NeuroImage 112 (2015) 232–243

Contents lists available at ScienceDirect

NeuroImage

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

Gaussian process classification of Alzheimer's disease and mild cognitive impairment from resting-state fMRI

Edward Challis [a], Peter Hurley [a], Laura Serra [c], Marco Bozzali [c], Seb Oliver [a], Mara Cercignani [b,*]

[a] Department of Physics and Astronomy, University of Sussex, Falmer, East Sussex BN1 9QH, UK
[b] Clinical Imaging Sciences Centre, Brighton and Sussex Medical School, University of Sussex, Falmer, East Sussex BN1 9PR, UK
[c] Neuroimaging Laboratory, Santa Lucia Foundation, Via Ardeatina 306, Roma, Italy

ARTICLE INFO

Article history:
Accepted 17 February 2015
Available online 28 February 2015

Keywords:
Machine learning
Functional connectivity
Dementia

ABSTRACT

Multivariate pattern analysis and statistical machine learning techniques are attracting increasing interest from the neuroimaging community. Researchers and clinicians are also increasingly interested in the study of functional-connectivity patterns of brains at rest and how these relations might change in conditions like Alzheimer's disease or clinical depression. In this study we investigate the efficacy of a specific multivariate statistical machine learning technique to perform patient stratification from functional-connectivity patterns of brains at rest. Whilst the majority of previous approaches to this problem have employed support vector machines (SVMs) we investigate the performance of Bayesian Gaussian process logistic regression (GP-LR) models with linear and non-linear covariance functions. GP-LR models can be interpreted as a Bayesian probabilistic analogue to kernel SVM classifiers. However, GP-LR methods confer a number of benefits over kernel SVMs. Whilst SVMs only return a binary class label prediction, GP-LR, being a probabilistic model, provides a principled estimate of the probability of class membership. Class probability estimates are a measure of the confidence the model has in its predictions, such a confidence score may be extremely useful in the clinical setting. Additionally, if miss-classification costs are not symmetric, thresholds can be set to achieve either strong specificity or sensitivity scores. Since GP-LR models are Bayesian, computationally expensive cross-validation hyper-parameter grid-search methods can be avoided. We apply these methods to a sample of 77 subjects; 27 with a diagnosis of probable AD, 50 with a diagnosis of a-MCI and a control sample of 39. All subjects underwent a MRI examination at 3 T to obtain a 7 minute and 20 second resting state scan. Our results support the hypothesis that GP-LR models can be effective at performing patient stratification: the implemented model achieves 75% accuracy disambiguating healthy subjects from subjects with amnesic mild cognitive impairment and 97% accuracy disambiguating amnesic mild cognitive impairment subjects from those with Alzheimer's disease, accuracies are estimated using a held-out test set. Both results are significant at the 1% level.

Introduction

A broad goal of neuroimaging research is to develop effective, reliable clinical tools for the early detection and diagnosis of a range of neurological conditions such as dementia, depression and attention deficit hyperactivity disorder. Machine learning (ML) seems a promising route to help achieve such objectives. ML is the study of algorithms and computational techniques that use previous examples in the form of multivariate datasets to help make future predictions. One application of a ML prediction algorithm in the context of neuroimaging would be to make clinical diagnoses from subject's functional MRI (fMRI) scans. Alongside providing a computational and statistical framework within which to make predictions from multivariate observations, ML can also provide insights into what multivariate features of the data are most relevant for making accurate predictions. In the context of neuroimaging for patient stratification those features correspond to biomarkers of disease states.

In this paper we present a ML technique to perform patient stratification between healthy control subjects and either amnesic mild cognitive impairment (a-MCI) or Alzheimer's disease subjects. Subject classifications are made from the functional-connectivity scores of their brains inferred from resting state fMRI (rsfMRI) scans. Previous rsfMRI patient stratification studies have applied support vector machines (SVMs) to make inter-group classifications. Our approach here

Abbreviations: AC, accuracy; AD, Alzheimers Disease; a-MCI, amnestic mild cognitive impairment; ARD, automatic relevance determination; AUC, area under the ROC curve; DMN, default mode network; GP-LR, Gaussian process logistic regression; ICA, independent component analysis; LOOCV, leave one out cross validation; MMSE, mini-mental state examination; NC, normal control; PCA, principal component analysis; SC, specificity; SS, sensitivity; SVM, support vector machine

* Corresponding author at: Clinical Imaging Sciences Centre, University of Sussex, Falmer, Brighton BN1 9RR, UK. Fax: +44 1273 876721.
E-mail address: m.cercignani@bsms.ac.uk (M. Cercignani).

http://dx.doi.org/10.1016/j.neuroimage.2015.02.037

is to use Gaussian process logistic regression (GP-LR) models. Whilst examples of application of Gaussian processes to neuroimaging data exist (e.g. Marquand et al., 2010), we are unaware of any other functional-connectivity studies using Gaussian process models to make such group level predictions.

The motivations of our work are two-fold. First, to show how GP-LR models can be applied to inter-group rsfMRI classification problems and to highlight some advantages of this approach. Second, to investigate what classification accuracy such a technique can achieve in distinguishing between healthy, a-MCI and AD subjects, and to identify what features of the data are most relevant in driving those predictions.

Machine learning fMRI studies

A neuroimaging problem to which ML can be applied is to predict whether a subject belongs to one of a number of different subject groups, for example to predict if a subject is healthy versus diseased or young versus old. Due to inter-subject, inter-scan and inter-centre variability, and the often limited number of example scans that are available to researchers, this is typically a hard statistical inference problem. In this study we seek to address the inter-group prediction problem. We apply ML methods with the specific aim to automatically disambiguate healthy control subjects from subjects exhibiting symptoms of amnesic mild cognitive impairment (a-MCI) and Alzheimer's disease (AD). Importantly, our data is not longitudinal; each scan corresponds to a different individual.

A good introduction to machine learning methods applied to neuroimaging problems can be found in the review articles by Pereira et al. (2009), Lemm et al. (2011) and Ashburner and Klöppel (2011). A more general introduction to probabilistic machine learning and Bayesian methods can be found in Barber (2012).

Resting-state functional MRI

We seek to perform patient stratification from the application of ML algorithms to resting-state fMRI scans. RsfMRI data refers to fMRI scans that are recorded whilst the subject is at rest; that is, the subject is not performing any particular task and is not asleep. From a practical perspective, resting-state scans have the advantage of being easier to acquire than scans recorded whilst the subject is performing a task because fewer experimental variables have to be controlled for. Thus, inter-scan differences that are not attributable to the subjects' mental state are minimised and group differences will be easier to infer. Furthermore, since many subjects, such as those that have Alzheimer's disease or dementia, are often incapable of carrying out cognitive tasks required by task-based studies, resting-state studies have the benefit of being able to include such subjects without biasing the experimental design.

RsfMRI voxel blood-oxygenation-level-dependent (BOLD) signal time-courses exhibit low frequency, ($\approx$0.1 Hz), oscillations. These spontaneous BOLD signal oscillations exhibit temporal correlations across spatially distinct brain regions. Such patterns of activity are now commonly believed to mirror the functional-connectivity patterns of the brain (Van Den Heuvel and Hulshoff Pol, 2010). Assuming that these patterns reflect specific resting state networks, one of them, namely the default mode network (DMN) has received particular attention. Evidence suggests that during goal directed behaviour the DMN correlations are suppressed (Buckner et al., 2008; Gusnard and Raichle, 2001). Multiple studies have observed that changes to the DMN may be biomarkers for various neurological conditions such as Alzheimer's disease (Koch et al., 2012; Greicius et al., 2004b), attention deficit hyperactivity disorder (Uddin et al., 2008; Liddle et al., 2011) and depression (Zeng et al., 2012; Sheline et al., 2009; Bluhm et al., 2009) amongst other studies.

Functional-connectivity

Functional-connectivity is commonly defined as the temporal dependence of neuronal activity patterns of anatomically separated brain regions (Friston et al., 1993). As such, functional-connectivity is a property of the brain that is static and independent of time. Whole brain resting-state functional-connectivity patterns are obtained by studying the coactivation between the time-courses of voxels, or collections of voxels, that are spatially distributed. Typically, the methods that are employed to discern functional-connectivity relations from rsfMRI data fall into two categories: Model-free methods such as independent component analysis or principal component analysis and model-based methods such as region of interest or seed correlation analysis. See Cole et al. (2010) for an introductory review of these techniques.

Model-free methods aim to find a reduced set of temporal basis functions such that each voxel's BOLD time-course can be well approximated by a liner combination of these temporal bases. The temporal basis functions are most frequently estimated using either the principal component analysis (PCA) or the independent component analysis (ICA) statistical models. Having applied PCA or ICA, functional-connectivity between two anatomically distinct regions is inferred if the two groups share similar temporal basis function coefficients. ICA and PCA methods are thought of as model-free in the sense that no brain region atlas is defined by the researcher a priori. However, the temporal bases are found by fitting a statistical model which makes certain assumptions about the data generating process, for example PCA finds the basis functions that span the directions of maximum variance and ICA finds basis functions that span the directions that maximise the kurtosis (or some other proxy of statistical independence). In this sense the model-free label is misleading. A practical consideration when using ICA or PCA methods is that the temporal bases can be difficult to interpret — deciding whether a basis is due to 'noise' or neuronal variability is typically decided by a human expert. Automatically ordering and labelling the temporal bases is the subject of on-going methodological research (Tohka et al., 2008; De Martino et al., 2007). A further difficulty with applying model-free methods as a data preprocessing step for making inter-group predictions is that it is unclear whether the temporal bases calculated from one group generalise to another. These issues make it difficult to apply model-free methods as a functional-connectivity preprocessing step in a ML system designed to make inter-group predictions.

An alternative to model-free methods are so called model-based methods. Model-based methods infer functional-connectivity by inspecting the temporal dependence in BOLD signals between anatomically distinct brain regions. Whilst many different time-course dependence metrics could be used to infer functional-connectivity (Zhou et al., 2009), a commonly used and simple metric is the spontaneous correlation in BOLD signals between brain regions. In such an analysis regions that have highly correlated time-courses are inferred to be functionally connected. We refer to this approach as the regions of interest (ROIs) method. Other names used in the literature include volumes of interest or seed based correlation analysis. These techniques are thought of as model-based because the seed ROIs need to be specified a priori and so connectivity patterns are not directly inferred from the data. The primary strength of this approach is the ease with which it can be implemented and the results interpreted. Thus, the ROI approach is the favoured functional-connectivity preprocessing technique for patient stratification ML studies (Craddock et al., 2009; Zeng et al., 2012; Meier et al., 2012; Anderson et al., 2011). There is also some evidence, specific to the problem of disambiguating healthy versus AD subjects, that model-based methods may have more diagnostic power than model-free methods (Koch et al., 2012). In this work, the authors hypothesise that model-based methods may have more diagnostic power due to correlational analysis being more robust to BOLD signal variability that is observed to increase with age or the partial volume effects of grey matter loss.

Discerning group differences

Previous research has explored using ML classifiers to perform patient stratification from rsfMRI scans. Although techniques such as quasinearest neighbour and random forest analysis have been used as classifiers (e.g. Shen et al., 2010b; Venkataraman et al., 2010, 2012), most studies have used support vector machines: Craddock et al. (2009) and Zeng et al. (2012) applied SVMs to disambiguate healthy control from depressed subjects, Fan et al. (2011) applied SVMs to disambiguate healthy control from patients with schizophrenia, and Meier et al. (2012) applied SVMs to disambiguate elderly versus young subjects. Methodologically, these studies differ in the data preprocessing steps and the kernel functions used. Craddock et al. (2009) defined the feature vectors as the correlation scores between 15 selected ROIs, the authors studied the performance of a variety of feature selection methods using a linear kernel SVM. Zeng et al. (2012) created feature vectors from the correlation scores between 116 ROIs covering the whole brain, features that achieved the strongest Kendall tau correlation, a metric of statistical dependence we describe in the Feature subset selection section, with the class label were included in a linear kernel SVM. Fan et al. (2011) applied ICA to all subjects in the group to obtain functional-connectivity scores, the ICA features were then projected on to a non-linear manifold and classifications made using a non-linear sigmoid kernel SVM. Meier et al. (2012) created feature vectors from the correlation scores of 100 ROIs, feature selection was implemented by only including the 200 features that had the strongest group difference statistics, binary classification was performed using the radial basis function kernel SVM.

Zhang et al. (2011) performed a multimodal group comparison using structural MRI, functional PET and CSF protein level data to disambiguate healthy controls versus MCI or AD subject-groups. Combining features from each of these modalities, the authors report classification accuracies of 93% and 76% for the NC versus AD and NC versus MCI tasks. Classifications were made using a multiple-kernel SVM. The MCI subjects were then partitioned into those subjects that either had or had not converted to AD in the following 18 months, for the MCI converter subjects their model achieved 91% accuracy.

Various, non-ML, functional-connectivity studies have been conducted that look for group differences between healthy controls, MCI and AD subject-groups. Greicius et al. (2004b) conducted a group, rsfMRI, ICA study comparing healthy controls with AD subjects. The authors reported that the AD group had comparatively less activity in the posterior cingulate and hippocampus and suggest that changes to DMN activity may prove to be a useful biomarker of incipient AD. Koch et al. (2012) applied both ICA and ROI functional-connectivity techniques to make group level comparisons between healthy controls, MCI and AD subjects from rsfMRI scans. The authors report that no significant changes in DMN connectivity could be observed between healthy controls and MCI subjects. However, DMN connectivity was significantly depressed in AD subjects as compared to the control group, in particular the connections between the posterior cingulate and the superior frontal cortex were found to exhibit the largest differences. With the aim of building an fMRI based diagnostic tool to detect a-MCI and AD subjects these studies are promising since they support the hypothesis that detectable changes in the functional connectivity patterns of these subject groups exist. However, since these studies are not predictive in nature they fail to provide evidence of how accurately an automatic diagnostic system might perform using functional-connectivity dependent measures.

Overview and structure of paper

The remainder of this paper is structured as follows. In the Data collection and preprocessing section we describe the subjects recruited for the study, the data collection process and how each subject's fMRI scan was converted to a functional-connectivity feature vector. In the Machine learning classifier section we briefly introduce the GP-LR model and some design choices that were made. In the Classification experiments section we describe the experiments that were performed, the metrics we used to analyse performance and the results obtained. In the Functional-connectivity analysis section we present a brief analysis of which features of the data were driving the classifier's performance. In the Summary and discussion section we discuss and summarise our core contributions. Appendices give more details of the methodology and results from an alternative choice of feature normalisation.

Data collection and preprocessing

In total, 77 participants were enrolled for this study; 27 with a diagnosis of probable AD (proportion of females (F/N) = 0.81, age mean = 68 years and standard deviation (s.d.) = 6 years, Mini Mental State Examination (MMSE) score mean = 19 and s.d. = 5) and 50 with a diagnosis of a-MCI (F/N = 0.44, age mean = 66 s.d. = 7 years and MMSE mean = 26 s.d. = 4). Local ethical committee approval and written informed consent were obtained before study initiation. Additionally, 39 age and gender matched healthy controls (NC) were also recruited (F/N = 0.46, age mean = 63 s.d. = 9 years and MMSE mean = 26 s.d. = 9).

The diagnosis of probable AD was defined according to the clinical criteria established by the National Institute of Neurological and Communicative Disorders and Stroke-Alzheimers Disease and Related Disorders Association (NINCDS-ADRDA) (McKhann et al., 1984). The diagnosis of a-MCI was performed according to current criteria (Petersen et al., 2001; Petersen, 2004). All participants had to be right-handed as assessed by the Edinburg Handedness Inventory (Busch et al., 2010), in order to reduce any potential source of variability due to hemispheric dominance.

All subjects underwent a MRI examination at 3 T (Siemens, Medical Solutions, Erlangen, Germany), including the following acquisitions: 3D modified driven equilibrium Fourier transform (MDEFT) scan (TR = 1338 ms, TE = 2.4 ms); and T2* weighted echo planar (EPI) sensitised to BOLD contrast (repetition time TR = 2080 ms, echo time TE = 30 ms, 32 axial slices, matrix = 64 × 64, pixel size = 3 × 3 mm2, slice thickness = 2.5 mm, flip angle: 70°) for rsfMRI. BOLD EPIs were collected during rest for a 7 minute and 20 second period, resulting in a total of 220 volumes. During this acquisition, subjects were instructed to keep their eyes closed, not to think of anything in particular and not to fall asleep. Data are available upon request.

Data preprocessing

The fMRI scans were processed using Matlab7 and SPM8.[1] The preprocessing steps included smoothing, correction for head motion, compensation for slice-dependent time shifts, and normalisation. With respect to motion correction, each data set was checked to ensure that the maximum absolute shift did not exceed 2 mm, and the maximum absolute rotation did not exceed 1.5°. Smoothing was applied using a 3D Gaussian kernel with 8 mm^3 FWHM. In house software was used to remove the global temporal drift using a 3rd order polynomial fit. The data was then filtered by regressing out movement vectors, and average white matter and cerebrospinal fluid signal. The data was filtered further by a phase-insensitive band-pass filter (pass band 0.01 to 0.08 Hz) to reduce the effect of low frequency drift and high frequency physiological noise. To avoid saturation effects the first four volumes of each scan were discarded.

Each subject's rsfMRI scan was then converted to a brain region connectivity feature vector defined as the variance-covariance in BOLD signals between 82 anatomically distinct regions of interest (ROIs). ROIs were defined according to the Automated Anatomical Labeling (AAL) brain atlas (Tzourio-Mazoyer et al., 2002). This was obtained by applying region-specific masks from the AAL atlas (after reslicing to the same

[1] The SPM8 Matlab package can be downloaded from www.fil.ion.ucl.ac.uk/spm.

resolution as the data) to each subject's warped and smoothed fMRI scan, and extracting the mean time course for each region. ROIs corresponding to the cerebellum (numbered 90–116 inclusive) and the precentral and postcentral gyrus, calcarine cortex, and lingual gyrus, bilaterally, were excluded a priori since we did not expect these regions to aid discriminatory performance (Braak and Braak, 1995, 1996). The benefit of this a priori selection is in the reduction of the feature vector dimensionality, which is much larger than the number of examples.

The covariance between each ROI was calculated as follows. ROI time-courses were obtained by averaging the BOLD signal recorded at each voxel in that ROI for each time-point, denoting $r_i{}^t$ as the activation of ROI i at time point t we defined

$$r_i^t := \sum_{j \in \mathrm{ROI}_i} v_j^t,$$

where vjt corresponds to the BOLD signal of voxel j at time-point t. Having calculated all ROI time-courses, the covariance between ROI k and ROI l, which we denote σ_{kl}, was calculated using

$$\sigma_{kl} := \frac{1}{T}\sum_{t=1}^{T}\left(r_t^k - \bar{r}_t^k\right)\left(r_t^l - \bar{r}_t^l\right),$$

$T = 216$ after having excluded the first 4 volumes. Thus each subject scan was mapped to a $82 \times (82 + 1) \times 1 = 3403$ dimensional feature vector describing the covariance matrix between all the included ROIs. Covariance, rather than correlation, was used as the connectivity metric since we wanted to include variances, i.e. σ_{ii} terms, because we hypothesised that ROI activity itself could be a relevant biomarker.

The final feature vector for each subject was then constructed by appending their age and mini mental state examination (MMSE) score to the ROI variance–covariance vector resulting in a 3405 dimensional feature vector describing each subject. We denote the feature vector of subject n by the column vector

$$x^n := \left[\text{age},\ \text{mmse},\ \sigma_{1,1}, \ldots, \sigma_{82,82}\right]^{\mathrm{T}} \in \mathbb{R}^{3405}.$$

Machine learning classifier

In this section we describe each stage of the machine learning data processing pipeline from feature selection, feature normalisation and model training to making predictions, detailing how each of these steps were implemented and the range of different model settings that were investigated.

Feature subset selection

Since the dataset has many more features, $D = 3405$, than examples, $N \approx 60$, we employed two simple strategies to reduce the number of active features in the classifier and hence to reduce the risk of over-fitting. Both strategies are carried out in the cross-validation stage described in the Model selection experiments section. First we supplied the GP-LR model with only the features that obtained the largest absolute Kendall tau correlation coefficients versus the class label. This feature reduction strategy has been used successfully in other machine learning neuroimaging studies (Zeng et al., 2012). Second we used GP covariance functions with automatic relevance determination (ARD) parameterisations to down weight the contribution of less relevant features. Details about the GP-LR model, including the specifics of the ARD covariance functions, are presented in Appendix A.

Kendall's tau correlation coefficient is a statistical measure of the correlation between the ordering of two variables (Kendall, 1938). As such it is often used as an approximate measure of statistical independence. Kendall's tau correlation coefficient is bounded such that $\tau \in (-1, 1)$; when $\tau = \pm 1$ the two variables have perfectly correlated or anti-correlated rank and so are statistically dependent; two independent random variables have $\tau = 0$. Similar to some previous rsfMRI studies, we use the Kendall tau measure to select a subset of the features that we will pass to the classifier (Shen et al., 2010a; Zeng et al., 2012). For each of the $D = 3405$ features, we calculated the Kendall tau coefficient versus the binary class label giving us a set of coefficients $\tau_d{}^D{}_{d=1}$. We rank the coefficients according to their largest absolute value and create a new feature vector $\tau_d{}^D{}_{d=1}$ such that $i_1, \ldots, iD'$ index the D' largest absolute correlation scores. At the model selection stage we investigate the performance of models trained using $D' = 5$, 10, 15 and 20 included features. We expected this to be a reasonable range of feature set sizes for a training dataset consisting of ≈ 60 data points, given more training data we would increase this range.

Feature normalisation

To reduce the numerical burden of optimising the parameters associated with the GP-LR and possibly increase classification accuracy we investigated using two different feature normalisation procedures.

The first method was simple feature-wise scaling; each feature was linearly transformed such that $\tilde{x}_d^n \leftarrow \left(\tilde{x}_d^n - \mu_d\right) / \sigma_d$ where μ_d and σ_d are the empirical mean and standard deviation of feature $\tilde{x}_d^n$ in the training data. Having applied the transformation, each feature has approximately zero mean and unit standard deviation. The second method was to project the data on to its principal components; each input vector was linearly transformed so that $\tilde{x}^n \leftarrow C^{-½}\left(\tilde{x}^n - \mu\right)$ where μ and C are the empirical mean and covariance matrix of the training subset inputs $\tilde{x}^n$. Having applied this second data transformation the inputs will have approximately zero mean and identity covariance matrix. To reduce the notational burden, in what follows we drop the tilde and denote the feature subset and normalised input data vectors $\tilde{x}$ as x and assume that $D = D'$.

Gaussian process logistic regression model

Gaussian process (GP) regression models are a Bayesian non-parametric approach to solving regression and classification supervised ML problems. The Gaussian process logistic regression (GP-LR) model is a technique to solve binary classification problems.

Given a training dataset of input output pairs, $\mathcal{D} = (X, y)$ where $\left[x^1, \ldots, x^{N_{trn}}\right] = X \in R^{D \times N_{trn}}$ and $\left[y^1, \ldots, y^{N_{trn}}\right]^T = y \in R^{N_{trn}}$, GP-LR is a Bayesian probabilistic approach to modelling the relation between the inputs x and the outputs y. In its simplest form, a GP regression model is defined by the likelihood distribution function, the prior distribution function and the covariance function. Using the rules of probability these terms can be manipulated to derive the probability distribution of a new unseen test point's class label y^* given its input vector x^*, the training data $\mathcal{D}$ and covariance function parameters θ, $p(y^*|x^*, \mathcal{D}, \theta)$, and hence to make predictions.

In Appendix A in the Supplementary material we provide a more complete introduction and overview of the elements of the model and methods. A thorough introduction to GP methods applied to ML problems can be found in Rasmussen and Williams (2006).

The flexibility and performance of GP-LR models comes from choosing the appropriate covariance function for the data. In this paper we consider two simple covariance functions: (i) the linear ARD and (ii) the squared exponential ARD. Each of these covariance functions have 'hyper-parameters' that adapt how models behaviour. The hyper-parameters are: the additive Gaussian observation noise, s^2, a scaling parameter σf (square exponential only) and a length-scale parameter for each dimension, ld, controlling the correlation scale along that dimension of input space. We use the empirical Bayesian maximum likelihood-II procedure to select hyper-parameters.

Classification experiments

We applied the GP-LR model to i) classify healthy control (NC) versus a-MCI subjects, and ii) classify a-MCI versus AD subjects. For each problem we encode the 'healthier' state, NC or a-MCI, subjects as $y = -1$ and the 'disease' state subjects, a-MCI or AD, as $y = +1$. The aim of our first set of experiments is to perform model selection on the validation data; a discrete set of different GP-LR models are applied to the validation data and their performance is measured, on the basis of those results we select a single 'best' GP-LR classifier. In addition to the GP-LR classifier, we also perform a SVM classification during the validation stage in order to compare performance. The aim of our second set of experiments is to estimate predictive accuracy on a separate test set; using the optimal model settings found from the first set of experiments, we retrain the GP-LR model using all the validation data and then measure its performance on the held-out test data.

For each subject-group the data was randomly partitioned into a validation set and a ten subject test set. The demographic data of the subjects in each dataset partition are presented in Table 1. The number of years of formal education was significantly ($p = 0.01$) different between NC and MCI in the validation set, whilst the gender ratio (m/f) differed between AD and MCI in the validation set. No other between-group differences were found with respect to these variables.

The remainder of this section is organised as follows. In the Model selection experiments section we describe the experiments conducted to perform model selection, detailing how the model was fit to the data and how model performance was measured. In the Model selection results section we present the results from the model selection experiments and discuss how the final 'best' GP-LR models were chosen. In Test set results section we discuss how the final GP-LR model was trained and present the held-out test set results.

Model selection experiments

Since the number of training examples is limited, and since we want to obtain the most reliable estimate of predictive performance as possible, model selection results were obtained using an iterative leave- one-out cross-validation (LOOCV) procedure: For a binary classification task with N data points we optimise the model's hyper-parameters on a training set consisting of $N_{trn} = N - 1$ of the data points and validate the model on the single remaining data point, this is repeated N times for each training and validation partition.

For each of the N LOOCV folds we apply the following steps to the data: i) feature subset selection using Kendall tau correlation coefficient ranking ii) feature set normalisation, iii) hyper-parameter optimisation and iv) GP-LR validation point prediction. To each classification problem we applied each combination of the models settings we described in the Machine learning classifier section: feature subset sizes of $D = 5, 10, 15$ and 20; feature normalisation using the feature-wise linear scaling and the PCA projection; and GP priors using the linear ARD and the squared exponential ARD covariance functions. Thus, in total LOOCV was applied to 16 different GP-LR models and 8 SVM models, for both the NC vs a-MCI and the a-MCI vs AD classification problems.

Model fitting

For each binary classification problem, GP-LR model and LOOCV fold, the model was trained using the steps described in the Machine learning classifier section. Importantly, all pre-processing steps (feature selection, feature normalisation and hyper-parameter optimisation) were derived from the training data only so that our test data point predictions are unbiased. The Bayesian evidence and posterior distribution $p(f|\mathcal{D},\theta)$ were approximated using the EP algorithm. Hyper-parameter optimisation was performed using the conjugate gradient ascent algorithm terminating after 200 line searches. Hyper-parameters were initialised using the procedure described in the Supplementary material. The EP approximation, the hyper-parameter optimisation, and the predictive density approximation equation (A.4) were all implemented in our experiments using the free, GP Matlab package gpml developed by Rasmussen and Nickisch (2010). The SVM classifier was implemented using the free LIBSVM software developed by Chang and Lin (2011).[2]

Evaluation metrics

We applied five metrics to assess the performance of the GP-LR model: the accuracy, the specificity, the sensitivity, the predicted probability and the area-under-the-curve (AUC) scores. Below we describe each of these metrics and how they are calculated.

Accuracy. We derive the accuracy score by comparing the true class label, y^*, to the predicted class label, $\hat{y}^*$, using the symmetric threshold of $\eta = 0.5$. If the true and predicted class labels match we assign a score of $+1$ otherwise a score of 0. The final score is calculated by averaging the accuracy score over each of the predictions made. Thus, the total accuracy is defined

$$AC = \frac{TP + TN}{TP + FN + TN + FP},$$

where TP, FP, TN and FN denote the number of true positive, false positive, true negative and false negative predictions respectively.

Sensitivity and specificity. The accuracy is a weighted average of the specificity and sensitivity scores of the classifier. The sensitivity score is defined as

$$SS = \frac{TP}{TP + FN},$$

which measures the accuracy of the classifier at detecting 'disease' state (i.e. $y = +1$) subjects. The specificity score is defined as

$$SC = \frac{TN}{TN + FP},$$

which measures the accuracy of the classifier at detecting 'healthy' or 'control' state (i.e. $y = -1$) subjects. The specificity is equal to one minus the false positive rate.

Predicted probability. To measure both the accuracy and the confidence of the model we record the predicted probability of the true class label $p(\hat{y}^* = y^*|x^*,X,y,\hat{\theta})$. Averaging these values over the LOOCV folds we obtain the predicted probability score. From a probabilistic modelling perspective this metric is optimal since it measures how accurate the entire predictive distribution is.

AUC. If we plot the false positive rate (equivalent to one minus the specificity score) versus the sensitivity score as we vary the classification threshold η from zero to one we obtain the receiver operated characteristic (ROC) curve of the classifier. The ROC curve is a useful way to inspect the accuracy of a binary classifier over all classification thresholds. Two examples of ROC curves are presented in Fig. 1. The AUC metric is defined as the area under the ROC curve. The AUC is bounded between 0 and 1, a uniform random classifier will achieve an AUC of 0.5 on average, a perfect classifier will achieve an AUC of 1. Since the AUC integrates classification accuracy over all possible classification thresholds it is invariant to class imbalances and is thus an informative measure of the overall performance of a classifier. See Fawcett (2004) for an introduction to ROC curves and the AUC metric as applied to classification problems.

[2] LIBSVM Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Table 1
Age, gender, handedness and education statistics of each group in the validation and test datasets: the N row reports the total number of subjects in that group.

	Validation					Test				
Group	NC vs MCI p	NC	MCI	AD	AD vs MCI p	NC vs MCI p	NC	MCI	AD	AD vs MCI p
N		29	40	17			10	10	10	
Age [years]	0.26	64.0 ± 9.5	66.5 ± 7.3	68.6 ± 5.5	0.29	0.18	61.4 ± 8.5	66.2 ± 6.8	68.3 ± 6.4	0.49
Gender [m/f]	0.62	16/13	25/15	3/14	**0.003**	0.6	5/5	3/7	2/8	1.0
Handedness [R/L]	–	29/0	40/0	17/0	–	–	10/0	10/0	10/0	–
Education [years]	**0.01**	13.0 ± 3.3	10.5 ± 4.6	9.7 ± 3.4	0.49	0.06	14.0 ± 3.7	10.1 ± 4.9	8.6 ± 4.4	0.48

For age and education, we report the mean and standard deviation ages (years) of that group, and the p-value of a two sample, two tailed, t-test comparing the NC and AD groups versus the MCI group. For gender, a Chi-square test was used. p-values in bold indicate significant between-group differences.

Model selection results

First we review the LOOCV results for the a-MCI versus NC classification task, second we review the AD versus a-MCI results.

Healthy controls versus a-MCI subjects

The task of discriminating between healthy controls (NC) and the a-MCI subjects is typically harder than distinguishing AD subjects. Whilst multiple studies have reported AD discriminatory performance at the ≈90% level (Laakso et al., 1998; Klöppel et al., 2008; Cuingnet et al., 2011), authors typically report either no significant difference between a-MCI and control subjects (Koch et al., 2012; Cuingnet et al., 2011) or predictive accuracy performance at ≈75% (Zhang et al., 2011). LOOCV results for all model combinations are presented in Table 2 for the non-PCA normalised models and Table 6 of the additional on-line material for the PCA normalised feature models.

The first observation that can be made from inspecting these results is that the PCA normalised models perform significantly worse than the non-PCA models. We hypothesise that this is most likely due to there being insufficient training data to accurately estimate the principal components.

For the non-PCA normalised experiments, predictive accuracy scores using the squared exponential covariance marginally outperform the linear covariance results. Amongst the squared exponential covariance models, the strongest results are achieved using either D = 10 or D = 15 included features. Since the performance difference between these two models is relatively small, we choose the more parsimonious of the two, D = 10, as the 'best' GP-LR model. This model achieves AUC = 0.81, accuracy AC = 77%, specificity SC = 62% and sensitivity SS = 88% with $\eta = 0.5$.

To assess the significance of the accuracy scores we apply a one sided Binomial test: the null hypothesis assuming predictions are made by assigning all data points the majority class label — this naive classification rule has an expected classification accuracy of AC = 40/69 ≈ 0.60. Applying this test to the LOOCV accuracy score of the chosen GP-LR model we obtain a p-value of $\approx 3 \times 10^{-4}$. Despite the widespread use of this procedure, the trial independence assumptions of the Binomial significance test are not valid for LOOCV results (Kohavi, 1995). However, the trial independence assumption is valid for the held-out test set results. In the following section we are able confirm statistical significance at the 1% level for the held-out test set results using the Binomial significance test.

The accuracy, specificity and sensitivity scores presented in Tables 2 and 3 were all calculated using a classification threshold of $\eta = 0.5$. As we described in the Evaluation metrics section, this threshold can be tuned to achieve different accuracy, specificity and sensitivity profiles. However, to tune the classification threshold we would need a distinct dataset to both the training and test datasets. Since we do not have enough data points to do so we do not provide an in depth analysis of how this parameter might be tuned and the results that could be achieved by doing so. Here we adopt the naive strategy of only reporting results for a threshold of $\eta = 0.5$ — which is suboptimal. However, we note that the AUC metric is quite strong for the best GP-LR model suggesting that significant improvements could be achieved. To give an indication of the increased accuracy that could be obtained by tuning η, in Fig. 1 we plot the receiver operated characteristic (ROC) for this model. As we show in the figure, changing the threshold can lift the accuracy, specificity, or sensitivity scores to AC = 78%, SC = 97% or SS = 88%.

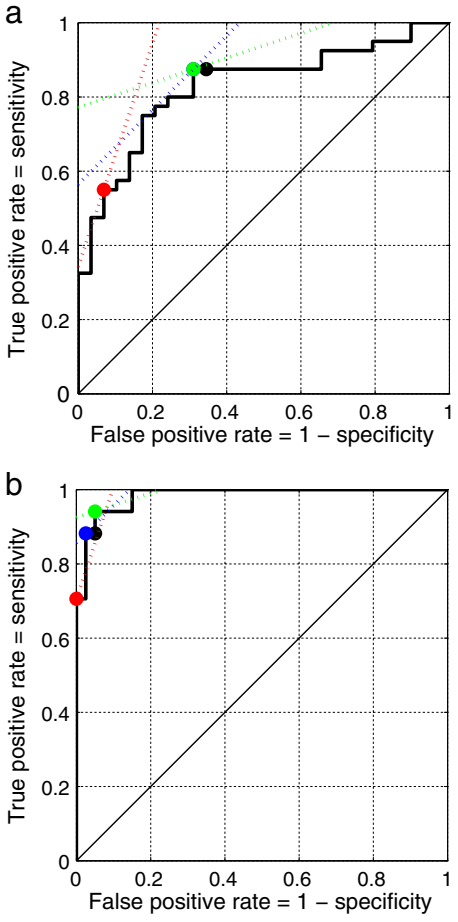

Fig. 1. ROC curves for the 'optimal' LOOCV models for the NC vs a-MCI and the a-MCI vs AD classification tasks. Subplot (a) correspond to the GP-LR model with the squared exponential ARD covariance, D = 10 included features and feature-wise scaling preprocessing. The AUC for this model and dataset is 0.82. For a symmetric miss-classification cost function the optimal classification threshold is at $\eta = 0.52$, at this setting the classifier has classification performance AC = 80%, SS = 88% and SC = 69% (blue dot). For a stronger sensitivity score we can set the threshold to $\eta = 0.52$ and achieve accuracies: AC = 80%, SS = 88% and SC = 69% (green dot). For a stronger specificity score we can set the threshold to $\eta = 0.91$ and achieve accuracies: AC = 71%, SS = 55% and SC = 93% (red dot). Using a threshold of $\eta = 0.5$ we obtain the results presented in Table 2 (black dot). Subplot (b) corresponds to the GP-LR model with D = 15 included features, non-PCA normalisation and the linear ARD covariance function. The AUC for this model is 0.98. For this model $\eta = 0.5$ gives performance figures of AC = 93%, SS = 88% and SC = 95%; setting $\eta = 0.37$ give performance figures of AC = 94%, SS = 94% and SC = 95%; setting $\eta =$ 0:99 give performance figures of AC = 91%, SS = 71% and SC = 100%.

Table 2
Model selection results for the LOOCV experiments using feature-wise scaling preprocessing.

Cov.	D	Nc vs a-MCI							a-MCI vs AD						
		logZ	AUC	p(y)	AC(%)	SC(%)	SS(%)	p-value	logZ	AUC	p(y)	AC(%)	SC(%)	SS(%)	p-value
K_{lin}	5	−28.9	0.72	62	68	62	73	3.21×10^{-2}	−6.2	0.95	90	91	93	88	2.9×10^{-5}
K_{se}	5	−31.3	0.73	64	68	62	73	3.21×0^{-2}	−9.4	0.91	85	88	90	82	4.9×10^{-4}
K_{lin}	10	−27.0	0.81	69	71	69	73	9.13×10^{-3}	−5.1	0.98	91	89	90	88	1.3×10^{-4}
K_{se}	10	−27.7	0.83	71	78	66	88	1.17×10^{-4}	−8.9	0.94	88	93	93	94	4.4×10^{-6}
K_{lin}	15	−26.6	0.82	70	74	76	73	1.99×10^{-3}	−4.8	0.98	93	93	95	88	4.4×10^{-6}
K_{se}	15	−26.5	0.85	74	74	66	80	1.99×10^{-3}	−8.5	0.94	87	91	93	88	2.6×10^{-5}
K_{lin}	20	−26.2	0.79	67	71	76	68	9.13×10^{-3}	−4.7	0.99	94	93	95	88	4.4×10^{-6}
K_{se}	20	−25.8	0.77	70	75	62	85	8.38×10^{-4}	−8.4	0.95	88	91	93	88	2.6×10^{-5}
SVM	5				70	64	73	1.77×10^{-2}				93	90	100	4.37×10^{-6}
SVM	10				75	71	78	8.38×10^{-4}				91	93	88	2.64×10^{-5}
SVM	15				80	76	83	3.89×10^{-5}				91	93	86	2.64×10^{-5}
SVM	20				81	83	79	1.18×10^{-5}				95	95	93	5.36×10^{-7}

Healthy control (NC) versus amnesic mild cognitive impairment (a-MCI) results are presented in the left hand section of the table, a-MCI versus Alzheimer's disease (AD) results are in the right hand section. All performance metric values are averaged over the N LOOCV folds, where N = 69 for the NC vs a-MCI task and N = 57 for the a-MCI vs AD task. Results are presented for the linear, K_{lin}, and squared exponential, K_{se}, ARD covariance functions. logZ denotes the average log marginal likelihood value over each of the LOOCV folds log $p(y|X, \theta)$. The AUC column reports the area under the ROC curve metric. $p(y)$ denotes the predicted probability score (see text). AC, SS and SC columns report the accuracy, sensitivity and specificity scores (%). The final column in each section reports the p-value of a one-sided Binomial significance test of the AC score under the null hypothesis that the classifier picks the majority class label i.e. $p(y^{*} = y^{*}) = 40/69$ for the NC vs a-MCI task and $p(y^{*} = y^{*}) = 17/57$ for the a-MCI vs AD task. For comparison, we add the AC, SS and SC scores and p-value obtained using LIBSVM (Chang and Lin, 2011), a linear SVM classification tool. As with the GPs, we use feature-wise scaling preprocessing and LOOCV.

For a comparison to the GP-LR models, Table 2 includes the accuracy, specificity and sensitivity scores achieved by a linear SVM classifier. Although the performance of the SVM is the same if not marginally better than the GP-LR models, we note these are the optimal scores for the SVM classifier. For GP-LR models, the classification threshold can be tuned to optimise performance in a specific measure.

A-MCI versus AD subjects

The task of discriminating between healthy control subjects and those with a-MCI is typically much harder than discriminating between a-MCI and AD subjects e.g. a good doctor could probably distinguish the latter from other clinical and neuropsychological criteria. Our results confirm this; the worst LOOCV scores over all non-PCA models implemented achieved an AUC = 0.94, p(y) = 0.87, AC = 91%, SC = 93%, and SS = 88%. For this classification task the linear and squared exponential covariance function obtained broadly similar predictive performance results, however the linear covariance function achieved substantially stronger, approximately 50 times more probable, marginal likelihood values (logZ in the results table). Because of this we consider only the non-PCA, linear covariance function GP-LR models, amongst these the models with D = 15 or D = 20 included features obtained the strongest, and broadly similar LOOCV scores. Again with a preference for more parsimonious models we choose the model with D = 15 included features to evaluate on the test set.

In Fig. 1(b) we present the ROC curve for the a-MCI vs AD classification problem for the model with non-PCA feature normalisation, D = 15 included features and the linear ARD covariance function. The strength of the classifier is immediately apparent. For this model, tuning the classification threshold has little effect, for each of the miss-classification cost functions used the classification threshold was the same at which point the model achieved an accuracy, specificity and sensitivity score of 0.96.

An interesting observation is that the non-linear squared exponential covariance function achieves the strongest performance for the NC vs a-MCI task whereas the linear covariance seems to be the strongest for the a-MCI vs AD problem. This might be a reflection of the difficulty of the a-MCI vs AD classification problem — if the inputs for either class significantly overlap a squared exponential covariance is likely to achieve a stronger classification accuracy since it can capture small differences in the class input distributions. In the a-MCI vs AD setting, the inputs for either group may have little significant overlap and a linear decision boundary is sufficiently expressive to capture the difference between the two groups.

The optimal performance of the SVM classifier is very similar to that obtained by the GP-LR models with classification threshold of $\eta = 0.5$.

Test set results

Having selected a single GP-LR model for either classification problem, in this section we investigate each of these model's performance on a held-out test set. Since model selection was performed over a relatively small set of candidates, we expect the LOOCV results to broadly match those obtained on a held-out test set. The primary motivation for conducting these test set experiments is to obtain a reliable measure of the significance of the classifier's performance. As mentioned in the previous section, the assumptions of the binomial significance test are not valid for LOOCV accuracy scores. However, on a held-out test set the Binomial model's assumptions hold and we can obtain a reliable measure of statistical significance.

For the NC vs a-MCI classification problem we apply the GP-LR model with non-PCA feature normalisation, D = 10 included features and the squared exponential ARD covariance function. For the a-MCI vs AD classification problem we apply the GP-LR model with non-PCA feature normalisation, D = 15 included features and the linear ARD covariance function. For both classification tasks we re-train the respective models using all the validation data; for the NC vs a-MCI task $N_{trn} = 69$ and for the a-MCI vs AD task $N_{trn} = 47$.

All data preprocessing steps are applied to the entire validation dataset. Covariance hyper-parameters are optimised using conjugate gradients with a maximum of 500 line search evaluations. The test set predictive accuracy scores are measured by applying the trained models to each of the input points in the held-out test set. Results are presented in Table 3.

For the NC vs a-MCI classification task the final model achieves an AUC = 0.7. Using a classification threshold of $\eta = 0.5$ the model obtains an accuracy AC = 75%, specificity SC = 50% and sensitivity SS = 100%. Applying the Binomial test, the accuracy score is significant at the 1% level. For the a-MCI vs AD classification task the final model achieves an AUC = 0.89, accuracy AC = 80%, specificity SC = 90% and sensitivity SS = 0.7, similarly with $\eta = 0.5$. As we show in the table, all the accuracy scores can be improved if we tune the classification threshold. We expect that the low specificity score (50%) on the NC vs a-MCI task is likely

Table 3
Test set results for the optimal GP-LR models applied to the NC vs a-MCI and a-MCI vs AD classification problems.

Task	logZ	AUC	p(y)	AC(%)	SC(%)	SS(%)	AC*(%)	SC*(%)	SS*(%)	p-value
NC vs a-MCI	−26.8	0.7	0.68	75	50	100	75	70	80	6×10^{-3}
a-MCI vs AD	−4.0	0.89	0.83	80	90	70	90	90	90	1×10^{-3}

The NC vs a-MCI results are obtained using the GP-LR model with non-PCA feature normalisation, D = 10 included features and the squared exponential ARD covariance function. The a-MCI vs AD results are obtained using the GP-LR model with non-PCA feature normalisation, D = 15 included features and the linear ARD covariance function. The columns report the training data marginal likelihood logZ, the AUC, the predicted class probability p(y), the accuracy (AC), the specificity scores (SC) and the sensitivity (SS) scores averaged over the test set (AC, SS and SC calculated using $\eta = 0.5$). The final three columns (AC^*, SC^* and SS^*) report the predictive accuracy scores using a classification threshold tuned to optimise AC^*. The final column reports the p-value of the AC score using a one sided Binomial test under the null hypothesis that label predictions are uniformly random distributed p = 0.5. The assumptions of this test are valid since the same model is applied to each test point and the test data has an equal number of subjects from each group. Both models achieve statistically significant accuracy scores at the 1% level.

the result of $\eta = 0.5$ being an inappropriate threshold value for this problem — supported by the fact that the tuned specificity score $SC^* = 70\%$. However, there is insufficient test data to be able to both tune this parameter and test its performance on a separate dataset. Again, the accuracy score for this problem is significant at the 1% level. For both classification problems, as is reasonable to expect, the test set results are slightly poorer than the LOOCV results.

In summary, our methods that have been trained on one data set perform significantly well on an unseen data set, providing a robust validation of our technique.

Functional-connectivity analysis

In this section we analyse what aspects of the data the GP-LR model was using to drive classification performance. We can infer what features of the dataset were most relevant to the GP-LR model by analysing the strength of the optimised covariance function length-scale parameters ld. In Tables 4 and 5 we list each of the features used in the optimal GP-LR models in the NC vs a-MCI and a-MCI vs AD LOOCV classification tasks. Included features are ordered by the average magnitude of the corresponding ℓ_d^{-1} parameter: the larger the inverse length scale parameter is, ℓ_d^{-1}, the greater the contribution feature x_d makes to the predictive distribution.

For both covariance functions, a small length scale (or large inverse length scale) means small movements in that particular dimension can have a large effect on probability, however the interpretation of the two is subtly different due to the way the kernels act.

For the linear ARD covariance function, ℓ_d is a direct measure of how much feature d, contributes to the classification.

Table 4
A list of all the features included in all the models trained during the LOOCV NC vs a-MCI classification experiments using the D = 10 included features, non-PCA feature normalisation and the squared exponential covariance function GP-LR model.

Feature	ℓ_d^{-1}
Rolandic oper L ↔ cingulum mid L	32.5
Cingulum mid L ↔ Heschl R	18.5
Temporal sup L ↔ temporal pole Mid R	17.1
Cingulum ant R ↔ caudate L	16.7
Occipital sup R ↔ putamen R	8.5
Heschl R ↔ temporal pole sup R	5.9
Cingulum mid L ↔ temporal Inf R	3.2
Cingulum ant L ↔ caudate L	1.3
Cingulum mid R ↔ fusiform L	0.5
mmse bl	0.5
Cingulum mid R ↔ Heschl R	0.5
Rolandic oper L ↔ temporal pole sup R	0.3
Cuneus R ↔ paracentral lobule L	0.2
Cingulum mid L ↔ temporal Inf L	0.2
Occipital sup R ↔ putamen L	0.2
Occipital sup R ↔ pallidum R	0.1
Cingulum ant L ↔ pallidum R	0.0
Cingulum mid L ↔ fusiform L	0.0

Features are ranked by the average magnitude of their inverse length scale parameters ℓ_d^{-1}; the larger this value is the greater the contribution of this feature to the model.

Table 5
A list of all the features included in all the LOOCV a-MCI vs AD classification experiments using the D = 15 included features, non-PCA feature normalisation and the linear covariance function GP-LR model.

Feature	ℓ_d^{-1}
mmse bl	53,392.8
Frontal inf orb L ↔ temporal Inf R	32,874.0
Frontal inf orb L ↔ temporal inf L	12,573.6
Frontal sup orb R ↔ occipital inf R	9251.4
Frontal sup R ↔ fusiform L	6914.3
Frontal sup L ↔ occipital inf R	5426.2
Occipital inf L ↔ temporal inf R	3371.7
Frontal sup R ↔ fusiform R	2663.3
Frontal sup orb R ↔ occipital Inf L	339.5
Frontal sup medial L ↔ parietal sup R	117.1
Rectus L ↔ temporal inf L	64.4
Occipital inf R ↔ parietal sup R	57.3
Frontal mid L ↔ frontal inf orb R	49.7
Occipital inf R ↔ temporal inf L	49.3
Frontal sup orb L ↔ temporal inf R	48.6
Frontal sup orb R ↔ fusiform L	44.7
Frontal inf orb R ↔ temporal inf R	29.8
Frontal sup medial L ↔ temporal inf R	24.5
Temporal inf L ↔ temporal inf R	20.1
Occipital inf L ↔ temporal inf L	15.4
Frontal mid L ↔ occipital Inf R	13.0
Frontal sup R ↔ occipital inf L	9.8
Frontal mid R ↔ occipital inf R	7.5
Rectus L ↔ fusiform L	6.4
Rectus L ↔ temporal inf R	5.7
Frontal sup orb R ↔ temporal inf R	5.0
Frontal mid R ↔ temporal inf R	3.5
Frontal sup orb R ↔ fusiform R	3.1
Rectus R ↔ temporal inf L	2.8
Rectus L ↔ fusiform R	2.7
Frontal sup medial R ↔ temporal inf R	2.6
Frontal sup L ↔ frontal inf orb R	2.5
Frontal sup L ↔ fusiform R	2.2
Frontal sup medial R ↔ occipital inf R	1.8
Fusiform L ↔ paracentral lobule L	1.5
Frontal sup medial L ↔ occipital inf L	1.5
Frontal mid L ↔ fusiform R	1.3
Frontal sup R ↔ temporal inf L	1.3
Occipital inf L ↔ temporal mid R	1.0
Frontal sup orb L ↔ occipital inf L	1.0
Frontal sup medial L ↔ occipital inf R	0.7
Occipital inf R ↔ temporal inf R	0.7
Frontal mid R ↔ fusiform L	0.7

Features are ranked by the average magnitude of their inverse length scale parameters ℓ_d^{-1}; the larger this value is the greater the contribution of this feature to the model.

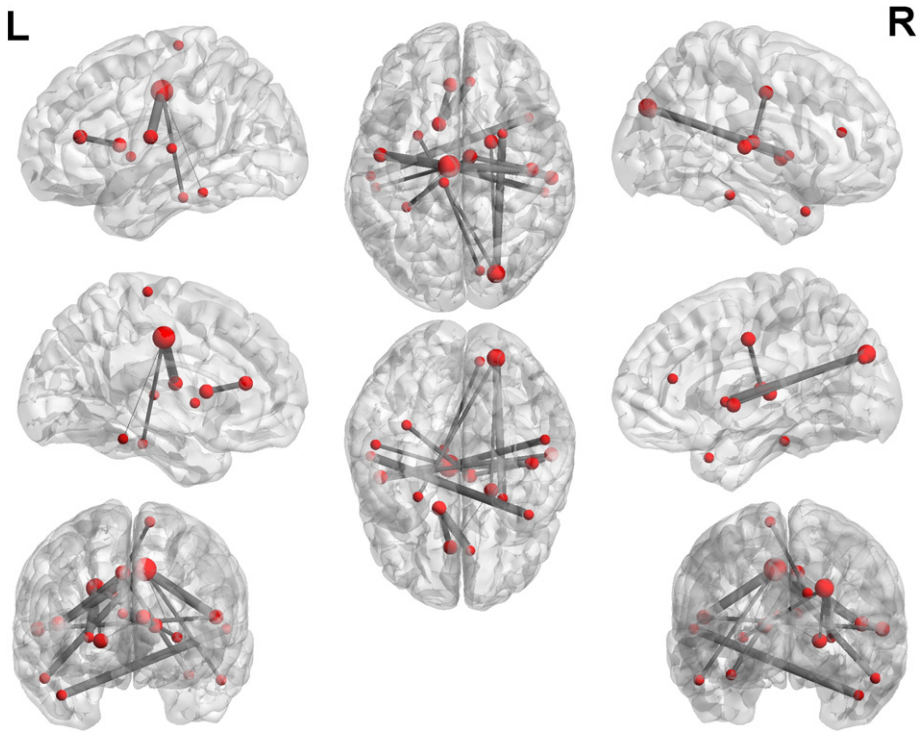

Fig. 2. Visualization of GP features included in all the models trained during the LOOCV NC vs a-MCI classification experiments using the D = 10 included features, non-PCA feature normalisation and the squared exponential covariance function GP-LR model. The graph was created using Brain-NetViewer (Xia et al., 2013, http://www.nitrc.org/projects/bnv/). The node size is proportional to the number of significant connections that particular ROI is involved with, whilst the edge thickness is proportional to the log of the average magnitude of the inverse length scale of each feature parameters (l_d). Together with these connectivity features, the MMSE score was also listed with $\ell_d^{-1} = 0.5$.

For the squared exponential covariance function, ℓ_d tells you how far you need to move for the function value to become uncorrelated i.e. if ℓ_d is large then (or small inverse length scale) then feature d is highly correlated and so changing its value will not change the function value much, and therefore not have a significant effect on probability.

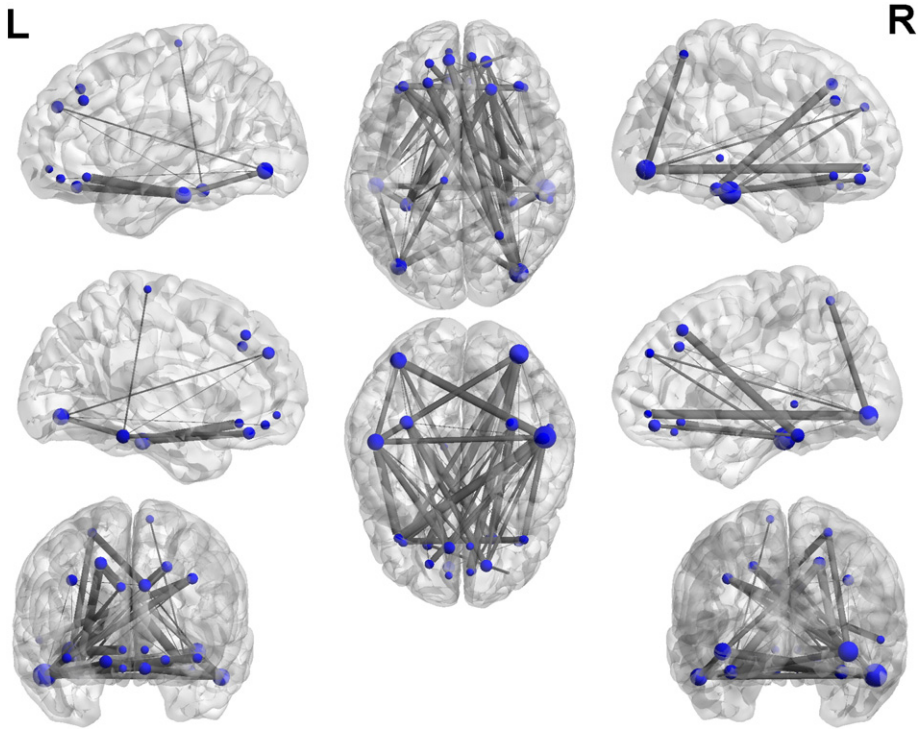

Fig. 3. Visualization of GP features included in all the models trained during the LOOCV AD vs a-MCI classification experiments using the D = 15 included features, non-PCA feature normalisation and the linear covariance function GP-LR model. The graph was created using BrainNetViewer (Xia et al., 2013), http://www.nitrc.org/projects/bnv/). The node size is proportional to the number of significant connections that particular ROI is involved with, whilst the edge thickness is proportional to the log of the average magnitude of the inverse length scale of each feature parameters (ℓ_d). Of note, the MMSE also was included, with the highest $\ell_d^{-1} = 53{,}392.8$.

To visualise the information provided by $\measuredangle_d$, we show the connectivity figures for our two classification tasks in Figs. 2 and 3. The size of nodes reflects how many connections that particular node is involved with, whilst edge thickness reflects the $\ln \measuredangle_d^{-1}$.

Consistent with previous literature, our results suggest that the loss of functional connectivity between the medial structures (cingulate cortex and cuneus) and temporal and subcortical regions of the brain can best classify patients with a-MCI. By contrast, the set of features that best classifies AD patients include the connectivity strength between frontal areas and the rest of the brain, indicating the spread of pathology typical of the disease.

Summary and discussion

We have presented a novel approach to performing patient stratification from resting state fMRI scans. The problem of patient classification in the early phases of the disease is particularly relevant in AD, as the disease is known to begin many years before the appearance of the first symptoms (Jack et al., 2013). Amnesic MCI is considered as a frequent prodromal state of AD, with an estimated conversion rate of 10–15% per year. However, only a part of these subjects eventually convert to AD. Some of them remain stable, some of them convert to other forms of dementia, whilst some of them recover to normality. The assessment of brain connectivity at rest was shown to be sensitive to AD progression (Greicius et al., 2004a) and is therefore a potentially useful non-invasive biomarker of the disease. However, its use for diagnostic or prognostic purposes at the individual level remains a challenging issue. Machine learning approaches have the potential to overcome this problem, providing a tool able to combine several features in a single classifier.

The application of machine learning classification techniques to perform patient stratification from resting state fMRI scans is not entirely novel. However, the significant majority of these previous studies have applied support vector machines (SVMs) to make classifications. One of the principle aims of our work was to show how the Gaussian process logistic regression model, a Bayesian probabilistic analogue of kernel SVMs, could also be applied to this problem. This is a relatively unexplored approach in neuroimaging studies, although Young et al. (2013) also applied it with type-II maximum likelihood to a MCI classification problem. Their goal was to predict conversion from MCI to AD using volumetric MRI, FDG-PET, cerebrospinal fluid, and APOE genotype data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database. Their study provides strong support for the Bayesian GP-LR approach employed here. Our work differs from theirs due to our novel application of the GP-LR model to rsfMRI data, aiming to make disease state predictions from the BOLD signal covariance scores between a set of predefined regions of interest.

As shown in the validation stage, the GPs performance (at a classification threshold of $\eta = 0.5$) is similar to the optimal performance of linear SVMs. We note that optimal performance of our GP classifiers can be increased by tuning the classification threshold.

Despite similar performance, the GP-LR approach has two significant advantages over kernel SVM: it provides a principled estimate of predicted class membership and a differentiable objective function by which to set hyper-parameters and make modelling decisions. The primary advantage of predicted probability estimates is that we know how confident the model is in its predictions. In the clinical setting such a value could have significant value in decision making. However, we note that to accurately estimate this value would require many more training subjects than the current study. Additionally, by tuning the classification threshold the probability estimate provided by the same GP-LR model can be set to achieve either strong specificity or sensitivity scores. This option is particularly important in the clinical context, where, depending on the question to be addressed, it might be more important to have either high specificity or high sensitivity. For example, at preclinical stages (as in MCI), differentiating normal ageing from pathological cognitive decline in large populations requires highly specific discriminators. Later on, highly sensitive discriminators might be more important to predict the time of conversion from preclinical to clinical AD.

Although one can interpret the SVM output probabilistically (e.g. Platt, 1999), this approach is ad-hoc. It does not take into account the predictive variance of $f(x)$, the discriminant function which describes distance from the hyperplane boundary (or classification boundary). Better performance can be obtained when the effect of this uncertainty is taken into account (Seeger, 2003). The Platt scaling is also known to produce meaningless results on small datasets.

Another advantage of having a differentiable objective by which to set hyper-parameters is that more complicated covariance functions can be used than kernel functions in SVMs, since numerically demanding grid search methods can be avoided. As we saw in our experiments, non-linear covariance functions with many hyper-parameters are needed in difficult classification tasks such as discriminating between healthy controls and amnesic mild cognitive impairment subjects. Additionally, as we have shown, the parameters of the optimised GP-LR model provide insights as to what features of the data are most relevant to the discrimination task; such an analysis would not be possible using a non-linear kernel SVM approach.

As is common in most machine learning neuroimaging studies, the number of training examples is small compared to the number of covariates or features describing each scan. In such a data regime there is significant risk of over-fitting any statistical model. We employed two strategies to reduce the risk of over-fitting: first we supplied the machine learning classifier with only the features that achieved the highest ranking Kendall tau correlation coefficient versus the class label, a technique previously applied in patient stratification neuroimaging studies (Shen et al., 2010a; Zeng et al., 2012); and second, we employed covariance functions with automatic relevance determination parameterisations that can automatically find a subset of the features that are most relevant to the classification task. We applied the GP-LR model to either distinguish between healthy controls and subjects exhibiting symptoms of mild cognitive impairment or distinguish between mild cognitive impairment subjects and those diagnosed with Alzheimers disease. For both these tasks, the GP-LR model achieved statistically significant classification accuracy at the 1% level as calculated on an independent held-out test set.

In this work we used only features derived from functional connectivity. Despite not including information on focal atrophy, we obtained good results in terms of classification. This must be considered in view of the recent revision of the Braak and Braak paradigm of AD evolution, which is now believed to account for a proportion of AD cases only (Lam et al., 2013). Measuring functional disconnection in a completely unbiased way opens the perspective of capturing different subcategories of AD at early stages and, possibly, other forms of dementia. In principle, however, the use of this classifier is not limited to a single type of biomarker, and it could be trained to combine the information available from many different biomarkers to provide a more accurate staging process. Future work should therefore focus on including measurements of atrophy in key areas of the brain (e.g., temporal lobe), cerebrospinal fluid assessments, β-amyloid measures using PET, and genotype information. Future work could also compare the classification of a-MCI patients against their clinical follow-up, e.g., establishing whether there is any correspondence between the a-MCI patients correctly/incorrectly classified and those who will convert to AD in a shorter/longer time.

Our study also suffers from some limitations: the sample size was relatively small, especially compared to the number of features tested. In addition, the groups were not perfectly matched for year of formal education (AD vs MCI, validation set) and gender ratio (AD vs MCI, validation set). Education is often used as a proxy of cognitive reserve (Stern, 2009), as it is a factor occurring early in life, but likely to influence life-style and life events occurring many years later. Positron-

emission tomography (Morbelli et al., 2013) and fMRI studies (Bozzali et al., 2015) of brain connectivity at rest have shown that the cognitive reserve might play a role in modulating the effect of AD pathology on functional connectivity. Nevertheless, we failed to demonstrate a similar modulation in healthy individuals (Bozzali et al., 2015), thus suggesting that our results cannot be fully explained by this effect. With respect to gender, although differences in brain connectivity between males and females have been reported (Tomasi and Volkow, 2012), such differences are not expected to be larger than those between MCI and AD patients. To support this conclusion we have run further analyses to compare the connectivity matrices between males and females within our sample, irrespective of diagnosis, and we found no significant differences (data available on request).

Another potential confound of functional connectivity analyses is motion (Satterthwaite et al., 2013). This is particularly relevant when comparing clinical populations, such as patients with AD, who are more likely to move during their scans than healthy controls. We have applied a rather simplistic approach to adjust for this confound, whilst more sophisticated approaches, such as using higher order models of motion, or independent component analysis (ICA) to classify motor components on subject by subject basis (Griffanti et al., 2014). Future work should focus on evaluating the performance of our classifier with better pre-processing. Further, it is important to evaluate the potential contribution of local grey matter loss to our findings. Patients with AD are known to develop brain atrophy in the temporal, parietal and frontal lobes. Similarly, patients with MCI tend to show more atrophy in the medio-temporal structures than healthy controls of similar age (Bozzali et al., 2006). Given the methodological approach followed here (i.e., extraction of the mean time series from pre-defined anatomical ROIs), it is possible that local atrophy has contributed, by means of partial volume effects, to our findings of decreased connectivity. Whilst more sophisticated approaches could be devised to minimise this effect, we would to reiterate that the aim of this work was to stratify patients. If the cumulative effects of atrophy and functional disconnection allows us to better classify them than functional connectivity alone after removing any possible residual effect of atrophy, it should not be regarded as a major problem. It would be interesting to compare the performance of our classifier if based on atrophy-related only features. We confine this to future work.

In addition, it should be observed that pathologies other than AD, such as major depression, can mimic the symptoms experienced by patients with MCI, and have also been shown to induce changes in functional connectivity. Distinguishing between these disorders was beyond the scope of this work, but remains an issue to be addressed before these automated classification algorithms can make their way into the clinic.

Finally, post-mortem confirmation of the diagnosis was not available for any of the participants. The classifier training and the evaluation of results rely on well-defined populations and do not account for misdiagnosis. In the elderly population, a significant proportion of the cognitively normal controls are highly likely at presymptomatic stages of the disease. Moreover, sporadic AD has a significant misdiagnosis rate, as it can be confused with other forms of dementia, unlike genetically determined diseases. Nevertheless, the neuropsychological screening was performed at one of the best dementia clinics in Italy and we are confident that it was as accurate as possible. However, for both classification tasks considered, the GP-LR model achieved statistically significant accuracy at the 1% level as calculated on an independent held-out test set.

Acknowlegments

This work was supported by the Science and Technology Facilities Council [ST/K002279/1]

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.neuroimage.2015.02.037.

References

Anderson, J., Nielsen, J., Froehlich, A., DuBray, M., Druzgal, T., Cariello, A., Cooperrider, J., Zielinski, B., Ravichandran, C., Fletcher, P., 2011. Functional connectivity magnetic resonance imaging classification of autism. Brain 134 (12), 3742–3754 (6).

Ashburner, J., Klöppel, S., 2011. Multivariate models of inter-subject anatomical variability. Neuroimage 56 (2), 422–439 (4).

Barber, D., 2012. Bayesian Reasoning and Machine Learning. Cambridge University Press (4).

Bluhm, R., Williamson, P., Lanius, R., Théberge, J., Densmore, M., Bartha, R., Neufeld, R., Osuch, E., 2009. Resting state default-mode network connectivity in early depression using a seed region-of-interest analysis: decreased connectivity with caudate nucleus. Psychiatry Clin. Neurosci. 63 (6), 754–761 (5).

Bozzali, M., Filippi, M., Magnani, G., Cercignani, M., Franceschi, M., Schiatti, E., Castiglioni, S., Mossini, R., Falautano, M., Scotti, G., Comi, G., Falini, A., 2006. The contribution of voxel-based morphometry in staging patients with mild cognitive impairment. Neurology 67 (3), 453–460 (Aug, 25).

Bozzali, M., Dowling, C., Serra, L., Spano, B., Torso, M., Marra, C., Castelli, D., Dowell, N.G., Koch, G., Caltagirone, C., Cercignani, M., 2015. The impact of cognitive reserve on brain functional connectivity in Alzheimer's disease. J. Alzheimers Dis. 44 (1), 243–250.

Braak, H., Braak, E., 1995. Staging of Alzheimer's disease-related neurofibrillary changes. Neurobiol. Aging 16 (3), 271–278 (9).

Braak, H., Braak, E., 1996. Evolution of the neuropathology of Alzheimer's disease. Acta Neurol. Scand. 94 (S165), 3–12 (9).

Buckner, R., Andrews-Hanna, J., Schacter, D., 2008. The brain's default network. Ann. N. Y. Acad. Sci. 1124 (1), 5–38 (5).

Busch, D., Hagemann, N., Bender, N., 2010. The dimensionality of the Edinburgh Handedness Inventory: an analysis with models of the item response theory. Laterality 15 (6), 610–628 (Nov, 8).

Chang, Chih-Chung, Lin, Chih-Jen, 2011. Libsvm: a library for support vector machines. ACM Trans. Intell. Syst. Technol.2157-6904 2(3). Mcbeay, pp. 27:1–27:27. http://dx.doi.org/10.1145/1961189.1961199 (URL http://doi.acm.org/10.1145/1961189.1961199. 13, 15)

Cole, D., Smith, S., Beckmann, C., 2010. Advances and pitfalls in the analysis and interpretation of resting-state FMRI data. Front. Syst. Neurosci. 4, 5.

Craddock, R., Holtzheimer, P., Hu, X., Mayberg, H., 2009. Disease state prediction from resting state functional connectivity. Magn. Reson. Med. 62 (6), 1619–1628 6, 7 33.

Cuingnet, R., Gerardin, E., Tessieras, J., Auzias, G., Lehéricy, S., Habert, M., Chupin, M., Benali, H., Colliot, O., 2011. 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 15.

De Martino, F., Gentile, F., Esposito, F., Balsi, M., Di Salle, F., Goebel, R., Formisano, E., 2007. Classification of fMRI independent components using IC fingerprints and support vector machine classifiers. Neuroimage 34 (1), 177–194 6.

Fan, Y., Liu, Y., Wu, H., Hao, Y., Liu, H., Liu, Z., Jiang, T., 2011. Discriminant analysis of functional connectivity patterns on grassmann manifold. Neuroimage 56 (4), 2058–2067 6, 7.

Fawcett, T., 2004. ROC graphs: notes and practical considerations for researchers. Mach. Learn. 31 (1–38), 14.

Friston, K., Frith, C., Liddle, P., Frackowiak, R., 1993. Functional connectivity: the principal-component analysis of large (PET) data sets. J. Cereb. Blood Flow Metab. 13 (5–5), 5.

Greicius, M., Srivastava, G., Reiss, A., Menon, V., 2004a. Default-mode network activity distinguishes Alzheimer's disease from healthy aging: evidence from functional MRI. Proc. Natl. Acad. Sci. 5, 7.

Greicius, M., Srivastava, G., Reiss, A., Menon, V., 2004b. Default-mode network activity distinguishes Alzheimer's disease from healthy aging: evidence from functional MRI. Proc. Natl. Acad. Sci. U. S. A. 101 (13), 4637–4642 (March, 19).

Griffanti, L., Salimi-Khorshidi, G., Beckmann, C.F., Auerbach, E.J., Douaud, G., Sexton, C.E., Zsol- dos, E., Ebmeier, K.P., Filippini, N., Mackay, C.E., Moeller, S., Xu, J., Yacoub, E., Baselli, G., Ugurbil, K., Miller, K.L., Smith, S.M., 2014. ICA-based artefact removal and accelerated fMRI acquisition for improved resting state network imaging. Neuroimage 95, 232–247 (Jul, 24).

Gusnard, D., Raichle, M., 2001. Searching for a baseline: functional imaging and the resting human brain. Nat. Rev. Neurosci. 2 (10), 685–694 (5).

Jack, C., Knopman, D., Jagust, W., Petersen, R., Weiner, M., Aisen, P., Shaw, L., Vemuri, P., Wiste, H., Weigand, S., Lesnick, T., Pankratz, V., Donohue, M., Trojanowski, J., 2013. Tracking pathophysiological processes in Alzheimer's disease: an updated hypothetical model of dynamic biomarkers. Lancet Neurol. 12 (2), 207–216 (February, 19).

Kendall, M., 1938. A new measure of rank correlation. Biometrika 30 (1/2), 81–93 (10).

Klöppel, S., Stonnington, C., Chu, C., Draganski, B., Scahill, R.I., Rohrer, J., Fox, N., Jack, C., Ash-burner, J., Frackowiak, R., 2008. Automatic classification of MR scans in Alzheimer's disease. Brain 131 (3), 681–689 (15).

Koch, W., Teipel, S., Mueller, S., Benninghoff, J., Wagner, M., Bokde, A., Hampel, H., Coates, U., Reiser, M., Meindl, T., 2012. Diagnostic power of default mode network resting state fMRI in the detection of Alzheimer's disease. Neurobiol. Aging 33, 466–478 (5, 6, 7, 15).

Kohavi, R., 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. International Joint Conference on, Artificial Intelligence. volume 14, pp. 1137–1145 (16).

Laakso, M., Soininen, H., Partanen, K., Lehtovirta, M., Hallikainen, M., Hänninen, T., Helkala, E., Vainio, P., Riekkinen, P., 1998. MRI of the hippocampus in Alzheimers disease: sensitivity, specificity, and analysis of the incorrectly classified subjects. Neurobiol. Aging 19 (1), 23–31 (15).

Lam, B., Masellis, M., Freedman, M., Stuss, D., Black, S., 2013. Clinical, imaging, and pathological heterogeneity of the Alzheimer's disease syndrome. Alzheimers Res. Ther. 5 (1), 1 (24).

Lemm, S., Blankertz, B., Dickhaus, T., Muller, K., 2011. Introduction to machine learning for brain imaging. Neuroimage 56 (2), 387–399 (4).

Liddle, E., Hollis, C., Batty, M., Groom, M., Totman, J., Liotti, M., Scerif, G., Liddle, P., 2011. Task-related default mode network modulation and inhibitory control in ADHD: effects of motivation and methylphenidate. J. Child Psychol. Psychiatry 52 (7), 761–771 5.

Marquand, A., Howard, M., Brammer, M., Chu, C., Coen, S., Mourao-Miranda, J., 2010. Quantitative prediction of subjective pain intensity from whole-brain fMRI data using Gaussian processes. Neuroimage 49 (3), 2178–2189 (Feb, 7).

McKhann, G., Drachman, D., Folstein, M., Katzman, R., Price, D., Stadlan, E., 1984. Clinical diagnosis of Alzheimer's disease Report of the NINCDS-ADRDA Work Group* under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease. Neurology 34 (7), 939 (8).

Meier, T., Desphande, A., Vergun, S., Nair, V., Song, J., Biswal, B., Meyerand, M., Birn, R., Prab-hakaran, V., 2012. Support vector machine classification and characterization of age-related reorganization of functional brain networks. Neuroimage 60 (1), 601–613 6, 7.

Morbelli, S., Perneczky, R., Drzezga, A., Frisoni, G.B., Caroli, A., van Berckel, B.N., Ossenkoppele, R., Guedj, E., Didic, M., Brugnolo, A., Naseri, M., Sambuceti, G., Pagani, M., Nobili, F., 2013. Metabolic networks underlying cognitive reserve in prodromal Alzheimer disease: a European Alzheimer disease consortium project. J. Nucl. Med. 54 (6), 894–902 (Jun, 24).

Pereira, F., Mitchell, T., Botvinick, M., 2009. Machine learning classifiers and fMRI: a tutorial overview. Neuroimage 45 (1), S199–S209 4.

Petersen, R., 2004. Mild cognitive impairment as a diagnostic entity. J. Intern. Med. 256 (3), 183–194 (8).

Petersen, R., Doody, R., Kurz, A., Mohs, R., Morris, J., Rabins, P., Ritchie, K., Rossor, M., Thal, L., Winblad, B., 2001. Current concepts in mild cognitive impairment. Arch. Neurol. 58 (12) (8).

Platt, John C., 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Advances in Large Margin Classifiers. MIT Press, pp. 61–74 (21).

Rasmussen, C., Nickisch, H., 2010. Gaussian processes for machine learning (gpml) toolbox. J. Mach. Learn. Res. 11, 3011–3015 (13).

Rasmussen, C., Williams, C., 2006. Gaussian Processes for Machine Learning. The MIT Press, Cambridge, MA, USA (11, 26).

Satterthwaite, T.D., Elliott, M.A., Gerraty, R.T., Ruparel, K., Loughead, J., Calkins, M.E., Eickhoff, S.B., Hakonarson, H., Gur, R.C., Gur, R.E., Wolf, D.H., 2013. An improved framework for confound regression and filtering for control of motion artifact in the preprocessing of resting-state functional connectivity data. Neuroimage 64, 240–256 (Jan, 24).

Seeger, Matthias, 2003. Bayesian Gaussian Process Models: Pac-Bayesian Generalisation Error Bounds and Sparse Approximations (22).

Sheline, Y., Barch, D., Price, J., Rundle, M., Vaishnavi, S., Snyder, A., Mintun, M., Wang, S., Coalson, R., Raichle, M., 2009. The default mode network and self-referential processes in depression. Proc. Natl. Acad. Sci. 106 (6), 1942–1947 (5).

Shen, H., Wang, L., Liu, Y., Hu, D., 2010a. Discriminative analysis of resting-state functional connectivity patterns of schizophrenia using low dimensional embedding of fMRI. Neuroimage 49 (4), 3110 (10, 22).

Shen, Hui, Wang, Lubin, Liu, Yadong, Hu, Dewen, 2010b. Discriminative analysis of resting-state functional connectivity patterns of schizophrenia using low dimensional embedding of fMRI. NeuroImage (ISSN: 1053-8119) 49 (4), 3110–3121. http://dx.doi.org/10.1016/j.neuroimage.2009.11.011 (URL http://www.sciencedirect.com/science/article/pii/S1053811909011951).

Stern, Y., 2009. Cognitive reserve. Neuropsychologia 47 (10), 2015–2028 (Aug, 24).

Tohka, J., Foerde, K., Aron, A., Tom, S., Toga, A., Poldrack, R., 2008. Automatic independent component labeling for artifact removal in fMRI. Neuroimage 39 (3), 1227–1245 (6).

Tomasi, D., Volkow, N.D., 2012. Gender differences in brain functional connectivity density. Hum. Brain Mapp. 33 (4), 849–860 (Apr, 24).

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B., Joliot, M., 2002. Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. Neuroimage 15 (1), 273–289 9.

Uddin, L., Kelly, A., Biswal, B., Margulies, D., Shehzad, Z., Shaw, D., Ghaffari, M., Rotrosen, J., Adler, L., Castellanos, F., 2008. Network homogeneity reveals decreased integrity of default-mode network in ADHD. J. Neurosci. Methods 169 (1), 249–254 5.

Van Den Heuvel, M., Hulshoff Pol, H., 2010. Exploring the brain network: a review on resting-state fMRI functional connectivity. Eur. Neuropsychopharmacol. 20 (8), 519–534 (5).

Venkataraman, A., Kubicki, M., Westin, C.F., Golland, P., 2010. Robust Feature Selection in Resting-State fMRI Connectivity Based on Population Studies. pp. 63–70 (6).

Venkataraman, Archana, Whitford, Thomas J., Westin, Carl-Fredrik, Golland, Polina, Kubicki, Marek, 2012. Whole brain resting state functional connectivity abnormalities in schizophrenia. Schizophr. Res. (ISSN: 0920-9964) 139 (13), 7–12. http://dx.doi.org/10.1016/j.schres.2012.04.021 (URL http://www.sciencedirect.com/science/article/pii/S0920996412002538).

Xia, M., Wang, J., He, Y., 2013. BrainNet viewer: a network visualization tool for human brain connectomics. PLoS ONE 8, 68910. http://dx.doi.org/10.1371/journal.pone.0068910 (July, 20, 21).

Young, J., Modat, M., Cardoso, M., Mendelson, A., Cash, D., Ourselin, S., Alzheimer's Disease Neuroimaging Initiative, 2013. Accurate multimodal probabilistic prediction of conversion to Alzheimer's disease in patients with mild cognitive impairment. Neuroimage Clin. 2, 735–745 (20).

Zeng, L., Shen, H., Liu, L., Wang, L., Li, B., Fang, P., Zhou, Z., 2012. Identifying major depression using whole-brain functional connectivity: a multivariate pattern analysis. Brain 135, 1498–1507 (5, 6, 7, 10, 22).

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 (7, 15).

Zhou, D., Thompson, W., Siegle, G., 2009. Matlab toolbox for functional connectivity. Neuroimage 47 (4), 1590–1607 6.