NeuroImage: Clinical 2 (2013) 735–745

Contents lists available at SciVerse ScienceDirect

NeuroImage: Clinical

journal homepage: www.elsevier.com/locate/ynicl

Accurate multimodal probabilistic prediction of conversion to Alzheimer's disease in patients with mild cognitive impairment ☆

Jonathan Young [a,*], Marc Modat [a], Manuel J. Cardoso [a], Alex Mendelson [a], Dave Cash [a,b], Sebastien Ourselin [a,b], the Alzheimer's Disease Neuroimaging Initiative [1]

[a] Centre for Medical Image Computing, University College London, UK
[b] Dementia Research Centre, Institute of Neurology, University College London, UK

ARTICLE INFO

Article history:
Received 14 February 2013
Received in revised form 7 May 2013
Accepted 8 May 2013
Available online xxxx

Keywords:
Alzheimer's disease
Mild cognitive impairment
Gaussian process
Support vector machine
Multimodality
Probabilistic classification
Risk scores

ABSTRACT

Accurately identifying the patients that have mild cognitive impairment (MCI) who will go on to develop Alzheimer's disease (AD) will become essential as new treatments will require identification of AD patients at earlier stages in the disease process. Most previous work in this area has centred around the same automated techniques used to diagnose AD patients from healthy controls, by coupling high dimensional brain image data or other relevant biomarker data to modern machine learning techniques. Such studies can now distinguish between AD patients and controls as accurately as an experienced clinician. Models trained on patients with AD and control subjects can also distinguish between MCI patients that will convert to AD within a given timeframe (MCI-c) and those that remain stable (MCI-s), although differences between these groups are smaller and thus, the corresponding accuracy is lower. The most common type of classifier used in these studies is the support vector machine, which gives categorical class decisions. In this paper, we introduce Gaussian process (GP) classification to the problem. This fully Bayesian method produces naturally probabilistic predictions, which we show correlate well with the actual chances of converting to AD within 3 years in a population of 96 MCI-s and 47 MCI-c subjects. Furthermore, we show that GPs can integrate multimodal data (in this study volumetric MRI, FDG-PET, cerebrospinal fluid, and APOE genotype with the classification process through the use of a mixed kernel). The GP approach aids combination of different data sources by learning parameters automatically from training data via type-II maximum likelihood, which we compare to a more conventional method based on cross validation and an SVM classifier. When the resulting probabilities from the GP are dichotomised to produce a binary classification, the results for predicting MCI conversion based on the combination of all three types of data show a balanced accuracy of 74%. This is a substantially higher accuracy than could be obtained using any individual modality or using a multikernel SVM, and is competitive with the highest accuracy yet achieved for predicting conversion within three years on the widely used ADNI dataset.

1. Introduction

The most common form of dementia in the elderly population is Alzheimer's disease (AD), with prevalence expected to increase greatly in coming years largely due to ageing and expected improvements in care (Ferri et al., 2005). AD is a progressive condition initially associated with impairment of episodic memory, followed by other cognitive domains, leading to increasing dependence and ultimately to death. While familial AD is entirely genetic in aetiology, it constitutes only a small minority (1–2%) of all cases. The more prevalent sporadic form of AD has far more complex causes. While lifestyle and genetic risk factors are significant, ageing remains by far the greatest risk factor. Current treatments for AD are palliative in nature, relieving the symptoms to some degree without tackling the underlying causes of the disease. As such, they cannot halt or even slow down the disease process. However newer treatments are intended to

* Corresponding author at: UCL Department of Medical Physics and Bioengineering — Centre for Medical Image Computing, Gower Street, London, WC1E 6BT, UK. Tel.: +44 2076790485; fax: +44 2076790255.
E-mail address: jonathan.young@ucl.ac.uk (J. Young).

[1] 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://adni.loni.ucla.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

2213-1582/$ – see front matter © 2013 The Authors. Published by Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.nicl.2013.05.004

address this limitation by interfering with the amyloid cascade that is thought to be one of the underlying causes of AD (Robert and Wark, 2012). While these hold great promise, to be effective treatment must begin in a much earlier stage in the disease process than current treatments. As a result, there needs to be a shift in focus from diagnosis to prognosis of AD in patients showing very mild symptoms, and eventually to people with no symptoms at all. Previous work has focused on studying patients with mild cognitive impairment (MCI) (Petersen et al., 1999). MCI is typically defined as a state where patients have isolated memory deficits that are not severe enough to affect normal living. Studies have shown that MCI patients convert to AD at an annual rate of 10–15% per year (Braak and Braak, 1995). MCI patients who do not convert to AD either develop other forms of dementia, remain stable, or in a small minority, revert to a nondemented state. Therefore predicting which MCI patients will develop AD in the short term (i.e. within a few years) and which will remain stable is extremely relevant to future treatments. Although a definitive diagnosis of AD can be made only at autopsy, in practice expert clinicians diagnose AD based on clinical history and batteries of cognitive tests. However these standard clinical tests are not able to identify the more subtle patterns of the disease process at this early stage, so more advanced methods are required.

The automated methods used to discriminate between stable (MCI-s) and converter (MCI-c) patients are similar to those used for diagnosis of AD. These automated tests use imaging and other biomarker data, and can now diagnose AD with an accuracy of about 90%, as accurately as expert clinicians can using more traditional methods. (Beach et al., 2012). While a number of different imaging modalities have been proposed for this application, the majority have used structural MRI as atrophy in specific brain regions is one of the most established hallmarks of AD. The features used in classification derived from structural MRI can take a number of forms, including voxel level maps of grey matter density (Fan et al., 2008; Klöppel et al., 2008; Nho et al., 2010), volume or shape (Barnes et al., 2004; Ferrarini et al., 2009; Gerardin et al, 2009; Zhang et al., 2011), or cortical thickness measurements (Desikan et al., 2009; Eskildsen et al., 2013; Lerch et al., 2008; Querbes et al., 2009). These features can be calculated over the whole brain or specific structures known to be affected by AD, such as the hippocampus. A comprehensive review and comparison of these methods, focused mainly on the type of MRI-derived features used rather than which machine learning algorithm was implemented, is given in Cuingnet et al. (2010).

Looking beyond structural MRI, fluorodeoxyglucose positron emission tomography (FDG-PET) is capable of measuring the level of glucose metabolism in the brain. Studies have shown that glucose metabolism is reduced in some regions in patients before they develop AD (Drzezga et al., 2003; Mosconi et al., 2010) and this may be used to classify AD patients from controls or predict conversion from MCI to AD (Gray et al., 2012). Biomarkers extracted from cerebrospinal fluid (CSF) have shown utility in the diagnosis of AD or MCI. In particular, CSF levels of total tau protein (t-tau) and phosphorylated tau (p-tau) proteins, known to be implicated in the formation of neurofibrillary tangles that cause atrophy in AD, are elevated, in AD patients, while levels of the amyloid-β_{42} ($a\beta_{42}$) peptide in CSF fell (Fjell et al., 2010a; Holtzman, 2011). Measurements of amyloid load in the brain using amyloid PET have shown similar results (Rowe et al., 2010). Also, variants of the apolipoprotein E (APOE) gene affect the risk of developing AD (Corder et al., 1993, 1994).

These different types of biomarker data have been shown to be complementary, meaning that they provide superior classification when used in combination than when either is used individually, even if they are correlated (Fjell et al., 2010b; Landau et al., 2010). Thus a number of studies have sought to make use of multiple biomarker types in classification. Structural MRI is used in combination with genetic data in Vemuri et al. (2008) and with CSF biomarkers in Vemuri et al. (2009) and Davatzikos et al. (2011). Structural MRI data, FDG-PET and CSF data are used in Hinrichs et al. (2011), Walhovd et al. (2010) and Zhang et al. (2011). A noteworthy disadvantage of multimodal methods is that the problem of missing data is multiplied, as a subject must be discarded entirely or the missing data must be synthesised if it is not present in any one of the modalities used. An approach to tackle this issue is presented in (Yuan et al., 2012).

The most popular classification method is the support vector machine (SVM), due to its accuracy and ability to cope with very high dimensional data. Another advantage of the SVM is its ability to use the kernel, a matrix of size n by n that summarises the distances or covariances between n training subjects. This can be applied to learn from multimodal data. Rather than simply concatenating the features from different modalities into a single vector, an individual kernel can be formed from each modality and then a combined kernel generated as a weighted sum of the individual ones. Both Zhang et al. (2011) and Hinrichs et al. (2011) use this approach, but find the individual kernel weights in a different fashion, the former choosing them by a grid search for the weights giving the best accuracy in a nested cross validation loop, and the latter by optimising them alongside the standard SVM parameters and with the standard SVM objective function, a method more broadly known as multiple kernel learning (Bach et al., 2004).

In this paper we present a different method using a combination of structural MRI, FDG-PET, CSF and APOE data to classify MCI-s and MCI-c patients. Primarily, we use Gaussian process (GP) classification, which is a probabilistic classification algorithm. Bishop (2007) lists four general advantages of a probabilistic framework, however for this particular study we would add two more which we feel to be particularly relevant: firstly, the option to tune free parameters automatically from the training data, avoiding the need for computationally expensive cross-validation loops, and secondly, that the probabilistic decisions produced by GP classification allow a great deal of flexibility in their interpretation. Although for convenience, disease is frequently characterised as a binary distinction, such as healthy or AD patient, each subject in fact occupies a point on a continuous spectrum of disease severity, as is reflected by the concept of MCI. Probabilistic classification allows us to identify the position of subjects on this spectrum, enabling disease staging, stratification, or event based modelling (Fonteijn et al., 2012). Probabilistic decisions can also be made into a binary classification simply by thresholding, and our previous work shows that this method offers accuracy as good as an SVM on voxel level data for the diagnosis of AD (Young et al., 2012); hence no diagnostic information is lost by choosing a probabilistic classification algorithm. While an SVM's output can be interpreted probabilistically by transforming the decision value with a sigmoid function, this method is a rather ad hoc modification to a binary classifier, and does not offer the principled formulation and automatic parameter tuning of GP classification.

Our previous work is, to our knowledge, the only previous application of GP classification to AD. GPs have been used previously in a regression context with fMRI data by Marquand et al. (2009), and for classification of structural MRI data in Huntington's disease by Chu et al. (2010). They have not been previously applied for multimodal medical image classification. Here we use four types of data, and we compare two methods of setting the kernel weight, one very similar to that given by Zhang et al. (2011) and the other a probabilistic method that is more natural within the GP paradigm. We also compare our results to those obtained by an SVM on the same data, again using the method of Zhang et al. (2011) for setting kernel weights in the multikernel paradigm.

The training population comprises healthy controls and AD patients, allowing us to interpret the results in the MCI population as a risk score for conversion to AD. We introduce a new method for the validation of probabilistic predictions, which shows that the predicted probability of conversion is a good estimate of the actual chances

of conversion. As well as interpreting the results probabilistically, we also obtained a binary classification into MCI-s and MCI-c by adaptively thresholding the probabilities, resulting in a highly accurate prediction of conversion.

2. Materials and methods

All data were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database.[1] ADNI was launched in 2003 by the National Institute on Ageing (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the Food and Drug Administration (FDA), private pharmaceutical companies and non-profit organisations, as a $60 million, 5-year public/private partnership. The primary goal of ADNI has been to test whether serial MRI, PET, other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild MCI and early AD. Determination of sensitive and specific markers of very early AD progression is intended to aid researchers and clinicians to develop new treatments and monitor their effectiveness, as well as lessen the time and cost of clinical trials. The Principal Investigator of this initiative is Michael W. Weiner, MD, VA Medical Center and University of California at San Francisco. 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.

2.1. MRI data

Images were all T1 weighted structural MRI scans from 1.5 T scanners acquired using a 3D MPRAGE sequence, taken at the baseline time point for each subject. Back-to-back scans were taken for each subject, and the best scan of the pair for each subject determined by visual inspection. The images were then post-processed to correct for gradient warping, B1 non-uniformity and intensity non-uniformity and underwent phantom based scaling correction. Postprocessed images were downloaded as DICOM files, and were then converted to NIFTI format for further processing.

2.2. PET data

Images were again all taken from the baseline scan for each included subject. Images were acquired by scanning 30–60 min post injection using scanner-specific protocols. Six five minute frames are acquired for each subject, and then co-registered and averaged. The average images are then rigidly registered to a standard space, and the individual native space frames registered to the standard space average and averaged and intensity normalised in the standard space. Finally, the average images in the standard space are smoothed with a scanner-specific kernel (Joshi et al., 2009) to a uniform isotropic resolution of 8 mm FWHM, which is approximately the resolution of the lowest resolution scanners used in ADNI. The postprocessed scans were downloaded as DICOM images.

2.3. APOE data

Variants of the apolipoprotein E (APOE) genotype are known to affect the risk of developing sporadic AD in their carriers. Each individual has two copies of this gene, one inherited from each parent. The most common allele is APOE ε3, but carriers of the APOE ε4 variant are at heightened risk of AD, whereas the APOE ε2 variant confers some protection on carriers (Corder et al., 1993, 1994). The APOE genotype of each subject was recorded as a pair of numbers indicating which two alleles were present. APOE genotype is determined from a 10 ml blood sample taken at screening time, and sent overnight to the University of Pennsylvania AD Biomarker Fluid Bank Laboratory for analysis. APOE genotype was available for all subjects for which we had imaging data.

2.4. CSF data

CSF samples of 20 ml volume were obtained from subjects by a lumbar puncture with a 24 or 25 gauge atraumatic needle around the time of their baseline scan. All samples were sent on dry ice on the same day as they were drawn to the University of Pennsylvania AD Biomarker Fluid Bank Laboratory, where levels of the proteins ($a\beta_{42}$, total tau, and phosphorylated tau) were measured and recorded. By design, only a subset of ADNI subjects had measurement of CSF levels. All three measured protein levels (t-tau, p-tau, and $a\beta_{42}$) were used in constructing a CSF kernel.

2.5. Subjects

All ADNI subjects were between 55 and 90 years old, speak English or Spanish, and had a study partner able to provide an independent assessment of functioning. All subjects were willing to undergo neuroimaging and agreed to longitudinal follow up, and a subset was willing to undergo lumbar punctures. Subjects with specific psychoactive medication were excluded. Inclusion criteria for healthy controls (HC) are MMSE scores between 24 and 30, a CDR of 0, non-depressed and non-demented. Ages of the HC subjects were roughly matched to those of the AD and MCI subjects. For MCI subjects, the criteria are an MMSE score between 24 and 30, a memory complaint, objective memory loss measured by education adjusted scores on Wechsler Memory Scale Logical Memory II, a CDR of 0.5, absence of significant levels of impairment in other cognitive domains, essentially preserved activities of daily living, and an absence of dementia.

For AD subjects, the criteria are an MMSE score between 20 and 26, CDR of 0.5 or 1.0, and meeting NINCDS/ADRDA criteria for probable AD. Subjects are designated as HC, AD or MCI at the time of the baseline scan, and for the purposes of this study MCI conversion status is decided by whether subjects who were MCI at baseline were subsequently diagnosed as AD at any stage during the subsequent 36 month follow-up period.

A total of 682 subjects with baseline 1.5 T MRI scans were available. Of these, the image parcellation procedure was run on 679, the manually generated brain masks required for the parcellation being unavailable for three. Of these 679 subjects, FDG-PET scans were also available for 286. Seven of these were diagnosed as MCI at baseline but as healthy at follow-up time points and were excluded as reverters, leaving a total of 279 subjects available for the study. The demographics of this group (referred to as the PET group) are given in Table 1.

Table 1
Demographics of the PET group including 279 subjects. Disease status = diagnosis of AD or MCI at baseline, with MCI-s or MCI-c decided over 3 year follow-up, n = total number of subjects in group, n female = total number of female subjects in group, Mean age = mean age of group in years, Mean MMSE = mean MMSE score of group in years, SD = standard deviation of measurement.

Disease status	n (n female)	Mean age (SD)	Mean MMSE (SD)
Healthy	73 (27)	75.9 (4.6)	28.9 (1.2)
MCI-s	96 (34)	75.6 (7.0)	27.2 (1.7)
MCI-c	47 (17)	74.5 (7.4)	26.9 (1.8)
AD	63(24)	75.2(6.6)	23.6 (2.0)

Table 2
Demographics of the PET-CSF group including 143 subjects. Disease status = diagnosis of AD or MCI at baseline, with MCI-s or MCI-c decided over 3 year follow-up, n = total number of subjects in group, n female = total number of female subjects in group, Mean age = mean age of group in years, Mean MMSE = mean MMSE score of group in years, SD = standard deviation of measurement.

Disease status	n (n female)	Mean age (SD)	Mean MMSE (SD)
Healthy	36 (12)	74.2 (4.2)	28.8 (1.3)
MCI-s	42 (16)	75.4 (7.0)	27.3 (1.6)
MCI-c	30 (11)	75.5 (7.6)	26.5 (1.8)
AD	35 (12)	75.2 (6.7)	23.9 (2.0)

We also examined the effect of using CSF in the multimodal classification. As there was relatively little overlap between the groups of patients given CSF biomarker testing as well as FDG-PET scans, the subset of the PET group for which full CSF data was also available (referred to as the PET-CSF group) was much smaller at a total 143 subjects. The demographics of the PET-CSF group are given in Table 2.

In the PET group, 47 out of 143 (33%) of MCI subjects are converters. As conversion is defined over a 3 year follow-up period, this is equivalent to an annualised conversion rate of 12.5% per year, in line with other studies. Subjects diagnosed as MCI at baseline in ADNI are reassessed after approximately 6, 12, 18, 24 and 36 months which allow us to roughly find the time after which they converted. The conversion times for the 47 MCI-c subjects in the PET group are listed in Table 3.

2.6. MRI image processing

To produce grey matter (GM) probability maps in a common space for classification, we follow roughly the same procedure as Klöppel et al. (2008). However we use different image processing software, and also add a step of masking the images to include only regions known to be affected by AD.

2.6.1. Image segmentation

The native space, preprocessed scans were probabilistically segmented using the open source NiftySeg tool (Cardoso et al., 2011). Based on the expectation maximisation algorithm, this method produces probabilistic maps for five tissue types: white matter, cortical GM, external cerebrospinal fluid, deep GM and internal cerebrospinal fluid.

2.6.2. Image parcellation

The native space, preprocessed scans were also anatomically parcellated into 83 regions. This was with a multi-atlas segmentation propagation algorithm (Cardoso et al., 2012). A library of 30 atlases manually labelled with 83 anatomical regions (Gousias et al., 2008) was used as a basis for the segmentations. In order to segment a new image, all the atlases and respective manual labels were first nonrigidly registered to this image. After registration, the manual labels of the locally most similar atlases were fused using a label fusion strategy based on an extension of the STAPLE algorithm (Warfield et al., 2004) to produce a final parcellation. The regions used in the classification process were chosen according to Braak and Braak (1995) and are listed in Table 4. These regions were then intersected with the GM tissue segmentations obtained above.

Table 3
Conversion times of MCI-c subjects in the PET group. Conversion time = time after baseline scan when the subject was first diagnosed as AD, n = total number of subjects in group.

Conversion time t (months)	n
$t < 6$	5
$6 < t < 12$	15
$12 < t < 18$	9
$18 < t < 24$	14
$24 < t < 36$	4

Table 4
Selected regions for classification.

Label numbers	Regions
1, 2	Hippocampus (R and L)
3, 4	Amygdala (R and L)
5, 6	Anterior temporal lobe, medial part (R and L)
7, 8	Anterior temporal lobe, lateral part (R and L)
9, 10	Parahippocampal and ambient gyri (R and L)
11, 12	Superior temporal gyrus, posterior part (R and L)
13, 14	Middle and inferior temporal gyrus (R and L)
15, 16	Fusiform gyrus (R and L)
24, 25	Cingulate gyrus, anterior part (R and L)
26, 27	Cingulate gyrus, posterior part (R and L)

2.6.3. Image registration

All images were transformed into the same anatomical space in order to provide consistent anatomy at each voxel for the classifier. The images were masked to remove non-brain material, and then used to perform groupwise registration. All images were repeatedly registered to the same target image in an iterative procedure. At the end of each iteration, all registered images were averaged together to create an updated target image, with a randomly chosen image serving as the target in the first iteration. Initially, all images were rigidly registered to avoid bias towards the chosen target. This was followed by a single round of affine registration, and then by 10 rounds of nonrigid registrations. All registrations were performed using NiftyReg (Modat et al., 2010), a registration toolkit that performs fast diffeomorphic nonrigid registrations. When the registrations had all been completed, the resulting deformations from each image's native space to the final average image were applied to the anatomically masked native space segmentations to bring them into the groupwise space. The registered, anatomically masked segmentations were modulated by the Jacobian determinants of this final deformation. This ensures that the total volume of tissue remains constant (Ashburner and Friston, 2000). As a final step, the registered, anatomically masked and Jacobian modulated cortical GM and deep GM segmentations were summed to produce an overall GM density map for the AD relevant regions in a common space for all subjects.

2.7. PET image processing

The PET images had already been through substantial postprocessing, as discussed above. After downloading, they were registered to the native space MRI image of the same subject, again using the NiftyReg software. Then masks generated from the structural MRI parcellations were overlaid on each subject to calculate the total activity within each region from the PET image. The mean activity within each region was then used as a feature for classification.

2.8. Gaussian processes

The resulting high dimensional image and biomarker data were then used to construct a GP classifier based on HC and AD subjects. A full description of GP classification is beyond the scope of this study. We refer readers to Rasmussen and Williams (2006) for a detailed theoretical treatment, and to our earlier work (Young et al., 2012) for an example of their application to AD image classification. Instead, we give a brief description of GP classification and give further details on the aspects that pertain to multimodal classification.

All learning of hyperparameters and GP calculations were done using the GPML toolbox[2] for MATLAB[3] which was also used to analyze results.

2.8.1. Gaussian process classification

Gaussian process classification can be seen as kernelised Bayesian extension of logistic regression. A Gaussian process, essentially a multivariate Gaussian, forms the prior on the value of a latent function, which is then mapped to the (0,1) interval through a sigmoid, which represents the probability of a subject belonging to a particular class. The exact prior is a function of the training data and labels, and a set of hyperparameters that control the shape of the prior. During the training phase, the hyperparameters are learned from the training data and labels by type-II maximum likelihood. The likelihood of the training data and labels with respect to the hyperparameters is maximised using a conjugate gradient descent optimisation method. Once the hyperparameters have been set, predictions on unseen data are made by integrating across this prior. In the regression case, this is analytically tractable, but for classification it is not, due to the sigmoidal response function, so an approximation must be made instead. A number of different approximation schemes can be used, but all our experiments use the expectation propagation algorithm (Minka, 2001). This has been shown to offer a good compromise of accuracy and computation time for GP classification (Nickisch and Rasmussen, 2008).

2.8.2. Gaussian processes as multimodal kernel methods

Note that the GP classifier is based on a kernel K representing the covariance among training subjects. This is a symmetric positive definite matrix where entry (i,j) is a covariance or some function of distance between training subjects i and j. As such, this means that GP classification belongs to the family of kernel methods as do SVMs, and all the rules for constructing valid kernels apply: in particular, a positive sum of valid kernels is a valid kernel, and a valid kernel multiplied by a positive scalar is also a valid kernel. The covariance between the *i*th and *j*th subject, K_{ij}, is a kernel function k of the feature vectors for the *i*th and *j*th subject x_i and x_j and a hyperparameter or hyperparameters θ. We use a linear kernel function, which is the scalar product of x_i and x_j plus a single hyperparameter representing a bias term: $K_{i,j} = x_i \cdot x_j + \theta$. The hyperparameter is learnt from the training data by type-II maximum likelihood. For multimodal classification, the rules for producing new kernel mean that we can define our kernel function as the weighted sum of a number of subkernels, each of which has been calculated from a the feature vectors representing a particular type of data or modality for each subject. Each subkernel has a scaling hyperparameter representing the modality's weight in the overall kernel, which also includes a single bias term. So in the case of multimodal classification using each subject's MRI, PET and APOE data the overall kernel is

$$K_{i,j} = \alpha_{MRI}\left(x_{MRI,i} \cdot x_{MRI,j}\right) + \alpha_{PET}\left(x_{PET,i} \cdot x_{PET,j}\right) + \alpha_{APOE}\left(x_{APOE,i} \cdot x_{APOE,j}\right) + \beta$$

where α are hyperparameters representing the weight given to each modality subkernel, and β is a hyperparameter representing the bias in the combined kernel. Thus θ is now a set of four hyperparameters which are again learnt from the training data by maximum likelihood. In this way we can automatically set the kernel weights without needing to resort to a grid search with cross validation. This is possible as the GPML software allows complex covariance functions to be specified. It allows us to apply masks to include only certain columns of the training data to be used in a covariance function, so we can learn separate covariance kernels for the MRI, PET and APOE data. The APOE kernel is based on representing each subject as a vector of length two, encoding each allele as an element of the vector, so for a example a subject with one copy of the ε3 allele and one of the ε4 would be encoded as (3 4). More sophisticated kernels have been developed for genetic data and we plan to exploit these in future.

For the PET group, we also do a grid search for the kernel weights to compare the results of this method of setting the kernel weights to the maximum likelihood method and to Zhang et al. (2011). Each MCI test subject in turn is left out, and a GP classifier is trained on all AD and control training subjects for each legitimate combination of α. The best values of α are chosen empirically as the ones offering the most accurate classification on the n-1 remaining MCI test subjects. As accuracy is a coarse measure, any ties are broken with the information theory based metric of classification quality suggested by Rasmussen and Williams (2006). Finally the classifier offering the best accuracy was tested on the left out MCI subject, and the process repeated until all MCI test subjects had been classified. Due to the leave-one-out loop and the need to do one tuning classification for every combination of parameters within each iteration of the loop, this method is very time consuming if more than a handful of parameters have to be tuned. Hence to make the whole classification procedure tractable, values of α are constrained to be positive and sum to one, with no bias term, as in Zhang et al. (2011). The resulting two-dimensional parameter space is searched with increments of 0.1 for both parameters.

Fig. 1 represents the multikernel approach.

2.9. SVM classification

To put the results obtained by GP classification in context and compare them to a more widely used method we also performed SVM classification on the same datasets. We made use of the open source libsvm library,[4] with the C parameter left at its default setting and linear kernels, but used precomputed kernels both for the sake of speed and to facilitate multikernel classification. Training and testing kernels were constructed for all three modalities in the PET group (MRI, PET and APOE) and all four in the PET-CSF group (MRI, PET, APOE and CSF). Kernel weights are again set using the method of Zhang et al. (2011) as described in the previous section. The weight setting is done within a leave one out scheme, where the testing (MCI-s and MCI-c) subjects are repeatedly split into one subject used for testing and the remaining ones used for tuning the kernel weights until each MCI subject has been left out; in this way it is possible to tune on the testing population without introducing optimistic bias (Kriegeskorte et al., 2009). We also tried to set the kernel weights using the training (NC and AD) subjects for tuning, by performing a leave-one-out cross validation on the training subjects at each legitimate combination of kernel weights. To break ties between parameter settings showing equal accuracy, we use the mean distance from the margin of correctly classified test subjects minus the mean distance from the margin of incorrectly classified test subjects as a metric of SVM quality. We also experimented with normalising training and testing data using a z-score to help combine different modalities on the same scale.

2.10. Classification strategy

Rather than both training and testing the classifier on MCI-s and MCI-c subjects in a cross-validation loop, we train on AD and healthy subjects, and then obtain results by applying the resulting classifier to the MCI population. This approach to classification of MCI subjects is widely used and was adopted here as it obtained substantially better results than those obtained by the training on MCI regime in all our preliminary work. The hypothesis justifying this is that the subpopulation of MCI subjects that are stable are more healthy-like (although some will go on to convert beyond the follow up period used for

[2] http://www.gaussianprocess.org/gpmlcode/matlab/doc.
[3] MathWorks Inc., Natick, MA, 2011.
[4] http://www.csie.ntu.edu.tw/~cjlin/libsvm/.

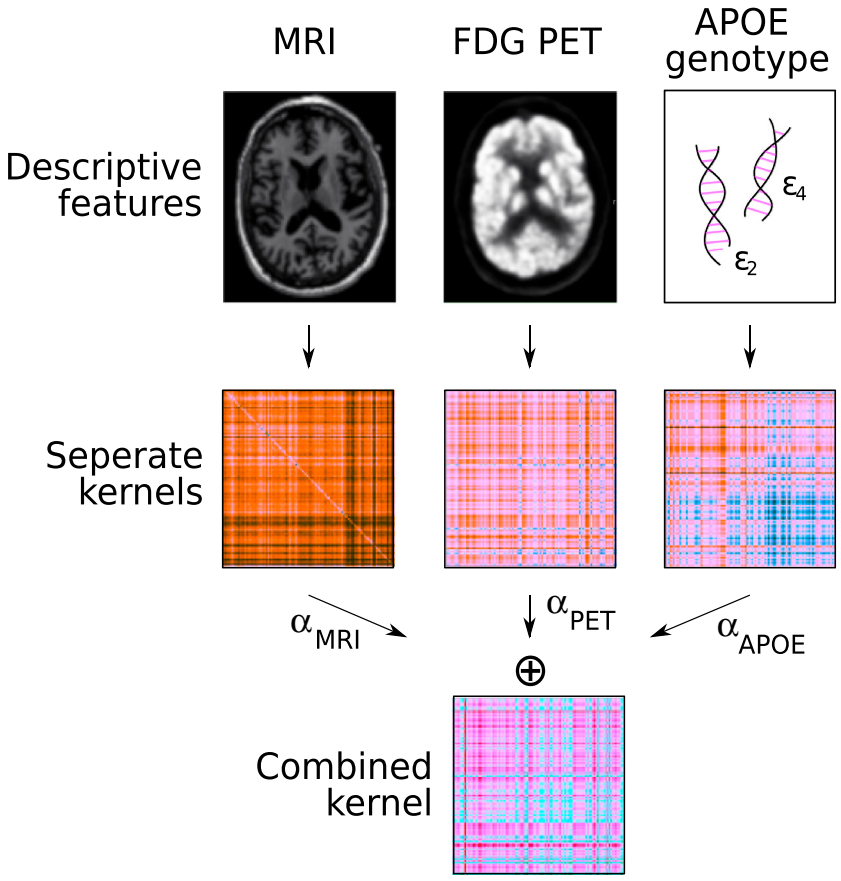

Fig. 1. Pipeline by which kernels are constructed from features extracted from each type of data, before being summed to produce a combined kernel.

defining conversion, which is probably a factor in the limited accuracy of predictions of MCI conversion), while those who go on to develop dementia are more AD-like, as is consistent with our contention that discrete disease states are an approximation to a continuous disease spectrum.

This means a classifier that successfully separates AD and control subjects will also be able to distinguish between MCI-c and MCI-s to some degree. This notion, illustrated in Fig. 2, has been used with some success for this problem previously (Ferrarini et al., 2009; Singh et al., 2012). As previously mentioned, however, when using a combined kernel with grid search we can use the MCI subjects not being classified to tune the kernel mixing parameters.

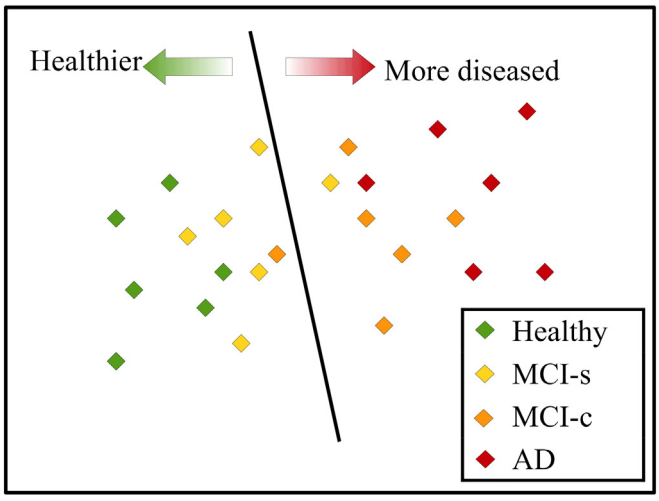

Fig. 2. Relation between AD and MCI classification.

2.11. Validation

The results of GP classification are numbers between 0 and 1 representing the estimated probability that a test subject belongs to a particular class, in our case the class of MCI-c. A simple way to binarise these probabilities is to threshold them at 0.5. We do this, and report the resulting accuracy, sensitivity and specificity. However this approach has two disadvantages. Firstly, as the model is trained on one population (AD and control) and tested on another (MCI-s and MCI-c), this would be the correct threshold value if the test population were in some sense exactly half way between the two classes of the training population, but there is no reason to believe this is necessarily the case. Secondly, setting the cut point at 0.5 leads to varying balances between sensitivity and specificity among the different methods, making them hard to compare. Because of this, we also use the test probabilities to determine the cut point that results in the closest possible value of sensitivity and specificity. We then determine the overall correct classification rate at this cut point and report only that, as by definition it will be very close to both the sensitivity and specificity. This is done in a leave-one-out framework to avoid optimistic bias in the balanced accuracy. We also use the probabilities to calculate the area under the receiver operating characteristics (ROC) curve, known as the AUC, for easy comparison with results from other studies. For both PET and PET-CSF groups we also report the balanced accuracy for classification using each modality alone, except for the APOE. This is left out because APOE data consists of pairs of alleles labelled 2, 3 or 4. As order does not matter this means each subject can be at one of only 6 points in two-dimensional APOE data

space (in practice 5 points as one combination does not occur in our data), so an APOE only classifier would produce probabilities that could only be one of five discrete values, making further analysis meaningless. We assess the significance of the difference in balanced accuracy between multimodal classification and unimodal classification for both the PET group and PET-CSF group with McNemar's test (McNemar, 1947) if there appears to be a substantial difference. The balanced accuracies are derived from the probabilities before they are corrected for bias with the procedure described in Section 3.2. We found that balanced binary accuracies derived from the corrected probabilities tended to be slightly lower.

However, to only do this would be to neglect the probabilistic information contained in the output of the GP. We can also treat the probabilities as risk scores for conversion to AD, and determine how well they function as estimates of the actual chances of conversion. As each subject either does or does not convert to AD, this cannot be assessed at the individual level. We instead bin all MCI subjects into eight equal intervals covering the range (0,1) by their risk score. For each of the eight intervals, the centre value of the interval is labelled the predicted risk. We then calculate the empirical risk for each interval as the proportion of patients in the interval that do in fact convert. Finally, the root mean square error between predicted and empirical risk is calculated as a measure of how well the risk scores from GP classification predict the actual risk of conversion. The number of intervals was chosen to provide the best balance between the demands for good statistics both within and between the bins.

The decision values obtained from SVM classification represent a signed distance from the optimal hyperplane determined from the training data, the sign indicating on which side of the hyperplane a test subject falls and thus to which class it is predicted to belong. We report the accuracy, sensitivity and specificity from the sign of the decision values (equivalent to thresholding the decision values at 0). We also perform a procedure to find the threshold producing the accuracy that best balances sensitivity and specificity in the same manner as we did for GP posterior probabilities, and finally calculate an AUC from the decision values.

3. Results

3.1. Accuracy of binary classification

The balanced accuracy, AUC, and p-value for comparison of multimodal methods with unimodal ones for the PET group are shown in Table 5 for the GP results, and in Table 6 for the SVM results.

The result in the last row of Table 6 was obtained by using the MCI subjects as a tuning set, and normalising training data with a z-score, and then normalising testing data using the mean and standard deviations from the un-normalised training data. All other combinations of choices of tuning set and normalisation produced inferior results.

The same accuracy measures for the PET-CSF group are shown in Table 7 for GP classification and Table 8 for SVM. For the GP, we do not perform the grid search method due to the increased computational demands of having to do a three dimensional grid search for four modalities, rather than a two dimensional grid search for three modalities as in the previous experiment. We do, however, report the results for multimodal classification both with and without the CSF data so it is possible to see its effect on classification with a consistent set of test subjects.

Table 5
Accuracy of methods on the PET group with GP classification. Acc = accuracy, sens = specificity, spec = specificity, balanced acc = balanced accuracy, AUC = area under the ROC curve, and p = significance of improvement in classification vs. indicated single modality.

Modality	acc	sens	spec	Balanced acc	AUC	p vs. MRI	p vs. PET
MRI	64.3%	53.2%	69.8%	61.5%	0.643	–	–
PET	65.0%	66.0%	64.6%	65.7%	0.767	–	–
MRI + PET + APOE (ML)	69.9%	78.7%	65.6%	74.1%	0.795	0.0162	0.0247
MRI + PET + APOE (GS)	67.1%	76.6%	62.5%	70.6%	0.751	0.0865	0.2301

Table 6
Accuracy of methods on the PET group with SVM classification. Acc = accuracy, sens = specificity, spec = specificity, balanced acc = balanced accuracy, and AUC = area under the ROC curve.

Modality	acc	sens	spec	Balanced acc	AUC
MRI	58.7%	53.2%	61.5%	58.7%	0.629
PET	69.9%	55.3%	77.1%	67.1%	0.762
MRI + PET + APOE (GS)	65.7%	68.1%	64.6%	67.8%	0.731

Again, the last two rows of Table 8 present results obtained using MCI subjects for tuning the kernel weights, and with the data normalised with a z-score as these provided the best accuracy.

The results show a clear advantage in accuracy for multimodal imaging. In the larger PET group, both multimodal algorithms are better than any single modality alone. This advantage is statistically significant at the 5% level for the type-II maximum likelihood method with GP classification, which outperforms the grid search method and outperforms the best single modality by over 8%. The AUC measure of accuracy shows how results must be interpreted with caution, as the multimodal grid search method has a higher balanced accuracy than using PET alone, but offers a slightly lower AUC. In the smaller group for which both PET and CSF data were available in all subjects, the same pattern applied in that multimodal methods outperformed all single modality methods.

To enable a side-by-side comparison, Table 9 shows the balanced accuracy for GP and SVM classification together with a p-value for the difference in accuracy. The p-value is generated by classifying all test subjects with the leave-one-out procedure used to generate the balanced accuracy figures, and comparing the resulting classifications, again using McNemar's test.

3.2. Accuracy of probabilistic classification

The predicted risk figures produced in the manner described in Section 2.11 exhibit some bias, in that the classifiers tend to overestimate the chances of conversion in general. This appears to be because of the transfer learning approach we use, where the classifier is trained on the AD and healthy population, and then applied to the MCI subjects. As the MCI subjects, in terms of the biomarker data we use, are not midway between the AD and control population but slightly closer to the AD subjects, this results in the classifier being somewhat biased in favour of predicting conversion. In order to remove this, we perform a correction procedure on the GP probabilities similar in approach to the one used to produce a balanced accuracy. We perform a logistic regression, using a leave-one-out approach again to avoid unduly optimistic results, on the GP probabilities and the labels indicating converter or stable status for the MCI subjects, with the label 0 indicating stable and 1 indicating converter. In this way we can learn the relationship between GP predicted risk and actual risk for the MCI subjects to correct for the bias. The resulting plots of empirical risk versus adjusted predicted risk for the PET and PET-CSF groups are shown in Figs. 3 and 4. Plotted points are labelled with the number of subjects in the corresponding bin. As not all the bins contain subjects, points for these bins are not included.

In these plots, a classifier producing accurate probabilities should have points plotted close to the diagonal. By inspection, the multimodal methods appear to perform well by this measure, and it is important to note that most points lying far away from the diagonal represent bins containing few subjects, making the empirical risk calculated for them less reliable. More broadly, the probabilities produced by GP classification procedure appear to be accurate in the

Table 7
Accuracy of methods on the PET-CSF group with GP classification. Acc = accuracy, sens = specificity, spec = specificity, balanced acc = balanced accuracy, AUC = area under the ROC curve, and p = significance of improvement in classification vs. indicated single modality.

Modality	acc	sens	spec	Balanced acc	AUC	p vs. MRI	p vs. PET	p vs. CSF
MRI	63.9%	76.7%	54.8%	61.1%	0.682	–	–	–
PET	66.7%	80.0%	57.1%	69.4%	0.789	–	–	–
CSF	55.6%	73.3%	42.9%	56.9%	0.575	–	–	–
MRI + PET + APOE (ML)	68.1%	83.3%	57.1%	72.2%	0.823	0.1860	0.7728	0.0725
MRI + PET + APOE + CSF (ML)	68.1%	90.0%	52.4%	72.2%	0.763	0.2012	0.8231	0.0153

Table 8
Accuracy of methods on the PET-CSF group with SVM classification. Acc = accuracy, sens = specificity, spec = specificity, balanced acc = balanced accuracy, and AUC = area under the ROC curve.

Modality	acc	sens	spec	Balanced acc	AUC
MRI	65.3%	76.7%	57.1%	63.9%	0.685
PET	69.4%	63.3%	73.8%	65.3%	0.782
CSF	56.9%	73.3%	45.2%	55.6%	0.575
MRI + PET + APOE (GS)	68.1%	76.7%	61.9%	68.1%	0.745
MRI + PET + APOE + CSF (GS)	66.7%	76.7%	59.5%	69.4%	0.727

sense that increased predicted risk of conversion does generally imply an increased chance of conversion actually taking place. The adjustment appears to be effective, with little bias exhibited in the predicted risks. Note that the only points which are plotted very far from the diagonal, and thus show a large difference between empirical and predicted risk, are of risk bins containing only 1 or 2 subjects and are simply the results of outliers.

4. Discussion

As previously stated, a clear advantage can be seen both for multimodal classification, and for the use of GP classification over the more widely used SVM. This applied to results for both the PET and PET-CSF groups. Moreover, there appears to be quite a strong interaction between the utility of multimodal classification and the type of classifier used. Looking at the balanced accuracy of classification on single modalities of data, there is little to choose between GP and SVM classification, with differences of one or two per cent in accuracy in either direction. Thus, it seems reasonable to conclude by this measure that there is little difference in discriminative ability on identical sets of data. However, the GP framework appears to be able to take much greater advantage of the availability of multimodal data. GPs offer much larger gains for multimodal versus unimodal classification, with gains of 8% in the PET group against the best single data (PET) as against only a 0.7% gain for the SVM approach. Similarly, the head-to-head comparisons between the GP and SVM methods using the same subjects and modalities, in Table 9, show the greatest differences in classification accuracy and greatest statistical significance are for the multimodal methods. While the difference is not quite significant at the 0.05 level, due to the relatively small number of subjects in the study, the advantage for GP against SVM classification is clear and consistent across all three multimodal classification experiments and we plan to verify it with a larger dataset.

Table 9
Statistical comparison of GP and SVM classification results.

Group	Modality	Balanced accuracy (GP)	Balanced accuracy (SVM)	p-Value for accuracy of GP vs. SVM
PET	MRI	61.5%	58.7%	0.3865
PET	PET	65.7%	67.1%	0.7893
PET	MRI + PET + APOE	74.1%	67.8%	0.1508
PET-CSF	MRI	61.1%	63.9%	0.6831
PET-CSF	PET	69.4%	65.3%	0.4497
PET-CSF	CSF	56.9%	55.6%	1
PET-CSF	MRI + PET + APOE	72.2%	68.1%	0.4497
PET-CSF	MRI + PET + APOE + CSF	72.2%	69.4%	0.8026

The improvement is most likely because the GP framework is better at finding a set of kernel weights for optimum classification. With an SVM we are restricted to finding these through a grid search, which has an inherently limited range and resolution if it is to be tractable, and is dependent on rather crude measures of accuracy to select an optimal parameter set. GPs offer tuning via the likelihood function, which seems to be both more robust and allows a wider search space — however this is not available for SVM classification, highlighting one of the advantages of a probabilistic framework mentioned in the introduction.

Adding CSF to multimodal classification did not increase the accuracy by any significant amount and in fact decreased the AUC, which is not surprising as CSF is the poorest single modality, offering accuracy little better than chance. The poor performance of CSF biomarkers alone, and their failure to add diagnostic value when used alongside other biomarkers, is perhaps explained by the fact that about a third of controls have a high amyloid load, suggesting they may be in fact at a presymptomatic stage of AD. In this case CSF is still a potentially valuable biomarker, but our choice of defining AD and control subjects purely by current symptoms and cognitive test results limits its applicability. This again suggests the need to treat AD as a spectrum rather than a set of discrete states, or at least to very carefully define such states.

Comparing the results presented here to other attempts to predict conversion in MCI patients is difficult. This is because, while the problem has been addressed in a large number of studies, these vary

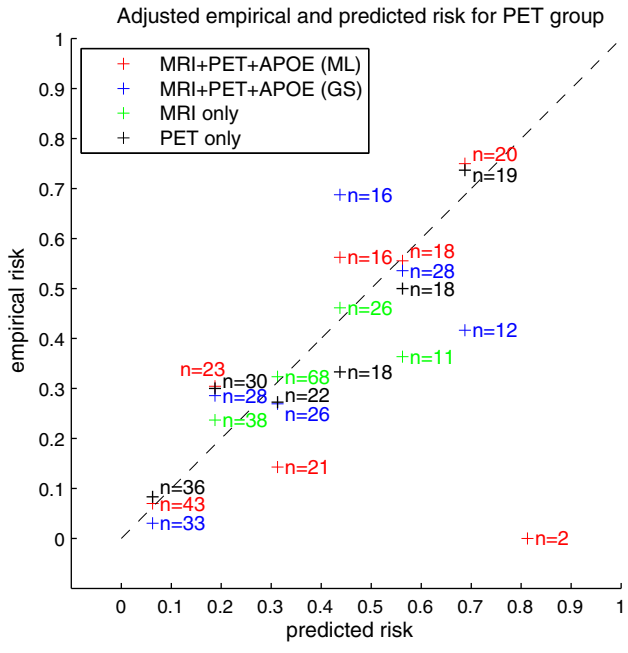

Fig. 3. Empirical risk vs. corrected predicted risk for the PET group.

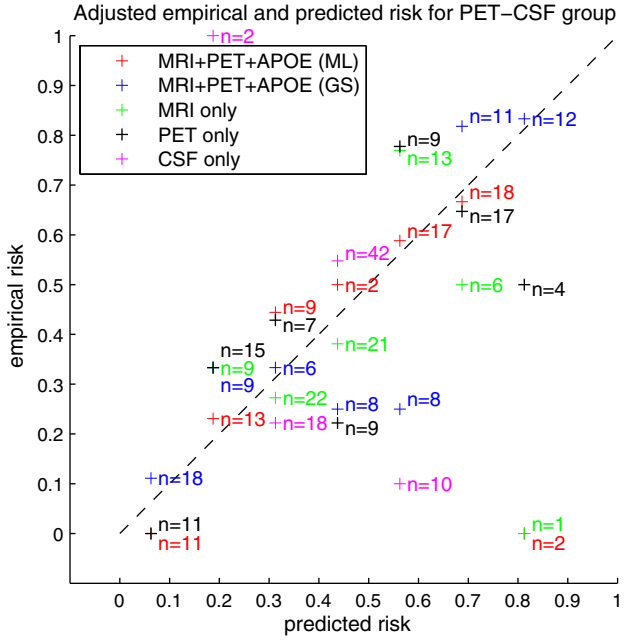

Fig. 4. Empirical risk vs. corrected predicted risk for the PET-CSF group.

widely in how MCI groups are defined, and the metric by which classification accuracy is assessed. However the method presented here certainly offers a high level of classification accuracy, especially considering studies that use ADNI data and offer higher accuracy make predictions over a time span of less than three years or make use of longitudinal data, which our algorithm does not need.

For MRI data, the most comparable methods are in Cuingnet et al. (2010). This study included a wide variety of types of feature, but those which used voxel level GM maps are quite similar to our work. Even within this definition, a wide range of options in image processing and feature extraction were used but the closest in methodology to ours is what they label as the Voxel-Direct-D-GM method. When applied to predicting MCI conversion this was found to have a specificity of 100% and sensitivity of 0%, i.e. The classifier simply assigned all subjects to the majority MCI-s class, possibly as a function of having trained on MCI-s and MCI-c rather than control and AD subjects. This paper did also find that the voxel based method in Fan et al. (2008) achieved a sensitivity of 62% and specificity of 67%, although this was found not to be significantly greater than chance. Our method achieves much greater accuracy than any in Cuingnet et al. (2010) for predicting MCI conversion, and moreover our accuracy is statistically significantly better than chance, which none of the methods assessed that study managed to achieve.

Other studies, however, have had much greater success in predicting conversion. For example, Coupé et al. (2012) and Eskildsen et al. (2013) have presented methods capable of predicting conversion with accuracies similar to ours. Coupé et al. (2012) uses a novel hippocampal grading biomarker. Using their most rigorous validation method, accuracy was slightly lower at 71% but their method needs no FDG-PET data and less computationally intensive image processing than the one presented here. Eskildsen et al. (2013) also achieves 74% accuracy by stratifying MCI-C subjects by conversion time and then combining the results of classifying each MCI-C subgroup against the MCI-s subjects. The classifier is rather unbalanced, with substantially higher specificity than sensitivity, a common problem with MCI classification, but again only structural image data is needed. Ye et al. (2012) report AUC values of up to 0.85 using MRI data, APOE genotypes, and a variety of cognitive measures with a sparse logistic regression procedure but do not list classification accuracy. Wee et al. (2012) use features based on correlations between mean thicknesses of cortical regions of interest with SVM classification, and obtain 75% accuracy and an AUC of 0.8426.

Among multimodal methods, Zhang et al. (2011) reports a specificity of 91.5% and a specificity of 73.4% for prediction of MCI conversion. While they do not report the proportions of MCI-s and MCI-c in their subjects and hence we cannot calculate the overall accuracy, it must be greater than our best result of 74%. However, they define conversion as a subject converting within 18 months rather than three years. Predicting over a short future timespan is an easier problem than over a longer one (Eskildsen et al., 2013) and less clinically useful. Moreover, defining conversion over a shorter time means a smaller proportion of MCI subjects will be converters, reducing the positive predictive value of the classification result. Additionally, their work uses CSF data in addition to MRI and FDG-PET, whereas our best performing classifier uses genetic data instead of CSF, which is easier and less invasively obtained. We are able to set our kernel weights by type-II maximum likelihood, avoiding the need for a computationally expensive grid search. The other previously published multikernel method to predict MCI conversion is by Hinrichs et al. (2011). Although they do define converters with a three year time span, direct comparison of results is again difficult, as they report only an AUC rather than accuracy. The best reported AUC was 0.791, similar to ours but this used longitudinal data, again effectively reducing the time span to predict conversion. They also found the method using only longitudinal image data was more effective than including non-imaging data in their multikernel learning approach. Methods based on features structural imaging alone are also capable of achieving high accuracy.

Table 10 summarises these results in comparison with our own.

Table 10 clearly shows the difficulty in making direct comparisons between results. For example, the time within which MCI conversion is defined has a strong effect on results. Vounou et al. (2012) used tensor based morphometry to define a set of voxels that are highly indicative of MCI conversion, and then applied an SVM to these. This

Table 10
Reported results from a variety of studies for predicting MCI conversion on ADNI data. n = number of subjects, conversion period = length of time over which MCI conversion is defined, acc = accuracy in predicting conversion, if reported, and AUC = area under ROC curve of predictions of conversion, if reported.

Article	Data used	n (MCI-s, MCI-c)	Conversion period	acc	AUC
Young et al.	MRI, FDG-PET, APOE	143 (96, 47)	0–36 months	74.1%	0.795
Eskildsen et al.	MRI	388 (227, 161)	0–36 months	73.5%	–
Ye et al.	MRI, APOE, cognitive scores	319 (177, 142)	0–48 months	–	0.8587
Wee et al.	MRI	200 (111, 89)	0–36 months	75.05%	0.8426
Zhang et al.	MRI, FDG-PET,CSF	99 (56, 43)	0–18 months	sens 91.5%, spec 73.4%	–
Hinrichs et al.	Longitudinal/baseline MRI, longitudinal/baseline FDG-PET, CSF, APOE, cognitive scores	119	0–36 months	–	0.7911
Coupé et al.	MRI	405 (238, 167)	0–36 months	73%	-
Wolz et al.	MRI	405 (238, 167)	0–36 months	68%	–
Nho et al.	MRI, APOE, family history	355 (205, 150)	0–36 months	71.6%	–
Davatzikos et al.	MRI, CSF	239 (170, 69)	0–36 months	61.7%	0.734

method was able to predict conversion with an accuracy of 82%. As this method uses both baseline MRI scans and 24 month follow-up MRI scans to generate Jacobian maps, it is effectively predicting conversion in only a 12 month period rather than three years as we do, and longitudinal data may not be available in all cases.

Parameterisations of the shape of the hippocampus have achieved a greater accuracy than our approach with conversion defined over three years (Costafreda et al., 2011; Ferrarini et al., 2009), however these used a small number of subjects scanned at a single centre, and also had autopsy confirmed AD subjects available, removing any uncertainty in the training labels. If conversion is defined over a three year period, we believe our method presented here has obtained an accuracy very competitive with the best methods yet published for prediction of conversion to date on ADNI data.

Moreover, our method offers the advantages of probabilistic classification listed in Section 1. The reject option is especially relevant in the case of computer-aided diagnosis. Having a probabilistic classification means that each diagnosis includes an attached degree of confidence rather than a simple binary decision. Clinical decision making is frequently hampered by overconfidence (Berner and Graber, 2008), so an estimate of the certainty of a diagnosis could be of great help, if only as a supplement to decisions made by more traditional methods.

5. Conclusion

We have shown that multimodal Gaussian process classifiers can be successfully applied to the prediction of conversion to AD in MCI patients. Prediction of conversion is significantly better for multimodal classification than for any single modality, and also better for GP compared to SVM classification, largely due to the GP's superior ability to exploit multimodal data. Accuracy is state-of-the-art, and to this we can add the advantages of probabilistic classification. In the future, we plan to take advantage of new subjects with FDG-PET and CSF data being added to the ADNI database and apply these methods to a larger group of subjects to show greater statistical significance for the advantage of our methods. We will perform more sophisticated feature extraction on FDG-PET data and to make use of more complex kernel covariance functions. We also plan to investigate other promising biomarkers such as hippocampal shape and cortical thickness, and will examine methods to overcome the problem of misdiagnosis leading to noisy training labels in ADNI data.

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.nicl.2013.05.004.

Acknowledgements

We wish to thank Professor John Ashburner for his help and guidance in the use of Gaussian process classification for neuroimaging data.

Sebastien Ourselin and Marc Modat received funding from the CBRC Strategic Investment Award (Ref. 168).

Sebastien Ourselin and Jorge Cardoso were also supported by EPSRC (EP/H046410/1).

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: Alzheimer's Association; Alzheimer's Drug Discovery Foundation; BioClinica, Inc.; Biogen Idec Inc.; Bristol-Myers Squibb Company; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; F. Hoffmann-La Roche 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.; NeuroRx Research; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; 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 Neuro Imaging at the University of California, Los Angeles. This research was also supported by NIH grants P30 AG010129 and K01 AG030514.

References

Ashburner, J., Friston, K.J., 2000. Voxel-based morphometry — the methods. NeuroImage 11 (6), 805–821 (Jun.).

Bach, F.R., Lanckriet, G.R.G., Jordan, M.I., 2004. Multiple kernel learning, conic duality, and the SMO algorithm. Proceedings of the twenty-first international conference on Machine learning, New York, NY, USA, pp. 41–48.

Barnes, J., Scahill, R.I., Boyes, R.G., Frost, C., Lewis, E.B., Rossor, C.L., Rossor, M.N., Fox, N.C., 2004. Differentiating AD from aging using semiautomated measurement of hippocampal atrophy rates. NeuroImage 23 (2), 574–581 (Oct.).

Beach, T.G., Monsell, S.E., Phillips, L.E., Kukull, W., 2012. Accuracy of the clinical diagnosis of Alzheimer disease at National Institute on Aging Alzheimer Disease Centers, 2005–2010. Journal of Neuropathology and Experimental Neurology 71 (4), 266–273 (Apr.).

Berner, E.S., Graber, M.L., 2008. Overconfidence as a cause of diagnostic error in medicine. American Journal of Medicine 121 (5 Suppl.), S2–S23 (May).

Bishop, C.M., 2007. Pattern Recognition and Machine Learning. Springer.

Braak, H., Braak, E., 1995. Staging of Alzheimer's disease-related neurofibrillary changes. Neurobiology of Aging 16 (3), 271–278 (May).

Cardoso, M.J., Clarkson, M.J., Ridgway, G.R., Modat, M., Fox, N.C., Ourselin, S., 2011. LoAd: a locally adaptive cortical segmentation algorithm. NeuroImage 56, 1386–1397 (Jun.).

Cardoso, M.J., Modat, M., Ourselin, S., Keihaninejad, S., Cash, D., 2012. Multi-STEPS: multi-label similarity and truth estimation for propagated segmentations. 2012 IEEE Workshop on Mathematical Methods in Biomedical Image Analysis (MMBIA), pp. 153–158.

Chu, C., Bandettini, P., Ashburner, J., Marquand, A., Klöppel, S., 2010. Classification of neurodegenerative diseases using Gaussian process classification with automatic feature determination. 2010 First Workshop on Brain Decoding: Pattern Recognition Challenges in Neuroimaging (WBD), pp. 17–20.

Corder, E.H., Saunders, A.M., Strittmatter, W.J., Schmechel, D.E., Gaskell, P.C., Small, G.W., Roses, A.D., Haines, J.L., Pericak-Vance, M.A., 1993. Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. Science 261 (5123), 921–923 (Aug.).

Corder, E.H., Saunders, A.M., Risch, N.J., Strittmatter, W.J., Schmechel, D.E., Gaskell Jr., P.C., Rimmler, J.B., Locke, P.A., Conneally, P.M., Schmader, K.E., 1994. Protective effect of apolipoprotein E type 2 allele for late onset Alzheimer disease. Nature Genetics 7 (2), 180–184 (Jun.).

Costafreda, S.G., Dinov, I.D., Tu, Z., Shi, Y., Liu, C.-Y., Kloszewska, I., Mecocci, P., Soininen, H., Tsolaki, M., Vellas, B., Wahlund, L.-O., Spenger, C., Toga, A.W., Lovestone, S., Simmons, A., 2011. Automated hippocampal shape analysis predicts the onset of dementia in mild cognitive impairment. NeuroImage 56 (1), 212–219 (May).

Coupé, P., Eskildsen, S.F., Manjón, J.V., Fonov, V.S., Pruessner, J.C., Allard, M., Collins, D.L., 2012. Scoring by nonlocal image patch estimator for early detection of Alzheimer's disease. NeuroImage: Clinical 1 (1), 141–152.

Cuingnet, R., Gerardin, E., Tessieras, J., Auzias, G., Lehéricy, S., Habert, M.-O., Chupin, M., Benali, H., Colliot, O., 2010. Automatic classification of patients with Alzheimer's disease from structural MRI: A comparison of ten methods using the ADNI database. NeuroImage 56 (2), 766–781 (May).

Davatzikos, C., Bhatt, P., Shaw, L.M., Batmanghelich, K.N., Trojanowski, J.Q., 2011. Prediction of MCI to AD conversion, via MRI, CSF biomarkers, and pattern classification. Neurobiology of Aging 32 (12), 2322.e19–2322.e27 (Dec.).

Desikan, R.S., Cabral, H.J., Hess, C.P., Dillon, W.P., Glastonbury, C.M., Weiner, M.W., Schmansky, N.J., Greve, D.N., Salat, D.H., Buckner, R.L., Fischl, B., Alzheimer's Disease Neuroimaging Initiative, 2009. Automated MRI measures identify individuals with mild cognitive impairment and Alzheimer's disease. Brain 132 (8), 2048–2057 (Aug.).

Drzezga, A., Lautenschlager, N., Siebner, H., Riemenschneider, M., Willoch, F., Minoshima, S., Schwaiger, M., Kurz, A., 2003. Cerebral metabolic changes accompanying conversion of mild cognitive impairment into Alzheimer's disease: a PET follow-up study. European Journal of Nuclear Medicine and Molecular Imaging 30 (8), 1104–1113 (Aug.).

Eskildsen, S.F., Coupé, P., García-Lorenzo, D., Fonov, V., Pruessner, J.C., Collins, D.L., 2013. Prediction of Alzheimer's disease in subjects with mild cognitive impairment from the ADNI cohort using patterns of cortical thinning. NeuroImage 65, 511–521 (Jan.).

Fan, Y., Batmanghelich, N., Clark, C.M., Davatzikos, C., 2008. Spatial patterns of brain atrophy in MCI patients, identified via high-dimensional pattern classification, predict subsequent cognitive decline. NeuroImage 39 (4), 1731–1743 (Feb.).
Ferrarini, L., Frisoni, G.B., Pievani, M., Reiber, J.H.C., Ganzola, R., Milles, J., 2009. Morphological hippocampal markers for automated detection of Alzheimer's disease and mild cognitive impairment converters in magnetic resonance images. Journal of Alzheimer's Disease 17 (3), 643–659 (Jan.).
Ferri, C.P., Prince, M., Brayne, C., Brodaty, H., Fratiglioni, L., Ganguli, M., Hall, K., Hasegawa, K., Hendrie, H., Huang, Y., Jorm, A., Mathers, C., Menezes, P.R., Rimmer, E., Scazufca, M., 2005. Global prevalence of dementia: a Delphi consensus study. Lancet 366 (9503), 2112–2117 (Dec.).
Fjell, A.M., Walhovd, K.B., Fennema-Notestine, C., McEvoy, L.K., Hagler, D.J., Holland, D., Brewer, J.B., Dale, A.M., 2010a. CSF biomarkers in prediction of cerebral and clinical change in mild cognitive impairment and Alzheimer's disease. Journal of Neuroscience 30 (6), 2088–2101 (Feb.).
Fjell, A.M., Walhovd, K.B., Fennema-Notestine, C., McEvoy, L.K., Hagler, D.J., Holland, D., Blennow, K., Brewer, J.B., Dale, A.M., 2010b. Brain atrophy in healthy aging is related to CSF Levels of Aβ1-42. Cerebral Cortex 20 (9), 2069–2079 (Sep.).
Fonteijn, H.M., Modat, M., Clarkson, M.J., Barnes, J., Lehmann, M., Hobbs, N.Z., Scahill, R.I., Tabrizi, S.J., Ourselin, S., Fox, N.C., Alexander, D.C., 2012. An event-based model for disease progression and its application in familial Alzheimer's disease and Huntington's disease. NeuroImage 60 (3), 1880–1889 (Apr.).
Gerardin, E., Chételat, G., Chupin, M., Cuingnet, R., Desgranges, B., Kim, H.-S., Niethammer, M., Dubois, B., Lehéricy, S., Garnero, L., Eustache, F., Colliot, O., 2009. Multidimensional classification of hippocampal shape features discriminates Alzheimer's disease and mild cognitive impairment from normal aging. NeuroImage 47 (4), 1476–1486 (Oct.).
Gousias, I.S., Rueckert, D., Heckemann, R.A., Dyet, L.E., Boardman, J.P., Edwards, A.D., Hammers, A., 2008. Automatic segmentation of brain MRIs of 2-year-olds into 83 regions of interest. NeuroImage 40 (2), 672–684 (Apr.).
Gray, K.R., Wolz, R., Heckemann, R.A., Aljabar, P., Hammers, A., Rueckert, D., 2012. Multi-region analysis of longitudinal FDG-PET for the classification of Alzheimer's disease. NeuroImage 60 (1), 221–229 (Mar.).
Hinrichs, C., Singh, V., Xu, G., Johnson, S.C., 2011. Predictive markers for AD in a multi-modality framework: an analysis of MCI progression in the ADNI population. NeuroImage 55 (2), 574–589 (Mar.).
Holtzman, D.M., 2011. CSF biomarkers for Alzheimer's disease: current utility and potential future use. Neurobiology of Aging 32 (Supplement 1, no. 0), S4–S9 (Dec.).
Joshi, A., Koeppe, R.A., Fessler, J.A., 2009. Reducing between scanner differences in multi-center PET studies. NeuroImage 46 (1), 154–159 (May).
Klöppel, S., Stonnington, C.M., Chu, C., Draganski, B., Scahill, R.I., Rohrer, J.D., Fox, N.C., Jack, C.R., Ashburner, J., Frackowiak, R.S.J., 2008. Automatic classification of MR scans in Alzheimer's disease. Brain 131 (Pt 3), 681–689 (Mar.).
Kriegeskorte, N., Simmons, W.K., Bellgowan, P.S.F., Baker, C.I., 2009. Circular analysis in systems neuroscience: the dangers of double dipping. Nature Neuroscience 12 (5), 535–540.
Landau, S.M., Harvey, D., Madison, C.M., Reiman, E.M., Foster, N.L., Aisen, P.S., Petersen, R.C., Shaw, L.M., Trojanowski, J.Q., Jack, C.R., Weiner, M.W., Jagust, W.J., 2010. Comparing predictors of conversion and decline in mild cognitive impairment. Neurology 75 (3), 230–238 (Jul.).
Lerch, J.P., Pruessner, J., Zijdenbos, A.P., Collins, D.L., Teipel, S.J., Hampel, H., Evans, A.C., 2008. Automated cortical thickness measurements from MRI can accurately separate Alzheimer's patients from normal elderly controls. Neurobiology of Aging 29 (1), 23–30 (Jan.).
Marquand, A., Howard, M., Brammer, M., Mourao-Miranda, J., 2009. Probabilistic classification of functional magnetic resonance imaging (fMRI) data using Gaussian process classification: application to pain perception. NeuroImage 47, S57 (Jul.).
McNemar, Q., 1947. Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika 12 (2), 153–157.
Minka, T.P., 2001. Expectation propagation for approximate Bayesian inference. Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence, San Francisco, CA, USA, pp. 362–369.
Modat, M., Ridgway, G.R., Taylor, Z.A., Lehmann, M., Barnes, J., Hawkes, D.J., Fox, N.C., Ourselin, S., 2010. Fast free-form deformation using graphics processing units. Computer Methods and Programs in Biomedicine 98 (3), 278–284 (Jun.).
Mosconi, L., Berti, V., Glodzik, L., Pupi, A., De Santi, S., de Leon, M.J., 2010. Pre-clinical detection of Alzheimer's disease using FDG-PET, with or without amyloid imaging. Journal of Alzheimer's Disease 20 (3), 843–854.
Nho, K., Shen, L., Kim, S., Risacher, S.L., West, J.D., Foroud, T., Jack, C.R., Weiner, M.W., Saykin, A.J., 2010. Automatic prediction of conversion from mild cognitive impairment to probable Alzheimer's disease using structural magnetic resonance imaging. AMIA Annual Symposium Proceedings 2010, 542–546.
Nickisch, H., Rasmussen, C.E., 2008. Approximations for binary Gaussian process classification. Journal of Machine Learning Research 9, 2035–2078 (Oct.).
Petersen, R.C., Smith, G.E., Waring, S.C., Ivnik, R.J., Tangalos, E.G., Kokmen, E., 1999. Mild cognitive impairment: clinical characterization and outcome. Archives of Neurology 56 (3), 303–308 (Mar.).
Querbes, O., Aubry, F., Pariente, J., Lotterie, J.-A., Demonet, J.-F., Duret, V., Puel, M., Berry, I., Fort, J.-C., Celsis, P., The Alzheimer's Disease Neuroimaging Initiative, 2009. Early diagnosis of Alzheimer's disease using cortical thickness: impact of cognitive reserve. Brain 132 (8), 2036–2047 (Aug.).
Rasmussen, C.E., Williams, C.K.I., 2006. Gaussian Processes for Machine Learning. MIT Press.
Robert, R., Wark, K.L., 2012. Engineered antibody approaches for Alzheimer's disease immunotherapy. Archives of Biochemistry and Biophysics 526 (2), 132–138 (Oct.).
Rowe, C.C., Ellis, K.A., Rimajova, M., Bourgeat, P., Pike, K.E., Jones, G., Fripp, J., Tochon-Danguy, H., Morandeau, L., O'Keefe, G., Price, R., Raniga, P., Robins, P., Acosta, O., Lenzo, N., Szoeke, C., Salvado, O., Head, R., Martins, R., Masters, C.L., Ames, D., Villemagne, V.L., 2010. Amyloid imaging results from the Australian Imaging, Biomarkers and Lifestyle (AIBL) study of aging. Neurobiology of Aging 31 (8), 1275–1283 (Aug.).
Singh, N., Wang, A., Sankaranarayanan, P., Fletcher, P., Joshi, S., 2012. Genetic, structural and functional imaging biomarkers for early detection of conversion from MCI to AD. Medical Image Computing and Computer-Assisted Intervention — MICCAI 2012, vol. 7510, pp. 132–140.
Vemuri, P., Gunter, J.L., Senjem, M.L., Whitwell, J.L., Kantarci, K., Knopman, D.S., Boeve, B.F., Petersen, R.C., Jack Jr., C.R., 2008. Alzheimer's disease diagnosis in individual subjects using structural MR images: validation studies. NeuroImage 39 (3), 1186–1197 (Feb.).
Vemuri, P., Wiste, H.J., Weigand, S.D., Shaw, L.M., Trojanowski, J.Q., Weiner, M.W., Knopman, D.S., Petersen, R.C., Jack Jr., C.R., 2009. MRI and CSF biomarkers in normal, MCI, and AD subjects: predicting future clinical change. Neurology 73 (4), 294–301 (Jul.).
Vounou, M., Janousova, E., Wolz, R., Stein, J.L., Thompson, P.M., Rueckert, D., Montana, G., 2012. Sparse reduced-rank regression detects genetic associations with voxel-wise longitudinal phenotypes in Alzheimer's disease. NeuroImage 60 (1), 700–716 (Mar.).
Walhovd, K.B., Fjell, A.M., Brewer, J., McEvoy, L.K., Fennema-Notestine, C., Hagler, D.J., Jennings, R.G., Karow, D., Dale, A.M., the Alzheimer's Disease Neuroimaging Initiative, 2010. Combining MR imaging, positron-emission tomography, and CSF biomarkers in the diagnosis and prognosis of Alzheimer disease. American Journal of Neuroradiology 31 (2), 347–354 (Feb.).
Warfield, S.K., Zou, K.H., Wells, W.M., 2004. Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. IEEE Transactions on Medical Imaging 23 (7), 903–921 (Jul.).
Wee, C.-Y., Yap, P.-T., Shen, D., the Alzheimer's Disease Neuroimaging Initiative, 2012. Prediction of Alzheimer's disease and mild cognitive impairment using cortical morphological patterns. Human Brain Mapping. http://dx.doi.org/10.1002/hbm.22156.
Ye, J., Farnum, M., Yang, E., Verbeeck, R., Lobanov, V., Raghavan, N., Novak, G., DiBernardo, A., Narayan, V.A., 2012. Sparse learning and stability selection for predicting MCI to AD conversion using baseline ADNI data. BMC Neurology 12 (1), 46 (Jun.).
Young, J., Modat, M., Cardoso, M.J., Ashburner, J., Ourselin, S., 2012. Classification of Alzheimer's disease patients and controls with Gaussian processes. 2012 9th IEEE International Symposium on Biomedical Imaging (ISBI), pp. 1523–1526.
Yuan, L., Wang, Y., Thompson, P.M., Narayan, V.A., Ye, J., 2012. Multi-source feature learning for joint analysis of incomplete multiple heterogeneous neuroimaging data. NeuroImage 61 (2), 622–632 (Jul.).
Zhang, D., Wang, Y., Zhou, L., Yuan, H., Shen, D., 2011. Multimodal classification of Alzheimer's disease and mild cognitive impairment. NeuroImage 55 (3), 856–867 (Apr.).