-classification predictions than bootstrap aggregating. We expect that this approach could be applied to other statistical prediction tasks where registration is important.**

***Index Terms*—Registration uncertainty, Ensemble learning, Alzheimer's disease**

I. Introduction

Medical imaging data is often used to make quantitative predictions about the current or future disease state of a subject. In the case of structural magnetic resonance (MR) images of the brain, image data can be analysed to identify differences in morphology which are distinctive between pathological and healthy states. The identification of these differences allows for a quantitative diagnostic measure, which have the potential to assist in early diagnosis of pathology. Several automated approaches have been demonstrated which allow the diagnosis and prediction of Alzheimer's disease (AD) [1].

Machine learning techniques such as statistical classifiers and regressors are often used to facilitate this objective. These approach predict the value of an outcome variable, such as disease score, or group, on the basis of some feature data, e.g. grey matter probability maps.

When analysing MR images of the brain for distinctive morphometric features across a group of subjects, most standard machine learning approaches require the image data to be transformed into a common frame of reference to facilitate comparison. This requires the use of an image registration tool to estimate the mathematical mapping between each subject image and the atlas space (usually an average, or representative subject), a process which is known as spatial normalisation. From the estimated mapping, spatially normalised feature data can be derived, which are used as a basis for making predictions.

Following spatial normalisation, statistical prediction may be performed using grey matter, or other tissue class probabilities as feature data. This is referred to as voxel based morphometry (VBM) [2]. Grey matter (GM) probability maps have previously been used in the discrimination of Alzheimer's disease [3], while other methods have incorporated white matter (WM) and cerebrospinal fluid (CSF) as well [4].

Alternatively, features of interest can be derived from the estimated mapping between subject and atlas space, a process which is known as deformation tensor based morphometry (TBM) [2]. Differences have been found in TBM data between subjects with Alzheimer's disease and age matched healthy controls [5].

The majority of machine learning techniques that are widely used for making predictions from medical imaging data are *supervised*, meaning that they require a set of training data with known outcome variables. This training set is used to derive a predictive model for estimating the mapping between feature data and outcome.

For any supervised learning approach any predictions on test data are highly dependent on the set of training data. As the feature data is required to be spatially normalised prior to analysis, each item of training data is dependent on

*I. J. A. Simpson is with the Institute of Biomedical Engineering, Department of Engineering Science, University of Oxford, OX3 7DQ, UK and the Oxford Centre for Functional MRI of the Brain, University of Oxford, OX3 9DU, UK. (email: ivor.simpson@eng.ox.ac.uk)

M. W. Woolrich is with the Oxford Centre for Human Brain Activity, Department of Psychiatry, University of Oxford and the Oxford Centre for Functional MRI of the Brain, University of Oxford, OX3 9DU, UK.

J. L. R. Andersson and A. R. Groves are with the Oxford Centre for Functional MRI of the Brain, University of Oxford, OX3 9DU, UK.

J. A. Schnabel is with the Institute of Biomedical Engineering, Department of Engineering Science, University of Oxford, OX3 7DQ, UK

 Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.ucla.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at http://www.loni.ucla.edu/ADNI/Collaboration/ADNI_Authorship_list.pdf.

the inferred image registration between the subject and atlas space. As has been previously shown in many studies, and is highlighted well in the comparison study in [6], inter-subject brain registration is far from an exact process. Therefore, we expect there to be some residual registration error in the feature data. Consequently, it is unlikely that any estimated statistical relationship derived from a given set of training data will be exactly correct and any residual mis-registration of data may contribute to errors in prediction.

The majority of registration methods estimate only the *maximum-a-posteriori* (MAP), which is the most likely mapping subject to some regularisation constraints. However, recent registration methods have emerged that provide estimates of the registration uncertainty [7][8]. This facilitates the consideration of a distribution of probable registration mappings, as opposed to just the MAP.

There are two published 3D medical image registration methods that we are aware of which infer a distribution of probable mappings: Risholm et al. [7] use Markov chain Monte Carlo (MCMC) to numerically estimate the full posterior distribution of transformation parameters, whilst marginalising over the regularisation parameters. This method allows the estimation of a non-parametric posterior distribution of mappings. However, the computational expense of this approach makes it impractical in the context of this work. An alternative approach, previously proposed by the authors [8], uses variational Bayes (VB) to infer an approximate posterior distribution of the set of transformation, regularisation, and noise parameters. This approach assumes that the true distribution of transformation parameters follows a multi-variate normal distribution making it computationally much more efficient.

Allassonnière et al. [9] describe a Bayesian deformable template registration framework which yields an approximate posterior distribution of transformation parameters. This approximative posterior may be a mixture of multi-variate normals, and is inferred using expectation-maximisation. Their work has the limitation that the strength of the prior, which greatly affects the posterior distribution, has to be hand-defined. This algorithm was demonstrated on 2D digit recognition.

An alternative view of registration uncertainty has been proposed by Van Leemput [10]. They describe an approach for creating a deformable labelled anatomical atlas through the use of a Bayesian statistical model. Their approach does not attempt to estimate the PDF of the true deformation field, but rather modelling the distribution of an image labelling. Such a method may provide complementary uncertainty information.

Some previous work has been performed on visualising transformation uncertainty in non-rigid registration [11][12]. More related work utilising the concept of uncertain registration includes: estimating local anisotropic smoothing kernels to compensate for uncertainty in registration when estimating spatially normalised statistics [13]. Very recently, Iglesias et al. [14] introduced an approach to hippocampal subfield segmentation, where registration uncertainty is integrated out of a combined registration/Gaussian mixture model approach to segmentation using MCMC. Risholm et al. [15] proposed an approach to calculating the uncertainty of a delivered dose in radiotherapy under uncertain registration. Here, we follow in a similar fashion by estimating the variability in statistical predictors that is due to the uncertainty in registration. The novel contribution of this work is to leverage this variability within a statistical learning framework to provide a more robust prediction.

We propose to incorporate the estimated registration uncertainty within an ensemble learning approach [16]. Ensemble learning methods have been demonstrated to be an effective mechanism to measure the uncertainty of the space of statistical predictors, providing a more robust prediction by combining estimates. To provide variability in predictive models, they need to be trained using different subsets of the training data. Each subset needs to be selected appropriately to encapsulate an appropriate level of variability in predictive models. In many settings, bootstrap aggregating or *bagging*, has proved itself to be an effective tool [17]. Bagging creates variability between predictors by sampling with replacement from the set of training subjects. However, the random selection of subjects in each training subset can lead to large differences in predictors due to inter-subject differences. Conversely, in this work we seek to leverage our knowledge of the derivation of the data to create sets of training data which encapsulate the intra-subject variability due to registration uncertainty.

In the case of spatially normalised feature data, the distribution of the data can be estimated from the set of probable mappings inferred from the registration algorithm. Samples can be drawn from this estimated distribution of feature data for each training subject, and used as a parametric variant of bootstrapping [18]. These samples of feature data can be used in place of the MAP observations to build up a set of training data sets, which all contain the same subjects, but with examples based on different probable registrations. Such an ensemble of statistical predictors accounts for the inherent uncertainty in the registration process, and therefore leads to a more robust prediction. A limited evaluation of this approach has been previously presented [19].

In this paper we first describe how the distribution of feature data can be derived using a probabilistic registration tool. Subsequently, we explain how this distribution can be used in an ensemble learning scheme. To demonstrate the performance of our approach, we apply it to discriminating between subjects with Alzheimer's disease and age matched healthy controls using a standard black box classifier, detailed in Section III. We find that the use of this scheme leads to an improvement in classification accuracy.

II. Methods

A. Probabilistic Registration Method

A probabilistic registration method that can estimate posterior transformation distributions is required to estimate the distribution of the feature data, which accounts for the uncertainty in registration. Standard registration procedures use a MAP approach to infer the mapping between images. These approaches do not provide any estimates on the confidence of the inferred mapping, and consequently do not lend themselves to this work. Therefore, we use our previously published

registration algorithm in this work [8], for which we now provide a summary.

Image registration can be described using a generative model:

$$\mathbf{y} = \mathbf{t}(\mathbf{x}, \mathbf{w}) + \mathbf{e} \tag{1}$$

where $\mathbf{y}$ is the target image, $\mathbf{t}(\mathbf{x}, \mathbf{w})$ is the transformed source image $\mathbf{x}$, where $\mathbf{w}$ parametrises the transformation. $\mathbf{e}$ models the image mismatch where $\mathbf{e} \sim \mathcal{N}(0, \phi^{-1}\mathbf{I})$, $\mathbf{I}$ is the matrix identity and ϕ is the global image noise precision (inverse variance). We use a free-form deformation (FFD) model [20] using cubic B-splines to provide a smooth mapping. $\mathbf{w}$ is the set of B-spline knot displacements. An FFD model was chosen due to the compact parameter representation, but any transformation model could be used.

Priors are included on all the unknown model parameters. Most importantly, a prior on $\mathbf{w}$ is required to provide regularisation of the mapping, $P(\mathbf{w}) = \mathcal{N}(0; (\lambda\Lambda)^{-1})$, where Λ encodes the bending energy regularisation model and λ is an inferred spatial precision parameter, controlling the strength of the spatial prior. Bending energy is defined as:

$$\int_0^X \int_0^Y \int_0^Z \sum_{d=1}^{3} \lambda \left\{ \left(\frac{\partial^2 G_d}{\partial x^2}\right)^2 + \left(\frac{\partial^2 G_d}{\partial y^2}\right)^2 + \left(\frac{\partial^2 G_d}{\partial z^2}\right)^2 + 2\left[\left(\frac{\partial^2 G_d}{\partial x \partial y}\right)^2 + \left(\frac{\partial^2 G_d}{\partial x \partial z}\right)^2 + \left(\frac{\partial^2 G_d}{\partial y \partial z}\right)^2\right]\right\} dxdydz \tag{2}$$

where G_d refers to the transformation, where the direction is indexed by d. x, y, z refer to locations within the co-ordinate system box, bounded by the origin and X, Y, Z. In this model, G_d is purely defined by $\mathbf{w}$, which are the B-spline coefficients of the FFD model. Λ is calculated by differentiating the effects of $\mathbf{w}$ on G_d, such that $\mathbf{w}^T\Lambda\mathbf{w}$ gives the bending energy of the transformation. The regularisation model used in this work uses free boundary conditions.

The priors on λ and ϕ are modelled using uninformative gamma distributions.

Using variational Bayes [21] the posterior probability distribution of the model parameters $\Theta = \{\mathbf{w}, \phi, \lambda\}$ are approximated using a parametric probability distribution function $q(\Theta) \sim P(\Theta|\mathbf{y})$. The mean-field approximation is used to assume independence in the posterior distribution of parameter groups, $q(\mathbf{w}, \phi\lambda) = q(\mathbf{w})q(\phi)q(\lambda)$. A set of iterative update equations can be derived which fit the parameters of approximate distributions such that they best resemble the data.

Of particular interest in this work is the approximate posterior distribution of transformation parameters, $q(\mathbf{w}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Upsilon}^{-1})$. The update equations for the two hyperparameters of this distribution are given as:

$$\boldsymbol{\Upsilon} = \alpha\bar{\phi}\mathbf{J}^T\mathbf{J} + \bar{\lambda}\boldsymbol{\Lambda} \tag{3}$$

$$\boldsymbol{\Upsilon}\boldsymbol{\mu}_{new} = \alpha\bar{\phi}\mathbf{J}^T(\mathbf{J}\boldsymbol{\mu}_{old} + (\mathbf{y} - \mathbf{t}(\mathbf{x}, \boldsymbol{\mu}_{old}))) \tag{4}$$

where $\boldsymbol{\mu}_{old}$ is the previous mean estimate of the transformation parameters $\mathbf{w}$, and $\mathbf{J}$ is the matrix of first order partial derivatives of the transformation parameters with respect to $\mathbf{t}(\mathbf{x}, \boldsymbol{\mu}_{old})$. $\bar{\lambda}$ and $\bar{\phi}$ are the expectation of the posterior regularisation $q(\lambda)$, and image noise distributions $q(\phi)$. α is a virtual decimation factor [22], which models the correlation in the residual image $(\mathbf{y} - \mathbf{t}(\mathbf{x}, \mathbf{w}))$.

$q(\mathbf{w})$ is an estimate of the posterior distribution of the final inferred mapping parameters. The shape and scale of this distribution is dependent on the structure of the image information, weighted by the noise precision, which indicates the level of mismatch in the fitted model. It also depends on the form of the spatial prior, e.g. the bending energy weighted by the spatial precision which is related to the similarity of the transformation to the spatial prior.

B. Statistical Prediction

Statistical predictors such as statistical classifiers and regressors take as input some feature data $\mathbf{d}$, and output an estimated outcome variable $\hat{o}$. Therefore, for a given class of statistical predictor, h, we can write $\hat{o} = h(\mathbf{d})$.

We are considering the class of supervised learners, which require a labelled training set of N data items, $\mathcal{L} = \{(o_n, \mathbf{d}_n), n = 1, 2, ..., N\}$ where n indexes the subjects in the training set, from which they learn about the relationship between $\mathbf{d}$ and o. Therefore, the trained predictive model provides an estimate for a test subject i based on $\mathcal{L}$, $\hat{o}_i = h(\mathbf{d}_i, \mathcal{L})$.

The relationship between a new test image $\mathbf{d}_i$, and its predicted outcome variable $\hat{o}_i$, is highly dependent on $\mathcal{L}$. Each training item of spatially normalised feature data $\mathbf{d}_n$ in $\mathcal{L}$ is dependent on the inferred image registration. Therefore, training a predictor using $\mathcal{L}$ is susceptible to mis-registration. Accordingly, mis-registration may contribute to errors in prediction.

C. Ensemble Learning to Incorporate Uncertain Registration

The novel aspect of this work is to incorporate the estimated registration uncertainty into statistical prediction using an ensemble learning approach. Ensemble learning methods [16] use a set of predictors to provide a more robust estimate of a prediction. To provide variability in predictive models, they need to be trained using different data sets, $\{\mathcal{L}_m\} \subseteq \{\mathcal{L}\}$, each of length N, where m indexes the different training sets.

The class of ensemble learning methods that we are concerned with use a linear combination of multiple statistical predictors to provide a more robust estimate:

$$\hat{o}_i = \Sigma_m^M \beta_m h(\mathbf{d}_i, \mathcal{L}_m) \tag{5}$$

where i is index of the test subject, M is the number of predictive models, and β is a vector containing the relative weights attributed to each trained prediction model. Only binary statistical predictors are used in this work, although any form of predictor can be used. In the ensemble learning schemes used in this paper, $\beta_m = \frac{1}{M}$, but alternative weighting schemes could be derived using, for example, Bayesian model averaging [23].

A standard approach to generating multiple predictors is bootstrap aggregating, or bagging [17]. In bagging, each $\mathcal{L}_m$ is selected by random uniform sampling with replacement from the set of training subjects. For a large number of bootstraps, each $\mathcal{L}_m$ would be expected to contain 63.2% of the unique training subjects [24]. This approach has been found to be

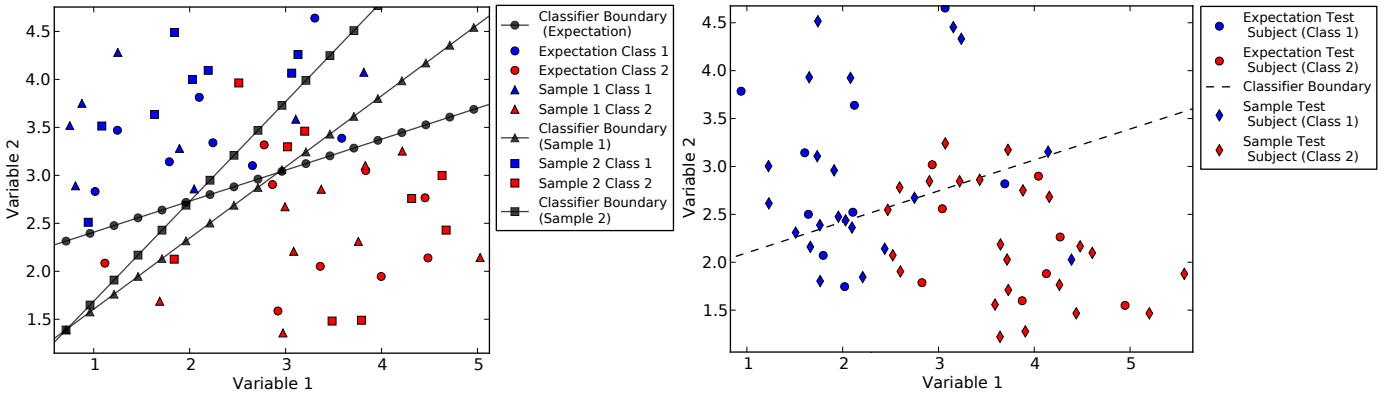

Fig. 1. Graphical examples of how sampled data can be used in classification. The left plot illustrates the scheme **Train+**, where multiple classification boundaries are estimated using random samples of each subject. In this illustration two random samples are drawn from each subject, and thus three classification boundaries can be drawn, two from sets of samples and one using the distribution expectation. The right plot shows the scheme **Test+**, where the variability in classification label can be calculated using random samples of each test subject with a fixed classification boundary. In this case, three random samples were drawn for each subject.

effective at sampling the space of prediction models based on the inter-subject variability. A limitation with bagging is that it only considers the variation between subjects to create different predictive models, whereas in some situations the intrinsic uncertainty of the measurements from which the model is derived may lead to a comparably large source of variability.

In this work, we leverage our knowledge of the derivation of the feature data when creating $\{\mathcal{L}_m, m = 1, M\}$ such that it considers the distribution of feature data as estimated from the set of probable registration mappings, $P(\mathbf{d}|\Theta)$. This is achieved by selecting $(\mathcal{L}_m)_n$ to contain a random sample drawn from each subject's feature data distribution $(P(\mathbf{d}_n|\Theta_n))$. This is a parametric variant of bootstrapping [18], where instead of using observations, new data is drawn from the distribution of observations. We refer to this scheme as **Train+**. A graphical illustration of how multiple classification models can be generated in such a fashion is given in the left plot of Fig. 1. Using a sufficiently large ensemble of statistical predictors should account for the inherent uncertainty in the registration process, and we expect that this exploitation of the instability of predictive models under different training data should lead to an improvement in prediction for all but the most stable of predictors and data.

Our knowledge of $P(\mathbf{d}|\Theta)$ can also be used to provide additional information on the prediction variability for each predictive model. This is achieved by averaging the predicted outcome for a set of random samples drawn from the test subject distribution $P(\mathbf{d}_i|\Theta_i)$, rather than simply testing using only the most likely observation. We refer to this scheme as **Test+**, which is graphically illustrated in the right plot of Fig. 1. **Train+** and **Test+** can be used separately, or can be combined together by using multiple test samples with each predictor in the ensemble. All of these variants can also be incorporated into a bagging framework which would help encapsulate both the inter, and intra-subject variability.

D. Feature Data

A voxel- or deformation tensor-based morphometry [2] approach can be taken to provide a framework for the classification of subjects into their respective groups.

VBM requires the registration of segmentation probability maps, for example grey matter probability maps, of each subject to an atlas space. In a conventional approach, e.g. [3], where a MAP registration tool is used, the image is transformed using only the expectation of the transformation distribution, $\mathbf{d} = \mathbf{t}(\mathbf{x}, \boldsymbol{\mu})$. The transformed grey matter probability map is modulated (multiplied) by the determinant of the warp field Jacobian to compensate for the expansion/contraction of voxels [25].

In TBM, instead of examining the spatially normalised image information, the assumption is made that the discriminative differences between subjects are contained in the deformation field which maps each subject to the atlas space [5]. Feature data is usually derived from the voxelwise 3×3 Jacobian matrix of the transformation of the mapping $\mathbf{J}_m$. Often a scalar measure of the Jacobian matrix is used for comparison, most commonly $\mathbf{d}_v = \log|\mathbf{J}_m|_v$, where v is a voxel index. $|\mathbf{J}_m|_v$ provides a measure of the expansion/contraction of a particular voxel as a result of the mapping. The $\log$ transform is commonly applied to make the data normally distributed [26].

E. Estimating a Distribution of Feature Data

The proposed method requires samples of the feature data under different probable spatial normalisations, rather than simply using the MAP estimate of the mapping. Therefore, the distribution of feature data according to inferred registration parameters needs to be calculated.

A distribution of feature data for each subject $P(\mathbf{d}|\Theta)$, can be estimated for either VBM or TBM data. This is achieved by drawing samples from the inferred approximate posterior distribution of transformation parameters $q(\mathbf{w})$, and calculating the resulting feature data for that sampled mapping. By sampling a large number of mappings, we can build up an estimate of the distribution $P(\mathbf{d}|\Theta)$.

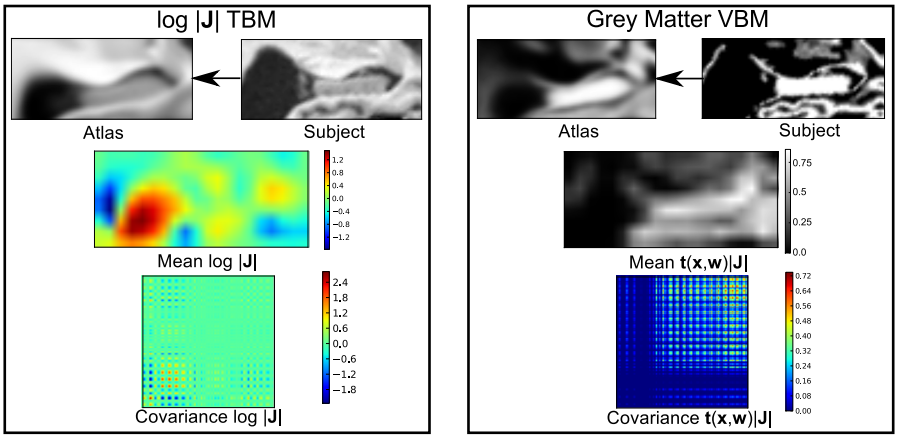

Fig. 2. Examples of the data features acquired when registering a region of interest around the left hippocampus from a subject with AD taken from the test set, to the left atlas image for both TBM (left) and grey matter VBM (right). For both TBM and VBM a single slice of the volume is illustrated. The images marked Atlas and Subject are the high-resolution cropped Atlas and Subject images on which the registration is performed. The mean feature images show the mean of the estimated distribution for the same slice from the sub-sampled feature volume. The covariance matrices illustrate the estimated voxelwise covariance of the feature data for the displayed slice. The covariance matrix is ordered as $y * max(x) + x$ where x and y are positions in the feature image. The bottom left and the top right of the covariance matrix corresponds to the bottom left and top right of the feature image accordingly. The TBM data shows that there is expansion of the ventricle, but also that the $\log |\mathbf{J}_m|$ in this region has a high degree of variance and covariance. The VBM feature data is a grey matter probability map which shows high variability across the image, except in the ventricles where there is no grey matter.

III. Experiments

A. Materials

1) ADNI: Data used in the preparation of this article were obtained from the Alzheimers Disease Neuroimaging Initiative (ADNI) database (adni.loni.ucla.edu). The ADNI was launched in 2003 by the National Institute on Aging (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the Food and Drug Administration (FDA), private pharmaceutical companies and non-profit organizations, as a $60 million, 5-year public-private partnership. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimers disease (AD). Determination of sensitive and specific markers of very early AD progression is intended to aid researchers and clinicians to develop new treatments and monitor their effectiveness, as well as lessen the time and cost of clinical trials.

ADNI is the result of efforts of many coinvestigators from a broad range of academic institutions and private corporations, and subjects have been recruited from over 50 sites across the U.S. and Canada. The initial goal of ADNI was to recruit 800 adults, ages 55 to 90, to participate in the research, approximately 200 cognitively normal older individuals to be followed for 3 years, 400 people with MCI to be followed for 3 years and 200 people with early AD to be followed for 2 years. For up-to-date information, see www.adni-info.org.

2) Subject Grouping: In this work a total of 311 subject images were taken from the ADNI database [27]. 149 of these subjects were patients with Alzheimer's Disease (AD) and 162 were age matched normal controls (NC). These images were broken up into training and testing sets, the properties of which are given in table I.

TABLE I
Statistics of the training and testing subject groups.

Subject group	Mean (years)	Standard deviation (years)	Mean MMSE
AD Train	77.4	7.47	19.4
NC Train	78.7	5.57	29.0
AD Test	76.4	7.50	20.4
NC Test	78.9	4.82	29.1

B. Pipeline

1) Processing: Each image was skull stripped using BET [28] with the option "-B" which deals with bias fields and help remove extra neck voxels. Each image was then registered using an affine registration algorithm [29] to the MNI152 atlas, and re-sampled to 1mm isotropic resolution. For VBM analysis, grey matter probability maps were extracted using FAST [30].

As drawing samples from multivariate normal distributions becomes computationally expensive for large number of transformation parameters, a region of interest (ROI) analysis was required. Atrophy in medial temporal lobe structures, particularly the hippocampus has been shown to be a sensitive marker of Alzheimer's disease [31]. Therefore ROIs of 40x80x36 voxels surrounding the left and right hippocampi were extracted from the structural MR and grey matter images for registration.

2) Atlas Creation: For spatial normalisation, most methods require the definition of an atlas space. The use of an exemplar subject to define an atlas space can induce a statistical bias. Therefore, in an effort to reduce the bias, a sharp atlas can be estimated estimated from the subjects in the training set in an iterative manner as in [32]. The initial atlas estimate is made by averaging the ROI feature data in the training set. Subsequent iterations non-rigidly register each of the images using our Bayesian non-rigid registration algorithm, with a 5mm knot spacing, to the current atlas estimate. The new atlas

estimate is created by averaging the intensities of the registered images, and deforming this image by the average warp from the atlas to the subjects. This is calculated by averaging the inverse of the transformations from all the subjects to the atlas. The B-spline FFD transformations are constrained to be diffeomorphic using the method of [33], which allows a smooth and well defined inverse to be estimated. This procedure is repeated until the estimated atlas image changes by less than 1% between iterations. This is used to create a structural MR and grey matter atlas for each ROI.

3) Feature Generation: To create the feature data, each subject image was non-rigidly registered to the relevant atlas image using either a 5mm FFD knot spacing for TBM to give high-resolution features, and a 10mm spacing for VBM to align the images, but still retain subject differences. Once the registration algorithm has converged, warp samples are drawn from $q(\mathbf{w})$ to characterise $P(\mathbf{d}|\mathbf{w})$. To avoid artefacts related to the edges of the ROI, the voxels within 4mm of the edge are removed, leaving a feature region of $32 \times 72 \times 28$ voxels. 4mm was chosen because the sampled deformations around the edge of the image are very unlikely to exceed this.

To allow the tractable storage of samples from $P(\mathbf{d}|\mathbf{w})$, the feature data needs to be sub-sampled by a factor of 4. This gives a total of 1008 voxels. To make the classification step computationally efficient, 3600 samples of the data feature are stored per subject to provide a sufficiently accurate description of the distribution, rather than fitting the data to a parametric distribution and later sampling from it. In the classification stages, samples are randomly chosen for each subject in both the Train+ and Test+ methodologies. In practice, all of these samples may be required for a run using Test+. An example illustration of VBM and TBM features for a subject in the test set with Alzheimer's disease is given in Fig. 2.

C. Classification

In the experiments we use a linear support vector machine (SVM) [34] as implemented in [35], using all 1800 feature voxels to classify between subject groups. Subject age is regressed out for each voxel based on the empirical expectation of $P(\mathbf{d}|\mathbf{w})$ for the healthy controls in the training set using ordinary least squares linear regression [36], and each voxel is given 0 mean and unit standard deviation.

To select the most appropriate SVM classifier parameter for use in the Original scheme, we use a leave-one-out cross-validation procedure on the training set to test. This tests a range of the soft margin penalty parameter values, $C = [2e^{-15}, 2e^{-14}, ..., 2e^{15}]$, to find the value with the best generalisation accuracy. The optimal parameter and its corresponding correct rate, defined as the ratio of correctly identified examples from the number of testing examples, are given in table II. For the SVM classifiers used within the ensemble learning schemes, $C = 2e^{15}$ which effectively removes the soft margin. This is beneficial as the training data is usually linearly separable, and removing the soft-margin introduces greater variability between classifiers as there is a greater dependence on the training data.

In our classification experiments we compare 7 different training schemes:

TABLE II
LINEAR SVM SOFT MARGIN PARAMETER C AS SELECTED BY LEAVE ONE OUT CROSS VALIDATION (LOOCV) USING THE EMPIRICAL EXPECTATION OF THE DATA FEATURES FOR EACH OF THE DIFFERENT DATA TYPES.

Feature data	LOOCV correct rate	C
L $\log\|\mathbf{J}\|$	0.821	$2e^{-11}$
R $\log\|\mathbf{J}\|$	0.821	$2e^{-12}$
L VBM	0.877	$2e^{-10}$
R VBM	0.827	$2e^{-10}$

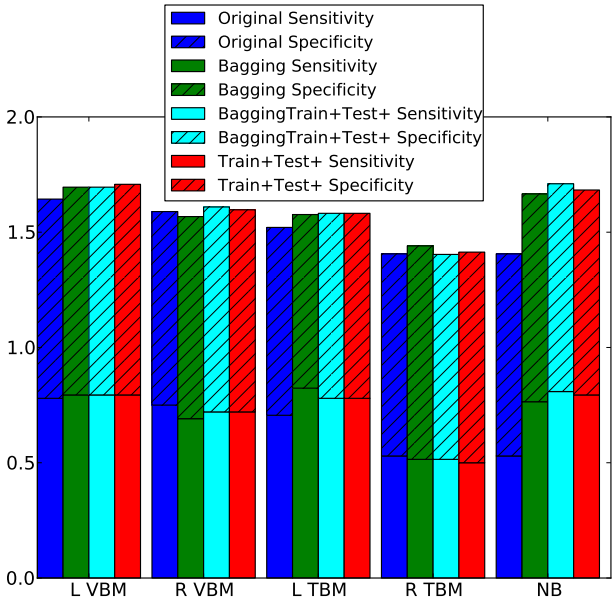

Fig. 3. Stacked bar chart illustrating the sensitivity and specificity of the classification of Alzheimer's disease in selected experiments. L and R represent left and right ROI images respectively. NB refers to the results of the naive Bayes combination of all the data features. It can be seen that BaggingTrain+Test+ and Train+Test+ outperform, or do as well as standard Bagging for all the features types, except R TBM, which is the lowest performing feature for all methods. BaggingTrain+Test+, when combining ensembles using naïve Bayes, gives the best trade-off of classification sensitivity, and specificity of any of the approaches considered.

- **Original**, train using the empirical expectation of the data in the whole training set, and test using the empirical expectation of the testing data.
- **Train+**, train using a random sample of each subject in the training set.
- **Bagging**, the standard bootstrap aggregating approach where the training set is sampled with replacement.
- **BaggingTrain+**, where a random subject sample is used within a bagging scheme.
- **Test+**, train using the empirical expectation of the whole training set and test using 20 random samples for each subject in the test set.
- **BaggingTest+**, the combination of bagging and test+.
- **Train+Test+**, the combination of train+ and test+.

In our experiments 300 classification models were generated to make an ensemble as this was sufficient for convergence for all methods. A summary of the results of these experiments is given in Fig. 3, and the classification correct rate for all of the methods is given in table III.

As we can see from table III, the left hippocampus provides

TABLE III

CLASSIFICATION CORRECT RATE USING THE DIFFERENT PREDICTOR TRAINING AND TESTING VARIANTS USING TBM AND VBM FEATURE DATA. L AND R INDICATE THE LEFT AND RIGHT HIPPOCAMPUS DATA, RESPECTIVELY. NAÏVE BAYES REFERS TO THE COMBINATION OF SOFT PROBABILITIES FROM DIFFERENT FEATURE DATA TYPES. TRAIN+TEST+ AND IT'S VARIANTS TEND TO DO AS WELL, OR BETTER THAN A STANDARD BAGGING APPROACH FOR BOTH VBM AND TBM. BAGGINGTRAIN+TEST+ PROVIDES THE BEST OVERALL CLASSIFICATION RESULTS.

Feature data	Original	Train+	BaggingTrain+	Bagging	Test+	Train+Test+	BaggingTrain+Test+
L TBM	0.765	0.792	**0.799**	0.785	0.765	0.792	**0.799**
R TBM	0.718	0.7181	0.725	**0.738**	0.7181	0.7248	0.718
L VBM	0.826	0.846	0.852	0.852	0.826	**0.859**	0.852
R VBM	0.799	0.805	0.805	0.792	0.799	0.805	**0.812**
Naive Bayes All	0.718	0.846	0.852	0.839	0.832	0.846	**0.859**

stronger features than the right for discriminating between Alzheimer's disease and age matched normal controls for all methods and both feature types. VBM provides good separation for both left and right hippocampi, whereas only TBM on the left hippocampus provided a reasonable level of discrimination. Fig. 3 provides a summary of some of the results illustrating the sensitivity and specificity of each ensemble. Sensitivity is defined as the proportion of correctly identified disease cases, out of the total number of disease cases. Specificity is the proportion of correctly identified control subjects, out of the total number of controls. This summary shows that ensemble learning approaches generally provide more accurate classification than the Original approach. All of the Train+Test+ approaches outperform, or do as well as Bagging for all features except right TBM, which produced the lowest classification results for all methods. This implies that the additional data variability provided by registration uncertainty assists in creating a more accurate ensemble.

Furthermore, we find the largest improvement over the original approach in the left hippocampus TBM feature data, particularly using the BaggingTrain+ schemes. The strength of this improvement is likely to be due to the data feature being derived from the warp field. Small changes in warp field which might have little effect on the image likelihood, may lead to more substantial changes in $\log|\mathbf{J}_m|$. This is likely to also be a contributing factor in the improvement in the VBM results, as the warped GM probability map is modified by the $|\mathbf{J}_m|$. The Test+ schemes do not have much impact on classification correct rates.

In terms of computational time, the Original scheme is fastest, as only 1 classifier is constructed. Including the overhead of loading, and pre-processing the data, the training takes approximately 5 seconds. The use of 300 bootstrapping samples in Bagging takes 5 minutes to complete. Train+ and BaggingTrain+ take about 5 minutes of CPU time. However, because of the slow speed of disk access which is required to load the samples, Train+ and BaggingTrain+ take 10-15 minutes. The use of Test+ schemes adds an additional 30 minutes to run-time. This extra time is almost entirely taken up by disk access. Once the ensembles have been created, classification of a new sample is very fast, ≈ 1 second, and 10 seconds using Test+. All of the experiments were conducted on a dual core 2.8GHz laptop with a serial ATA (7200RPM) hard-drive.

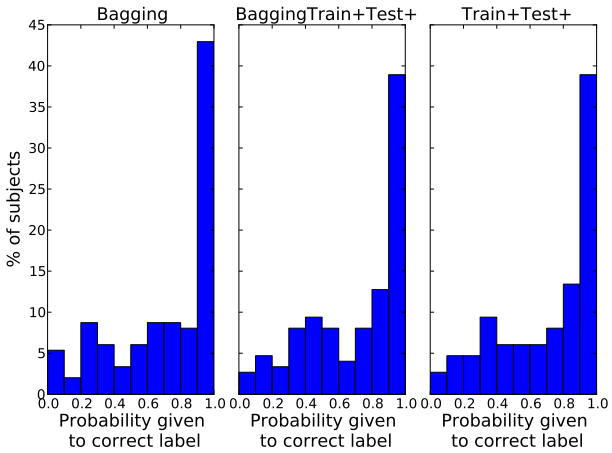

Fig. 4. Histogram of the probability estimates given by a classification ensemble, to the correct class label. These histograms are plotted for the right hippocampus TBM feature data. Although Bagging produces a more accurate classification for this data than the other schemes, it is also over-confident in some estimates which are incorrect. This can be seen in the larger number of subjects which are assigned a very low probability of belonging to the true class which they are from. This relatively poor estimate of uncertainty given by the Bagging ensemble classifiers is a likely cause of the worse performance of Bagging in the naïve Bayes classification results shown in table III.

D. Combining Soft Classification Probabilities

Each ensemble gives a soft classification result for each subject as it is an average of multiple predictions, given by:

$$P(o_i|\mathbf{d}_i^j,\{\mathcal{L}^j\}) = \Sigma_m^M \frac{1}{M} h(\mathbf{d}_i^j,\mathcal{L}_m^j) \tag{6}$$

where h is restricted to being a binary classifier and the superscript j denotes the feature data type. An illustration of the soft classification resulting from the ensemble in given in Fig. 4. This figure shows that Train+Test+ and BaggingTrain+Test+ give greater probability to the correct classification for subjects which they get wrong, than bagging does. To assess how reasonable the soft predictions given by each classifier ensemble are, they can be combined in a post-classifier analysis. Here, we combine the soft classification probabilities using a naive Bayesian classifier:

$$P(o_i|\mathbf{d}_i,\{\mathcal{L}\}) = \frac{\Pi_j^J P(o_i|\mathbf{d}_i^j,\{\mathcal{L}^j\})}{\Pi_j^J P(o_i|\mathbf{d}_i^j,\{\mathcal{L}^j\}) + \Pi_j^J P(\neg o_i|\mathbf{d}_i^j,\{\mathcal{L}^j\})} \tag{7}$$

where $\neg$ indicates the alternative classification label in a binary classification problem. Results for the naïve Bayes combination are provided in Table III and Fig. 3. The combination

of Bagging with Train+Test+ leads to the most reasonable soft-predictions, as shown by the highest correct rate and best sensitivity/specificity trade-off. All of the Train+ schemes also outperform Bagging in the naïve Bayes classification. This is likely to be caused by Bagging showing greater prevalence for over-confidence in incorrect predictions, as illustrated in Fig. 4. It can be seen that the use of Test+, on its own, or in combination with BaggingTrain+ is beneficial, implying the additional testing samples helps to estimate the uncertainty of prediction.

IV. Discussion and Conclusions

In this work we have demonstrated a scheme in which registration uncertainty can be incorporated into an ensemble learning scheme to provide more accurate prediction, with more reasonable estimates of classifier uncertainty, than standard approaches such as bootstrap aggregation. This was achieved by sampling probable registration transformations inferred from a probabilistic registration algorithm, and then estimating the distribution of a data feature given the uncertainty in the registration. Samples of the feature data distribution are used in place of the most likely observations in the training and testing phase for statistical predictors. The proposed approach generates prediction variability from the intra-subject uncertainty as opposed to the inter-subject variation, as is achieved by boostrapping. We describe a method of combining predictors trained using sampled data into an ensemble. In our experiments, we provide results on the problem of classification of subjects with Alzheimer's disease, from age matched healthy controls using a linear SVM.

Our results demonstrate that the proposed scheme tends to lead to an improvement in classification correct rate over a standard scheme and bootstrap aggregating when examining grey matter voxel based morphometry, and tensor based morphometry. This implies that the registration uncertainty contains more useful information for the discriminative problem than that obtained from bootstrapping. We also found that improved classification results can be achieved by the naive Bayesian combination of the ensembles created from the separate data features. Here, we showed that the variability induced by bootstrap aggregating provides less reasonable estimates of prediction uncertainty than the proposed approach, and this is reflected in our results.

In this work the strength of a prior with a fixed covariance structure, based on bending energy, is inferred from the data. A fixed prior covariance structure expects a similarly smooth transformation across the image. The choice of the prior covariance structure will have an effect on the posterior transformation distribution, particularly in regions where little image information is available. The structure of this prior distribution could be estimated from the data. Such an approach could provide a more biologically plausible prior. Some general approaches have been proposed which allow the estimation of the parameter covariance structure from the data using variational methods [37][38]. These methods are likely to be computationally expensive, sometimes necessitating the approximation of independence between regions [39].

If the prior distribution, $p(\mathbf{w})$, with any associated parameters is known, then it may be possible to draw samples of the true posterior distribution of transformation parameters, $p(\mathbf{w}|\mathbf{y}, \phi, \lambda)$, using MCMC in a computationally efficient manner. Such approaches have been previously described for registration [40][14], but in these cases λ was predetermined for a group of subjects, which may have an effect on the inferred distribution. In future work we will experiment with MCMC inference of $p(\mathbf{w}|\mathbf{y}, \phi, \lambda)$, fixing λ and ϕ based on the expectation of $q(\lambda)$ and $q(\phi)$ respectively.

The choice of atlas is also an important prior of this model, and therefore if it is biased towards a particular anatomy, it may affect the estimated distribution of probable registrations, $q(\mathbf{w})$. Future work will experiment with a groupwise registration approach, which should help to reduce any atlas induced bias.

An alternative approach to utilise the full estimated distribution of registration mappings would be to use the transformation parameters themselves as data features. As the inferred distribution is multivariate normal, there is a finite length description of the distribution. This description could be used directly, rather than through sampling the distribution. A disadvantage of using such an approach is that some of the interpretability associated with VBM/TBM is lost when using the transformation parameters directly as features.

An entirely different approach to mitigate mis-registration would be to use multiple atlases. This has been successfully applied in integrating out registration from propagated segmentations [41]. Koikkalainen et al. [42] suggest approaches to generating more robust TBM features by registering each subject to several atlases, which are all registered to the reference space. This work is complementary, and further investigation could compare the benefits of each approach.

Posterior probabilities can be directly estimated from a variety of classifiers, including SVMs, but in a more principled manner from logistic regression or relevance vector machines [43]. Such probabilistic classification estimates, or regression outputs, could be incorporated within the proposed framework. Future work could investigate including and comparing against uncertainty estimates from these classifiers.

In this work we have presented results discriminating subjects with AD from age matched healthy controls. A more interesting clinical problem lies in the diagnosis of subjects suffering from mild cognitive impairment [44], and the prognosis of those who may convert to AD. Imaging data of such subjects is available in ADNI, and this analysis will be investigated in future work.

We have shown that incorporating registration uncertainty leads to an improved classification performance over standard approaches for this particular experimental pipeline. However, this experimental process itself may not be optimal for the classification problem, and further work could be carried out to improve the overall classification accuracy to match current state of the art pipelines. Firstly, we could consider looking at multiple functional areas, and combining their soft-classification probabilities. We could consider a more flexible ensemble learning scheme, such as Bayesian model averaging. Feature selection could be used in place of, or following

voxel sub-sampling for creating the feature data distribution, to select the most discriminative voxels. In particular, the choice of classifier, and its associated parameterisation could be addressed as well as the data processing scheme. Nevertheless, the proposed approach has demonstrated that information that is pertinent to the classification problem can be exploited from the uncertainty of image registration.

V. Acknowledgment

I. J. A. Simpson would like to acknowledge funding from the EPSRC through the Life Sciences Interface Doctoral Training Centre, Oxford, UK.

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: Abbott; Alzheimers Association; Alzheimers Drug Discovery Foundation; Amorfix Life Sciences Ltd.; AstraZeneca; Bayer HealthCare; BioClinica, Inc.; Biogen Idec Inc.; Bristol-Myers Squibb Company; Eisai Inc.; Elan Pharmaceuticals Inc.; Eli Lilly and Company; F. Hoffmann-LaRoche Ltd and its affiliated company Genentech, Inc.; GE Healthcare; Innogenetics, N.V.;IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Medpace, Inc.; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Servier; Synarc Inc.; and Takeda Pharmaceutical Company. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for NeuroImaging at the University of California, Los Angeles. This research was also supported by NIH grants P30 AG010129 and K01 AG030514.

References

[1] R. Cuingnet, E. Gerardin, J. Tessieras, G. Auzias, S. Lehéricy, M. Habert, M. Chupin, H. Benali, and O. Colliot, "Automatic classification of patients with Alzheimer's disease from structural MRI: A comparison of ten methods using the ADNI database," *Neuroimage*, vol. 56, no. 2, pp. 766–781, 2011.

[2] J. Ashburner and K. Friston, "Voxel-based morphometry–the methods," *Neuroimage*, vol. 11, no. 6, pp. 805–821, 2000.

[3] S. Klöppel, C. Stonnington, C. Chu, B. Draganski, R. Scahill, J. Rohrer, N. Fox, C. Jack Jr, J. Ashburner, and R. Frackowiak, "Automatic classification of MR scans in Alzheimer's disease," *Brain*, vol. 131, no. 3, pp. 681–689, 2008.

[4] B. Magnin, L. Mesrob, S. Kinkingnhun, M. Plgrini-Issac, O. Colliot, M. Sarazin, B. Dubois, S. Lehricy, and H. Benali, "Support vector machine-based classification of Alzheimers disease from whole-brain anatomical MRI," *Neuroradiology*, vol. 51, pp. 73–83, 2009.

[5] X. Hua, A. Leow, N. Parikshak, S. Lee, M. Chiang, A. Toga, C. Jack Jr, M. Weiner, and P. Thompson, "Tensor-based morphometry as a neuroimaging biomarker for Alzheimer's disease: an MRI study of 676 AD, MCI, and normal subjects," *Neuroimage*, vol. 43, no. 3, pp. 458–469, 2008.

[6] A. Klein and et al., "Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration," *Neuroimage*, vol. 46, no. 3, pp. 786–802, 2009.

[7] P. Risholm, E. Samset, and W. Wells, "Bayesian estimation of deformation and elastic parameters in non-rigid registration," *WBIR*, pp. 104–115, 2010.

[8] I. Simpson, J. Schnabel, A. Groves, J. Andersson, and M. Woolrich, "Probabilistic inference of regularisation in non-rigid registration," *NeuroImage*, vol. 59, no. 3, pp. 2438–2451, 2012.

[9] S. Allassonniére, Y. Amit, and A. Trouvè, "Toward a coherent statistical framework for dense deformable template estimation," *Journal of the Royal Statistical Society, Series B*, vol. 69, no. 2, 2007.

[10] K. Van Leemput, "Encoding probabilistic brain atlases using bayesian inference," *Medical Imaging, IEEE Transactions on*, vol. 28, no. 6, pp. 822–837, 2009.

[11] M. Hub, M. Kessler, and C. Karger, "A stochastic approach to estimate the uncertainty involved in B-spline image registration," *IEEE Transactions on Medical Imaging*, vol. 28, no. 11, pp. 1708 –1716, 2009.

[12] P. Risholm, S. Pieper, E. Samset, and W. Wells, "Summarizing and visualizing uncertainty in non-rigid registration," in *MICCAI 2010*, ser. LNCS, T. Jiang, N. Navab, J. Pluim, and M. Viergever, Eds. Springer, Heidelberg, 2010, vol. 6362, pp. 554–561.

[13] I. J. A. Simpson, M. Woolrich, A. Groves, and J. Schnabel, "Longitudinal brain MRI analysis with uncertain registration," in *MICCAI 2011*, ser. LNCS, G. Fichtinger, A. Martel, and T. Peters, Eds. Springer, Heidelberg, 2011.

[14] J. Iglesias, M. Sabuncu, and K. Leemput, "Incorporating parameter uncertainty in bayesian segmentation models: Application to hippocampal subfield volumetry," in *MICCAI 2012*, ser. LNCS, N. Ayache, H. Delingette, P. Golland, and K. Mori, Eds. Springer, Heidelberg, 2012, vol. 7512, pp. 50–57.

[15] P. Risholm, J. Balter, and W. Wells, "Estimation of delivered dose in radiotherapy: the influence of registration uncertainty," in *MICCAI 2011*, ser. LNCS, G. Fichtinger, A. Martel, and T. Peters, Eds. Springer, Heidelberg, 2011, pp. 548–555.

[16] T. Dietterich, "Ensemble methods in machine learning," *Multiple classifier systems*, pp. 1–15, 2000.

[17] L. Breiman, "Bagging predictors," *Machine learning*, vol. 24, no. 2, pp. 123–140, 1996.

[18] B. Efron, "Bootstrap methods: another look at the jackknife," *The annals of Statistics*, vol. 7, no. 1, pp. 1–26, 1979.

[19] I. Simpson, J. Schnabel, J. Andersson, A. Groves, and M. Woolrich, "Ensemble learning incorporating uncertain registration," *Proceedings of MIUA 2012*, 2012.

[20] D. Rueckert, L. Sonoda, C. Hayes, D. Hill, M. Leach, and D. Hawkes, "Nonrigid registration using Free-Form Deformations: application to breast MR images," *IEEE Transactions on Medical Imaging*, vol. 18, no. 8, pp. 712–721, 1999.

[21] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul, "An introduction to variational methods for graphical models," *Machine learning*, vol. 37, no. 2, pp. 183–233, 1999.

[22] A. R. Groves, C. F. Beckmann, S. M. Smith, and M. W. Woolrich, "Linked independent component analysis for multimodal data fusion," *NeuroImage*, vol. 54, no. 3, pp. 2198 – 2217, 2011.

[23] J. Hoeting, D. Madigan, A. Raftery, and C. Volinsky, "Bayesian model averaging: a tutorial," *Statistical science*, pp. 382–401, 1999.

[24] B. Efron, "Estimating the error rate of a prediction rule: improvement on cross-validation," *Journal of the American Statistical Association*, vol. 78, no. 382, pp. 316–331, 1983.

[25] C. D. Good, I. S. Johnsrude, J. Ashburner, R. N. Henson, K. J. Friston, and R. S. Frackowiak, "A voxel-based morphometric study of ageing in 465 normal adult human brains," *NeuroImage*, vol. 14, no. 1, pp. 21 – 36, 2001.

[26] A. Leow, I. Yanovsky, M. Chiang, A. Lee, A. Klunder, A. Lu, J. Becker, S. Davis, A. Toga, and P. Thompson, "Statistical properties of jacobian maps and the realization of unbiased large-deformation nonlinear image registration," *Medical Imaging, IEEE Transactions on*, vol. 26, no. 6, pp. 822–832, 2007.

[27] S. Mueller, M. Weiner, L. Thal, R. Petersen, C. Jack, W. Jagust, J. Trojanowski, A. Toga, and L. Beckett, "Alzheimer's Disease Neuroimaging Initiative," *Advances in Alzheimer's and Parkinson's Disease*, pp. 183–189, 2008.

[28] S. Smith, "Fast robust automated brain extraction," *Human Brain Mapping*, vol. 17, no. 3, pp. 143–155, 2002.

[29] M. Jenkinson and S. Smith, "A global optimisation method for robust affine registration of brain images," *Medical image analysis*, vol. 5, no. 2, pp. 143–156, 2001.

[30] Y. Zhang, M. Brady, and S. Smith, "Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm," *Medical Imaging, IEEE Transactions on*, vol. 20, no. 1, pp. 45–57, 2001.

[31] C. Jack Jr, R. Petersen, Y. Xu, S. Waring, P. O'Brien, E. Tangalos, G. Smith, R. Ivnik, and E. Kokmen, "Medial temporal atrophy on MRI in normal aging and very mild Alzheimer's disease," *Neurology*, vol. 49, no. 3, pp. 786–794, 1997.

[32] A. Guimond, J. Meunier, and J. Thirion, "Average brain models: A convergence study," *Computer vision and image understanding*, vol. 77, no. 2, pp. 192–210, 2000.

[33] B. Karacali and C. Davatzikos, "Estimating topology preserving and smooth displacement fields," *IEEE Transactions on Medical Imaging*, vol. 23, no. 7, pp. 868–880, 2004.

[34] N. Cristianini and J. Shawe-Taylor, *An introduction to support Vector Machines: and other kernel-based learning methods*. Cambridge University Press, 2000.

[35] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011.

[36] J. Dukart, M. Schroeter, and K. Mueller, "Age correction in dementia–matching to a healthy brain," *PloS one*, vol. 6, no. 7, p. e22193, 2011.

[37] L. Harrison, W. Penny, J. Daunizeau, and K. Friston, "Diffusion-based spatial priors for functional magnetic resonance images," *Neuroimage*, vol. 41, no. 2, pp. 408–423, 2008.

[38] K. Friston, L. Harrison, J. Daunizeau, S. Kiebel, C. Phillips, N. Trujillo-Barreto, R. Henson, G. Flandin, and J. Mattout, "Multiple sparse priors for the M/EEG inverse problem," *NeuroImage*, vol. 39, no. 3, pp. 1104–1120, 2008.

[39] L. Harrison, W. Penny, G. Flandin, C. Ruff, N. Weiskopf, and K. Friston, "Graph-partitioned spatial priors for functional magnetic resonance images," *NeuroImage*, vol. 43, no. 4, pp. 694 – 707, 2008. [Online]. Available: http://www.sciencedirect.com/science/article/B6WNP-4T8SM1K-2/2/39dec6701b86958bbe8d8a6ab21559b3

[40] S. Allassonnière, E. Kuhn, and A. Trouvé, "Construction of bayesian deformable models via a stochastic approximation algorithm: a convergence study," *Bernoulli*, vol. 16, no. 3, pp. 641–678, 2010.

[41] T. Rohlfing, D. Russakoff, and C. Maurer, "Performance-based classifier combination in atlas-based image segmentation using expectation-maximization parameter estimation," *Medical Imaging, IEEE Transactions on*, vol. 23, no. 8, pp. 983–994, 2004.

[42] J. Koikkalainen, J. Lötjönen, L. Thurfjell, D. Rueckert, G. Waldemar, and H. Soininen, "Multi-template tensor-based morphometry: application to analysis of alzheimer's disease," *NeuroImage*, vol. 56, no. 3, pp. 1134–1144, 2011.

[43] M. Tipping, "Sparse bayesian learning and the relevance vector machine," *The Journal of Machine Learning Research*, vol. 1, pp. 211–244, 2001.

[44] R. Petersen, "Mild cognitive impairment as a diagnostic entity," *Journal of internal medicine*, vol. 256, no. 3, pp. 183–194, 2004.