doi:10.1093/brain/aws084 Brain 2012: 135; 1508–1521 | 1508

BRAIN
A JOURNAL OF NEUROLOGY

Multi-centre diagnostic classification of individual structural neuroimaging scans from patients with major depressive disorder

Benson Mwangi,[1] Klaus P. Ebmeier,[2] Keith Matthews[1] and J. Douglas Steele[1]

1 Division of Neuroscience, Medical Research Institute, University of Dundee, Level 5, Ninewells Hospital and Medical School, Dundee, DD1 9SY, UK
2 Department of Psychiatry University of Oxford Warneford Hospital, Oxford, OX3 7JX, UK

Correspondence to: Benson Mwangi,
Division of Neuroscience,
Medical Research Institute,
University of Dundee,
Level 5, Ninewells Hospital and Medical School,
Dundee, DD1 9SY, UK
E-mail: b.m.irungu@dundee.ac.uk

Quantitative abnormalities of brain structure in patients with major depressive disorder have been reported at a group level for decades. However, these structural differences appear subtle in comparison with conventional radiologically defined abnormalities, with considerable inter-subject variability. Consequently, it has not been possible to readily identify scans from patients with major depressive disorder at an individual level. Recently, machine learning techniques such as relevance vector machines and support vector machines have been applied to predictive classification of individual scans with variable success. Here we describe a novel hybrid method, which combines machine learning with feature selection and characterization, with the latter aimed at maximizing the accuracy of machine learning prediction. The method was tested using a multi-centre dataset of T_1-weighted 'structural' scans. A total of 62 patients with major depressive disorder and matched controls were recruited from referred secondary care clinical populations in Aberdeen and Edinburgh, UK. The generalization ability and predictive accuracy of the classifiers was tested using data left out of the training process. High prediction accuracy was achieved (~90%). While feature selection was important for maximizing high predictive accuracy with machine learning, feature characterization contributed only a modest improvement to relevance vector machine-based prediction (~5%). Notably, while the only information provided for training the classifiers was T_1-weighted scans plus a categorical label (major depressive disorder versus controls), both relevance vector machine and support vector machine 'weighting factors' (used for making predictions) correlated strongly with subjective ratings of illness severity. These results indicate that machine learning techniques have the potential to inform clinical practice and research, as they can make accurate predictions about brain scan data from individual subjects. Furthermore, machine learning weighting factors may reflect an objective biomarker of major depressive disorder illness severity, based on abnormalities of brain structure.

Keywords: major depressive disorder; predictive classification; relevance vector machines; support vector machines; feature based morphometry

Received September 30, 2011. Revised February 8, 2012. Accepted February 10, 2012

Abbreviations: 3D-SIFT = 3D scale invariant feature transform; BDI = Beck Depression Inventory; FBM = feature-based morphometry; HRSD = Hamilton Rating scale for Depression; RVM = relevance vector machine; SVM = support vector machine; VBM = voxel-based morphometry

Introduction

Major depressive disorder is characterized by abnormalities of mood, cognition and psychomotor function (Sobin and Sackeim, 1997). Many neuroimaging and post-mortem studies have reported group level anatomical brain volume reductions in patients with major depressive disorder (Shah *et al*., 1998; Fava and Kendler, 2000; Costafreda *et al*., 2009; Koolschijn *et al*., 2009). Volumetric reductions of the hippocampus, basal ganglia, subgenual prefrontal cortex and orbitofrontal cortex, particularly in patients with multiple episodes and longer durations of illness, have been repeatedly reported (Lorenzetti *et al*., 2009). The largest volume reductions are typically in the anterior cingulate and orbitofrontal cortex, with smaller reductions in the hippocampus, putamen and caudate nucleus (Koolschijn *et al*., 2009). Nevertheless, the diagnosis of major depressive disorder remains entirely clinical, based on history, symptoms and signs. Currently, there are no clinically useful biomarkers for the diagnosis of major depressive disorder, or for the prediction of treatment response, although there is active work in this area (Miller *et al*., 2009; Belzeaux *et al*., 2010; Schmidt *et al*., 2011; Schneider *et al*., 2011). Identification of biomarkers is of considerable interest as these could, ultimately, inform clinical decisions with individual patients. Importantly, they may also support improved diagnostic classification systems and provide a focus for empirical research into pathophysiological and therapeutic mechanisms. An important first step is being able to reliably discriminate scans from individuals with major depressive disorder from matched controls.

Clinically it is recognized that traditional qualitative radiological methods are not able to distinguish neuroanatomical scans of patients with major depressive disorder from healthy subjects. While not pathognomonic for major depressive disorder, white matter hyperintensities may be a predictor of depression in elderly patients (Godin *et al*., 2008). While low dimensional measurements, e.g. per-subject average grey matter probabilities from selected brain regions identified using voxel-based morphometry (VBM) can be significant in a group level analysis, attempting to classify individual novel data results in high classification errors. In order to build a high accuracy predictive model it is essential to use a technique that combines many measurements (Lao *et al*., 2004). Such computer-based learning techniques include relevance vector machines (RVM) and support vector machines (SVM), which have recently been applied to clinical neuroimaging research, with the aim of training classifiers that can reliably distinguish different clinical groups at an individual subject level, e.g. depressive illness (Fu *et al*., 2008; Costafreda *et al*., 2009; Gong *et al*., 2011), psychosis (Koutsouleris *et al*., 2009), Alzheimer's disease (Davatzikos *et al*., 2008; Kloppel *et al*., 2008; Magnin *et al*., 2009) and prediction of presymptomatic Huntington's disease (Kloppel *et al*., 2009).

Typically, this approach has two stages. First, image data from each subject is entered into a classifier together with a diagnostic label (e.g. major depressive disorder versus control) and the system learns to classify the data. Secondly, the accuracy of the trained classifier is estimated with additional image data not used for training. Relatively high accuracies of 89% have been reported when using structural MRI for discriminating Alzheimer's disease from normal ageing (Kloppel *et al*., 2008), >86% for predicting future psychoses versus controls (Koutsouleris *et al*., 2009), using functional MRI, 86% accuracy in discriminating depressed 'treatment responders' from 'non-responders' (Fu *et al*., 2008), although a much lower accuracy of 67% was reported for structural images discriminating depressed patients from controls (Fu *et al*., 2008). Major depressive disorder classification based on structural MRI data is of particular interest as these scans are easier, quicker and cheaper to obtain routinely than functional MRI data, which introduces complexities relating to the choice and design of the behavioural paradigm, plus comprehension and cooperation of subjects with the task. Depression can be associated with cognitive impairments (Austin *et al*., 2001) and some very ill patients may be unable to concentrate on a paradigm. In such cases, 'structural' scanning, which only requires patients to remain still, may still be a practical way to obtain brain image data.

Acquisition of multi-centre data in brain imaging is increasingly recognized as a desirable and feasible approach to facilitate recruitment and to increase study power, provided increased power outweighs the cost of between-scanner related variance (Moorhead *et al*., 2009; Gradin *et al*., 2010; Suckling *et al*., 2010). Additionally, multi-centre studies allow testing whether a given classification performance can be generalized, not just to different subjects scanned on the same scanner, but also to different subjects scanned on different scanners. This adds to the reliability of the results obtained (Ashburner, 2009). To our knowledge, no diagnostic classifier has yet been applied successfully to multi-centre scanner and clinical subject data, to discriminate patients with major depressive disorder from healthy matched volunteers.

Here T_1-weighted magnetic resonance images of brain structure were acquired from two imaging centres (one scanner per centre): University of Aberdeen (Cohort A) and University of Edinburgh (Cohort B) from different locally recruited cohorts of patients with major depressive disorder of at least moderate severity and matched healthy controls. Functional MRI data were also acquired and reported (Steele *et al*., 2007; Kumar *et al*., 2008).

The primary objective was to train a high accuracy predictive classifier that could discriminate scans from participants with major depressive disorder from scans obtained from controls. A second objective was to estimate the probability of the subject belonging to the predicted class, to determine the confidence of the prediction. Third, we investigated whether the SVM and RVM derived

Table 1 Demographic and clinical details

	Cohort A (Aberdeen)		Cohort B (Edinburgh)		Significance, *P*-value
	Patients mean (SD)	Controls mean (SD)	Patients mean (SD)	Controls mean (SD)	
Age (years)	46.1 (12.5)	40.6 (10.3)	44.7 (10.0)	43.0 (13.2)	0.60[a,*]
Females/total	9/15	11/18	10/15	7/14	0.41[b,*]
National Adult Reading Test	111.6 (8.4)	114.3 (8.6)	115.7 (5.1)	117.7 (5.7)	0.566[a,*]
BDI	22.9 (8.2)	3.1 (2.6)	38.0 (9.2)	1.1 (1.7)	<0.001[c,d] <0.001[d,e] <0.001[d,f] 0.015[d,g]
Spielberger State Anxiety Score	54.6 (11.5)	30.4 (8.2)	62.0 (6.8)	28.8 (9.0)	<0.001[c,d] 0.43[d,e,*] <0.001[d,f] 0.613[d,g,*]
Snaith–Hamilton	35.0 (6.8)	51.4 (4.0)	32.1 (6.3)	52.7 (4.2)	<0.001[c,d] 0.24[d,e,*] <0.001[d,f] 0.368[d,g,*]
HRSD	23.2 (4.3)		27.87 (5.8)		0.02[d]

Tests for differences between groups (two-tailed):
a Four group ANOVA.
b Chi-square.
c Centre A patients versus controls.
d two group independent *t*-test.
e Centre B patients versus controls.
f Patients versus patients.
g Controls versus controls.
*No significant difference between groups (significance defined as $P < 0.05$).

weighting factors (a single weighting factor being calculated for a single subject) correlated with illness rating measures. This was to investigate whether the predictive models (SVM or RVM) achieved accurate predictions on novel neuroimaging data using clinically meaningful measures.

Materials and methods

Sixty-two participants were recruited in total; 30 patients with major depressive disorder and 32 controls. Approval was obtained from Grampian and Lothian NHS Research Ethics Committees and written informed consent obtained from all participants. In both centres, patients met criteria for DSM-IV (Diagnostic and Statistical Manual, 4th edition) major depressive disorder without comorbidity. Preference was given to the recruitment of patients with continuing illness despite active treatment, as such patients are commonly referred to secondary care psychiatric outpatient services in the UK and structural brain changes may be more likely in treatment unresponsive patients (Ebmeier *et al*., 2006; Lorenzetti *et al*., 2009). All patients with major depressive disorder had a minimum duration of illness >3 months despite compliance with continuing antidepressant drug treatment, with a first episode of major depressive disorder typically diagnosed at least 5 years before recruitment.

In many cases it was difficult to determine with certainty the number of previous episodes of illness as residual symptoms tended to persist. All medications were stable for at least 1 month before scanning and, for recruitment, a minimum illness severity rating of 21 on the Hamilton Depression Rating Scale (HRSD) was required on initial assessment. Exclusion criteria were any other psychiatric diagnosis including personality disorder, a history of substance misuse, gross structural brain abnormality observable radiologically, use of non-antidepressant and related medication and neurological disorder. Patients were also excluded if they were likely to be intolerant of scanning (e.g. were agitated or had a history of claustrophobia). Control participants were matched on the basis of age, gender and premorbid intelligence as assessed by the National Adult Reading Test (Nelson and Wilson, 1991), with the same exclusion criteria as patients with major depressive disorder. There were no significant differences in ages between groups as shown in Table 1.

All subjects completed self-ratings of depression [Beck Depression Inventory (BDI)], anxiety (Spielberger state-trait anxiety inventory) and hedonic status (Snaith–Hamilton Pleasure Scale). HRSD ratings were obtained by the same experienced clinician (J.D.S.) immediately before scanning, which was always done in the morning. Centre A (Aberdeen) participants consisted of 18 controls and 15 patients with major depressive disorder. Centre B (Edinburgh) participants consisted of 15 patients with major depressive disorder and 14 controls. Patients were receiving a wide variety of antidepressant medication reflecting contemporary clinical practice. Supplementary Table 4 provides further detailed information.

Illness severity rating scales

Long established rating scales used to assess the severity of major depressive disorder symptoms include the BDI (Beck *et al*., 1961) and HRSD (Hamilton, 1960). The BDI and HRSD rating scales are questionnaires used to measure the severity of depression. The questionnaires are composed of specific items relating to symptoms of major depression such as hopelessness, irritability, guilt, low mood, agitation and suicidal ideation. The BDI is self-rated while HRSD is rated by a clinician and in both cases, higher scores indicate increased severity of major depressive disorder.

Major depressive disorder may be considered to have two core dimensions; anhedonia and agitation-anxiety. These were rated by the Snaith–Hamilton pleasure scale (Snaith *et al*., 1995) and Spielberger State Anxiety Inventory (Kvaal *et al*., 2005), respectively. The Snaith–Hamilton Pleasure Scale is a self-rating scale used to measure anhedonia, the absence of which is anhedonia (a loss of ability to experience pleasure). Lower Snaith–Hamilton Pleasure Scale scores are associated with increased depression severity (Nakonezny *et al*., 2010). The Spielberger state-trait anxiety inventory scale is a measure of state anxiety with higher scores reflecting higher levels of anxiety. Major depressive disorder is frequently associated with higher anxiety.

Image acquisition and preprocessing

The same manufacturer and field strength of scanner was used in both centres; 1.5 T GE Medical Systems Sigma. The T_1 acquisition parameters for Centre A were: repetition time 20 ms, echo time 6 ms, flip angle 35°, 124 contiguous 1.6 mm axial slices of 256 × 256 voxels with an in-plane resolution of 0.938 mm^2. Centre B: repetition time 20 ms, echo time 6 ms, flip angle 15°, 124 contiguous 1.7 mm axial slices of 256 × 256 voxels with an in-plane resolution of 0.938 mm^2.

T_1-weighted magnetic resonance images were visually inspected for artefacts and then preprocessed using the Diffeomorphic Anatomical Registration using the Exponential Lie Algebra (DARTEL) toolbox (Ashburner, 2007) as implemented in SPM8 (http://www.fil.ion.ucl.ac.uk/spm). The DARTEL approach involves creation of a study-specific template and segmentation of grey matter with modulation to control rescaling effects. Segmented grey matter images were smoothed with a 6-mm full-width half-maximum Gaussian kernel.

The data were then randomly divided into two sets of 31 subjects ('Split1' and 'Split2') with approximately equal numbers of controls and patients from each centre. One 'split' was used for training, the other for testing. The same data were not used for training and testing to avoid 'double dipping' (Kriegeskorte *et al*., 2009). The feature selection and classifier training image processing 'pipelines' were therefore performed twice; first using Split 1 as the training set and Split 2 as the testing set, then using Split 2 as the training set and Split 1 as the testing set. The results were collated to evaluate the performance of the classifier on the whole data set.

Machine learning

RVM (Tipping, 2001) and SVM (Vapnik, 1998) are machine learning methods used for making predictions about novel data given labelled example training data. Given a training set $\{X_n, t_n\}_{n=1}^{N}$, where X_n represents example training data and t_n diagnostic labels (major depressive disorder or control), the object was to train a model that was able to make accurate predictions of t for novel imaging data X (during the testing stage) to which the predictive classifier was never previously exposed to (during the training stage). The following sections provide a very general summary of both methods, also summarized in Figs. 1 and 3. More detail is provided in Appendix I and the Supplementary material.

Support vector machine

Given example training data (brain imaging data) of two groups with respective diagnostic labels (major depressive disorder or control), the SVM learning process identifies a boundary (hyperplane) that optimally separates the training example data into the two labelled groups (Fig. 1). This boundary sets a maximum margin 'd' between the two labelled classes and is later used during the testing stage to categorize

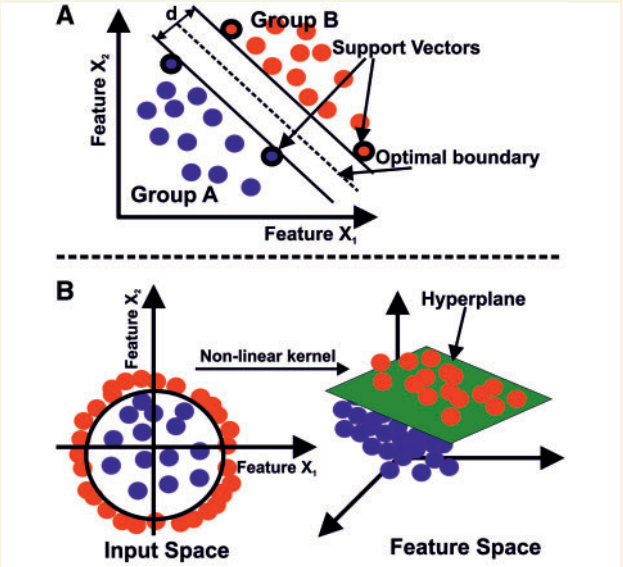

Figure 1 Two dimensional representation of the SVM algorithm. Each point represents a feature vector from a single subject. (**A**) SVM algorithm identifies a hyperplane boundary maximally (maximizing 'd') separating the two groups. (**B**) Feature vectors may not always be separable by a linear boundary so a kernel function can be used to map feature vectors into a non-linear feature space, facilitating separation.

novel scan data (major depressive disorder versus control). SVM is based on the concept of 'structural risk minimization' (Vapnik, 1998), which means that it aims to find a boundary that maximizes the distance between the two classes (major depressive disorder and controls), simultaneously minimizing the misclassification of data (scans).

Training stage data from individual subjects that are closest to the hyperplane are termed 'support vectors'. Notably, individual subjects (major depressive disorder or control) are categorized by SVM as contributing either support vector imaging data or non-support vector imaging data, with only the former being given weights by the SVM algorithm (used for making predictions).

An SVM model requires two parameters to be identified: a 'kernel' parameter and a 'regularization parameter'. The kernel allows the SVM model to calculate similarities between the training data sets (Bishop, 2007) and the kernel parameter controls how the similarities are calculated. Different types of kernel are possible and in the present study a non-linear Gaussian kernel was used. The regularization parameter allows the model to define a maximal margin between the two labelled classes, minimizing misclassification. The kernel and regularization parameters were identified using a 'grid search' method within a 'leave-one-out' cross-validation procedure during training (Theodoridis and Koutroumbas, 2006).

Relevance vector machine

Unlike SVM, RVM utilizes a probabilistic learning framework to make predictions (Tipping, 2001). The model estimates a probability $p(t|X)$, which is the probability of the class label (major depressive disorder versus controls) given novel data X. The RVM training procedure estimates a posterior distribution of weights, which describes the

'importance' of each training example data to the trained model. Specifically, the most important training example data (termed 'relevance vectors', analogous to 'support vectors' in SVM) are given weights for calculating predictions for novel data. Other less important training stage data are 'pruned' with their corresponding weights set to zero. The biggest difference between SVM and RVM is that predictions from the latter are probabilistic (e.g. the probability of a scan being from a subject with major depressive disorder or a control). Similar to SVM, RVM requires selection of a kernel function and a kernel width parameter. Here a non-linear Gaussian kernel was used with the kernel width parameter determined using 'leave-one-out' cross-validation of the training data set.

Previous neuroimaging studies have used SVM for predictive classification; however, RVM has some advantages. First, SVM does not make probabilistic predictions and therefore does not provide the level of confidence or uncertainty on individual predictions, which can be useful in some contexts. Second, SVM requires a separate estimation of the regularization parameter, while RVM estimates such 'hyperparameters' directly. Additionally, RVM is a relatively sparse algorithm, meaning it uses only a small percentage of the training data for making predictions. In contrast SVM uses many more training data for making predictions (Bishop, 2007). This makes RVM more computationally efficient, but conversely less useful for investigation of the relationship between weighting factors and clinical ratings of illness severity (as there are far fewer subjects with weighting factors).

For a more detailed mathematical introduction to both SVM and RVM see Appendix I and other texts (Vapnik, 1998; Tipping, 2001; Bishop, 2007). Custom Matlab (The Mathworks Inc.) software was written to implement the above calculations based on SVM (Schwaighofer, 2001) and RVM (Tipping, 2001) Matlab toolboxes.

Evaluation of support vector machine and relevance vector machine prediction performance

Both SVM and RVM models were evaluated using a 2×2 'confusion matrix', obtained from the classifier testing stage, and used to calculate accuracy, specificity, sensitivity and chi-square *P*-value, with the latter testing the null hypothesis of no significant classification. A receiver operating characteristic curve was also generated.

Relevance vector machine and support vector machine weighting factor analyses

Correlations between weighting factors (SVM and RVM) and illness severity ratings (BDI, Snaith–Hamilton Pleasure Scale and Spielberger state-trait anxiety inventory) were calculated. This was to explore whether the weighting factors reflected a clinically meaningful index of illness severity.

As above, the SVM and RVM algorithms set the weighting factors to zero for subjects designated as not contributing 'support' (or 'relevance') vector imaging data. This meant that the relationship between weights and illness scores could not be explored for some subjects. We therefore addressed this by inferring weighting factors for the SVM calculation.

Inference was done by identifying which support vectors were most similar to the non-support vectors, by calculating a pairwise L_1 norm distance (Appendix I) between a non-support vector and all support vectors. Non-support vector subjects were assigned weights from the most similar support vector weights; 'inferred weights'. Correlations between the inferred weights and corresponding actual illness severity ratings were then calculated.

The clinical characteristics (BDI, Snaith–Hamilton Pleasure Scale, Spielberger state-trait anxiety inventory) of patients and controls categorized as providing support vector imaging data were compared with subjects not categorized as providing support vector imaging data. This was to test the hypothesis that non-support vector subjects had more clinically intermediate characteristics (e.g. patients with lowest illness severity ratings, controls with highest BDI scores).

Enhancement of support vector machine and relevance vector machine performance using feature selection

A typical T_1-weighted scan contains many more voxels than numbers of subjects in a neuroimaging study, and neuroimaging data include 'noise' (random variation) and voxels that are redundant for making predictions. Consequently, it was important to preselect the most important brain regions for making accurate predictions. This was done using feature selection.

Feature selection is a process that preselects brain regions that help to distinguish different groups (e.g. major depressive disorder versus controls). Machine learning studies have demonstrated that feature selection is important for achieving a high predictive accuracy classifier (Guyon and Elisseeff, 2003). Importantly, feature selection is only performed using the training data. Once identified, the same brain regions identified during training are used for testing the classifier predictive accuracy.

We used a filter method for feature selection (Saeys *et al*., 2007; Hua *et al*., 2009). Specifically, using VBM (Ashburner and Friston, 2000) as implemented in SPM8, we performed a two sample *t*-test during training to identify the voxels that differed most in major depressive disorder versus controls. An optimal threshold of $P < 0.0009$ (uncorrected, extent threshold 0 voxels) was identified by empirically varying the threshold using a cross-validation procedure with the training data. The optimal threshold corresponded to the peak training accuracy as shown in Fig. 2.

As discussed in the Supplementary material, a multivariate 'Wrapper' feature selection method, Recursive Feature Elimination (Guyon and Elisseeff, 2003; Saeys *et al*., 2007; Martino *et al*., 2008; Hua *et al*., 2009) was also explored. However, this was not associated with as high a prediction accuracy as the method described here.

Enhancement of support vector machine and relevance vector machine performance using feature description

If as is usually the case, perfect predictive classification accuracy cannot be achieved using feature selection with machine learning, feature description [also referred to here as feature-based morphometry (FBM)] may enhance accuracy (Tommasi *et al*., 2008). The raw data, which in the present study consisted of grey matter probabilities at each selected voxel, were transformed into a different representation that aimed to enhance classifier performance. The FBM method used was a combination of two methods: the 3D Scale Invariant Feature Transform (3D-SIFT) descriptor (Scovanner *et al*., 2007) and the 'Bag of Words' method.

First, grey matter voxel probabilities calculated using DARTEL, from brain regions identified during feature selection, were transformed

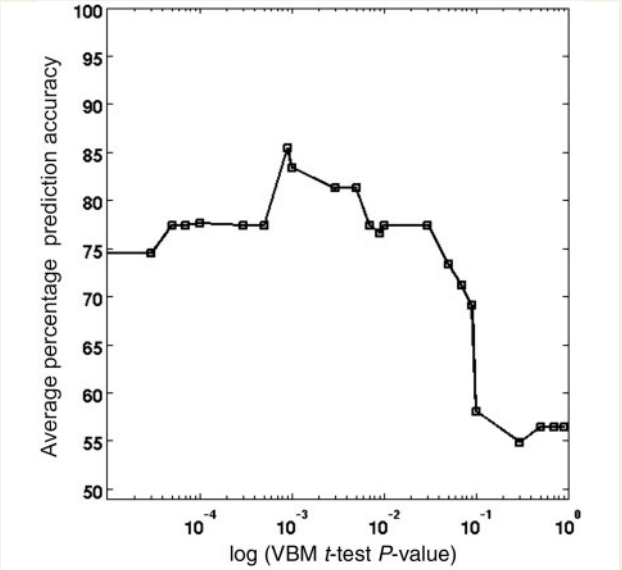

Figure 2 Curve showing the average percentage predictive accuracy of the two splits of data versus VBM *t*-test *P*-value threshold. Maximal predictive accuracy during the training stage occurred at an optimum threshold of $P < 0.0009$.

using the 3D-SIFT process. This transformation consisted of representing the grey matter probabilities as changes (gradients) of local grey matter probability, together with the anatomical orientations of the gradients (Appendix I and Supplementary material).

Second, as feature selection identified a number of distributed brain regions for each subject, each brain region had an associated 3D-SIFT descriptor, so it was necessary to combine all the descriptors from a single subject to form a single overall anatomical feature vector descriptor. This combination of descriptors was done using the 'Bag of Words' approach (Appendix I and Supplementary material). This method is popular in computer vision research (Fei-Fei and Pietro, 2005; Nielsen *et al*., 2005; Hardoon *et al*., 2007; Tommasi *et al*., 2008), neuroimaging (Nielsen *et al*., 2005; Hardoon *et al*., 2007; Toews *et al*., 2009, 2010; Castellani *et al*., 2012), medical imaging (Tommasi *et al*., 2008) and gene expression classification (Ji *et al*., 2009).

A summary of the combined FBM-FBM-SVM procedure and analogous VBM-FBM-RVM procedure is shown in Fig. 3.

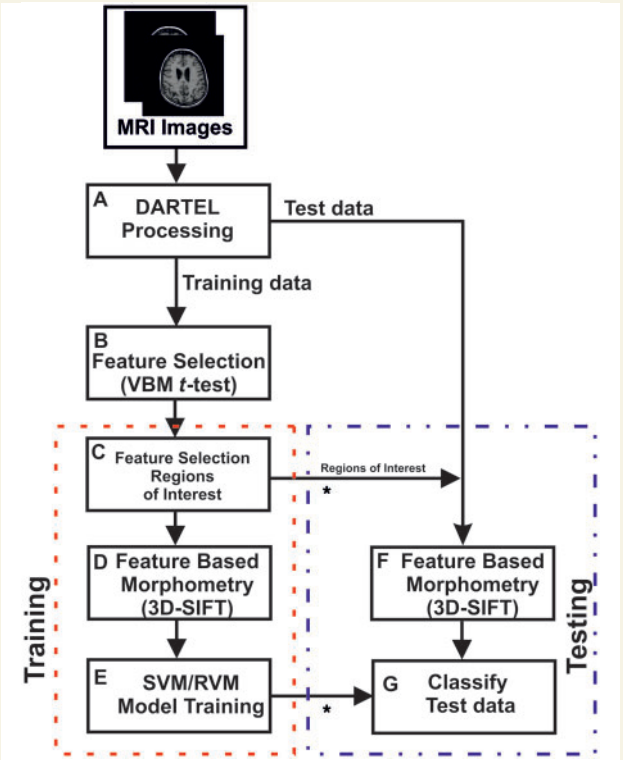

Figure 3 Flow diagram illustrating the combined VBM-FBM-RVM and equivalent SVM based procedure. *Information transferred to testing stage only after training complete.

Results

Table 1 summarizes sociodemographic and clinical details of subjects. There were no significant differences between groups with respect to age, gender ratio and estimated premorbid IQ. As expected, patients with major depressive disorder were rated as more depressed, anhedonic and anxious than controls, for each centre.

Anatomical feature selection

Tables 2 and 3 summarize brain regions that differed significantly between controls and patients with major depressive disorder (in both 'splits' of data, *t*-tests were performed separately). Figure 4 shows an average feature selection map from a VBM *t*-test for both splits. Grey matter reductions in major depressive disorder were identified in the dorsolateral prefrontal cortex, medial frontal cortex, orbitofrontal cortex, temporal lobe, insula, cerebellum and posterior lobe as shown in Fig. 4. No increases in major depressive disorder grey matter were identified.

Table 2 Regions of grey matter reductions in patients with major depressive disorder versus controls from Split 1 as training data

Anatomical region	Brodmann area	MNI coordinates (x, y, z)			VBM Z-value
Middle frontal gyrus	9	26	26	48	3.76
Superior parietal lobule	7	−32	−60	52	6.14
Inferior frontal gyrus	47	28	22	−26	4.31
Superior temporal gyrus	22	−62	−30	2	4.09
Superior temporal gyrus	22	−58	−6	0	3.86
Superior frontal gyrus	9	−26	38	30	3.76
Superior frontal gyrus	9	14	52	34	3.32

Individual scan classification

The combined VBM-FBM-RVM technique (Fig. 3) resulted in the best performance: accuracy 90.3%, specificity 87.5% and sensitivity 93.3%, chi-square $P < 1 \times 10^{-7}$. Figure 5 shows a receiver

operating characteristic curve with an area under curve totalling 0.904 with an *F*-measure of 0.903. This indicates excellent classifier performance across three performance metrics: accuracy, area under curve and *F*-measure. Figure 6 shows bar graphs depicting the distributions of predicted probabilities of classification for patient and control subjects. Notably, probabilities predicted for controls are skewed towards zero, while those from patients are skewed towards unity, indicating that predicted probabilities were typically well separated from 'indecision' (0.5). Subject specific probabilities are provided in Supplementary Table 1.

Table 3 Regions of grey matter reductions in patients with major depressive disorder versus controls from Split 2 as training data

Anatomical region	Brodmann area	MNI coordinates (x, y, z)			VBM Z-value
Superior parietal lobule	7	−32	−60	52	6.14
Superior frontal gyrus	9	20	40	42	3.52
Inferior frontal gyrus	47	28	22	−26	4.31
Superior temporal gyrus	22	−62	−30	2	4.09
Orbital gyrus	24	−2	38	−24	3.14
Superior frontal gyrus	9	−26	44	20	3.76
Superior frontal gyrus	9	6	58	32	3.29

The combined VBM-FBM-SVM technique (Fig. 3) achieved almost as good performance as the corresponding RVM method: accuracy 87.1%, specificity 87.5% and sensitivity 86.67%, chi-square $P < 1 \times 10^{-7}$.

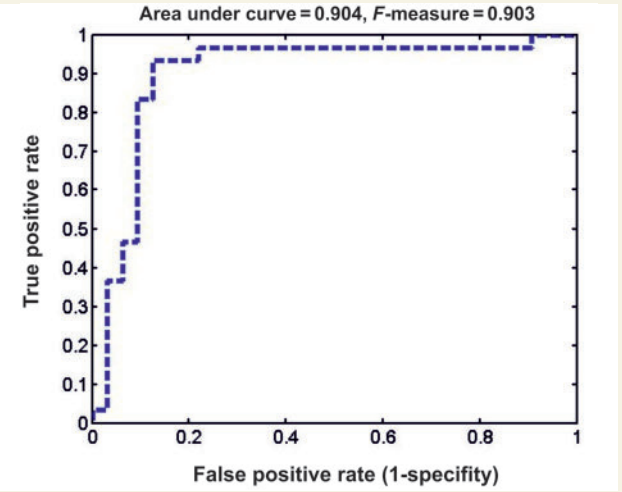

Figure 5 Receiver operating characteristic curve for VBM-FBM-RVM predictive classification.

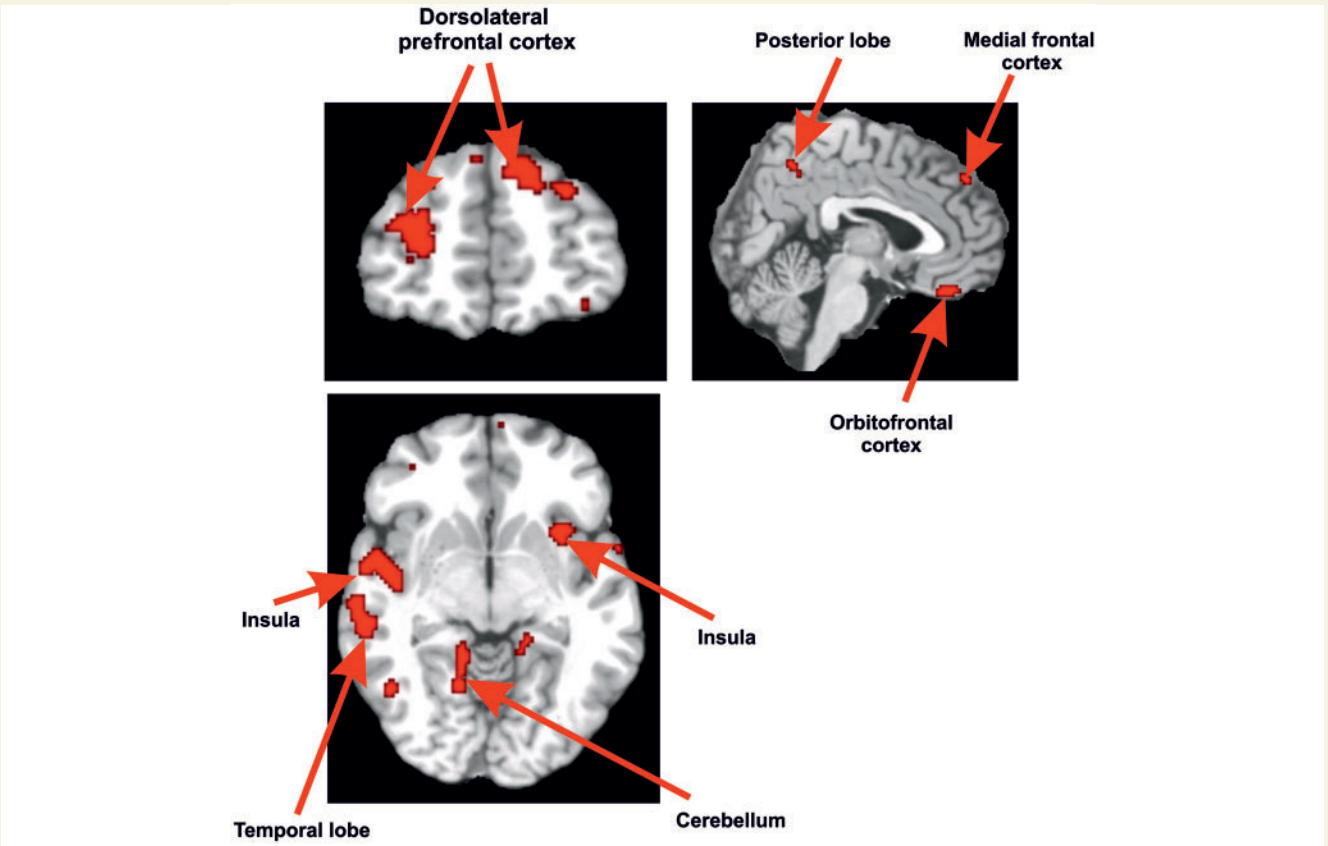

Figure 4 Brain regions selected by feature selection.

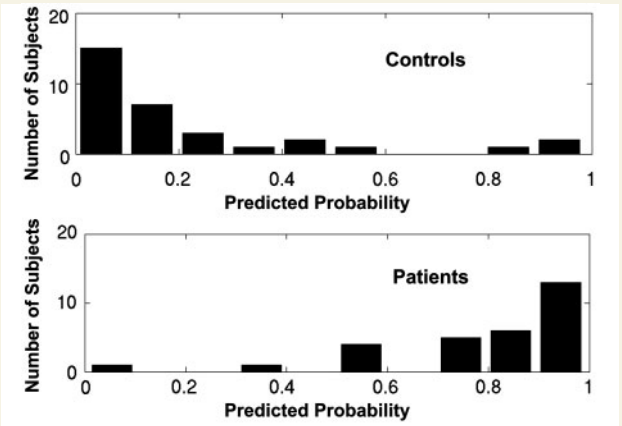

Figure 6 Histograms of distributions of RVM predicted probabilities. Probabilities for controls are skewed towards zero, probabilities for patients are skewed towards unity.

Effect of including feature-based morphometry

For the RVM-based calculation, the effect of excluding FBM (i.e. VBM-RVM) versus including FBM (i.e. VBM-FBM-RVM) was explored. The former had an accuracy of 85%, sensitivity 84% and specificity 85%. The latter had an accuracy of 90%, sensitivity 93% and specificity 87%. This indicates that while as expected, FBM did improve overall predictive accuracy, the improvement was modest at 5%.

Correlation between weighting factors and clinical ratings of illness severity

For the combined VBM-FBM-SVM method, the SVM support vector weights correlated strongly with illness severity ratings: BDI, $R = 0.82$, $P < 0.00001$, Spielberger state-trait anxiety inventory, $R = 0.76$, $P < 0.00001$, Snaith–Hamilton Pleasure Scale, $R = -0.80$, $P < 0.00001$. Inferred SVM weights (for non-support vector subjects) also correlated strongly with BDI, $R = 0.73$, $P = 0.00096$; Spielberger state-trait anxiety inventory, $R = 0.69$, $P = 0.002$; and Snaith–Hamilton Pleasure Scale, $R = -0.84$, $P = 0.0002$.

Correlations between the weighting factors calculated for the combined VBM-FBM-RVM method and illness severity ratings were of borderline significance: BDI, $R = 0.64$, $P = 0.07$; Spielberger state-trait anxiety inventory, $R = 0.69$, $P = 0.04$; and no correlation with the anhedonia ratings was found Snaith–Hamilton Pleasure Scale, $R = 0.02$, $P = 0.95$. These relationships are shown in Figs. 7 and 8.

Far fewer subjects in the RVM analysis were categorized as contributing relevance vector data and therefore weights (nine subjects) than with the SVM analysis (45 subjects) due to the relative sparsity of the RVM method (Tipping, 2001). This meant there were far less RVM weighting data available, which markedly reduced the power of the RVM correlation analysis to detect effects.

Clinical characteristics of support vector machine support and non-support categorized patients

We tested the null hypothesis of no difference between support vector patients and non-support vector patients using two group *t*-tests (BDI, $P = 0.5658$; Snaith–Hamilton Pleasure Scale, $P = 0.0374$; Spielberger state-trait anxiety inventory = 0.0695) shown in Fig. 9. The BDI test was therefore not significant but the Snaith–Hamilton Pleasure Scale test was significant at $P < 0.05$ and the Spielberger state-trait anxiety inventory test not far from significance. This indicates that patients contributing support vector imaging data were in some ways different (i.e. more anhedonic with a suggestion of increased anxiety) compared with patients who did not contribute support vector imaging data. The algorithm therefore selected some patients as contributing support vector data that were best for making predictions and these patients had some of the most marked clinical symptoms.

Discussion

Major depressive disorder is one of the most common psychiatric disorders and depressive symptoms are frequent in other psychiatric disorders such as schizophrenia, Alzheimer's disease and drug misuse. The World Health Organization ranks major depressive disorder as a major contributor to the global burden of disease in terms of 'disability-adjusted life years' (World Health Organization, 2001).

High accuracy predictive classification, as reported in the present study, is important from a clinical perspective, as major depressive disorder is still regarded by many as a 'functional' psychiatric disorder, lacking a robust neurological or structural basis. This view persists despite many reports of differences in brain structure at a group level, between patients with major depressive disorder and controls. However, group level differences provide little useful information for individual patients and there has, furthermore, been marked heterogeneity of reported group level abnormalities (Ebmeier *et al*., 2006). Consequently, these findings have had limited impact on actual clinical practice. Development of a high accuracy predictive classification technique that can discriminate individual subjects with major depressive disorder from controls, replicated across scanners and with different subjects, is therefore a significant advance.

Brain regions identified during feature selection were consistent with conventional group level studies that have reported structural grey matter reductions in patients with major depressive disorder. These reductions tend to be within the frontal lobe including the orbitofrontal and cingulate cortex, middle frontal gyrus, inferior and superior gyrus (Koolschijn *et al*., 2009). Previous studies have suggested the existence of neural circuits within the frontal-subcortical structures supporting executive function and emotional regulation. Dysfunction of these structures is associated with a variety of neuropsychiatric disorders including major depressive disorder (Tekin and Cummings, 2002; Koolschijn *et al*., 2009). Feature selection was performed to preselect a subset of

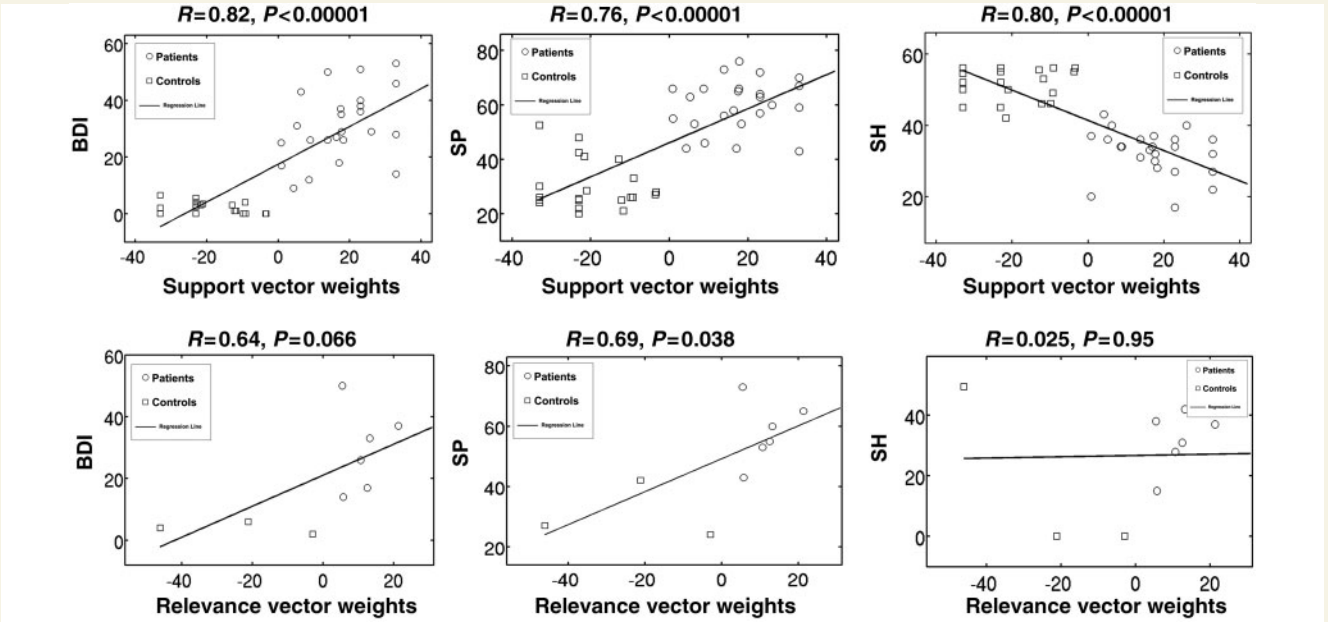

Figure 7 Scatter plots showing the relationship between relevance vector/support vector weights and illness severity ratings [depression (BDI), anhedonia (Snaith–Hamilton Pleasure Scale; SH), anxiety (Spielberger state-trait anxiety inventory; SP)].

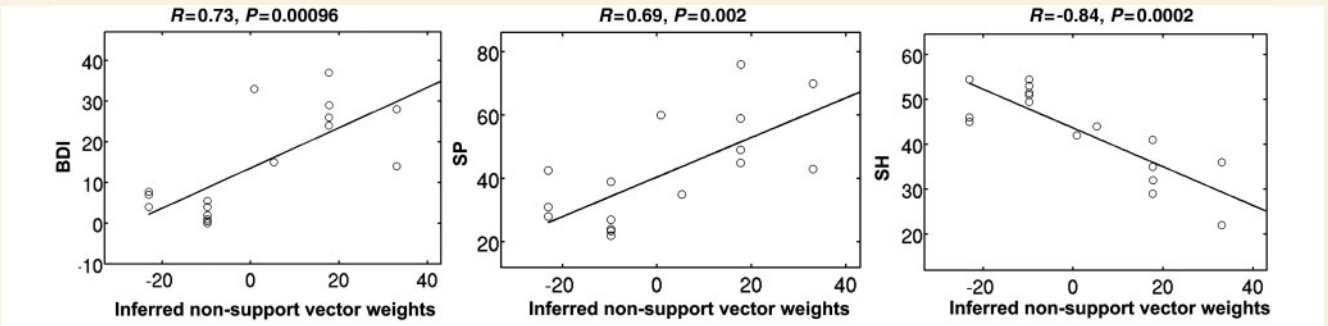

Figure 8 Scatter plots showing the relationship between the non-support vector 'inferred' weights and illness severity ratings [depression (BDI), anhedonia (Snaith–Hamilton Pleasure Scale; SH), anxiety (Spielberger state-trait anxiety inventory; SP)].

the most 'relevant' anatomical features so effectively discarding other brain regions that provided redundant information and 'noise'. Feature selection can significantly improve the accuracy of a predictive model (Guyon and Elisseeff, 2003; De Martino *et al*., 2008).

We found strong relationships between SVM (and to some extent RVM weighting values, though far fewer data were available) and illness severity ratings. To our knowledge, no other study on a psychiatric disorder has generated an illness severity index based just on categorical examples (patient versus control data). This index might be considered an objective biomarker of syndrome severity, as it has not been obtained from subjective procedures (e.g. a patient responding to questions in a rating scale and trying to match this with illness experience, or a clinician interpreting a patient's illness based on patient responses). Currently, there exist many different rating scales for depressive illness with little to guide their selection. Overlapping but different concepts of major depressive disorder result in an imperfect correlation between scores. In practice, rating scales used in research studies are often selected on the basis of popularity, to facilitate comparison with other studies. The present study suggests that it may be possible to derive an objective biomarker for major depressive disorder severity, generated from brain structure abnormalities. Further work is needed to establish whether the putative biomarker reflects low mood, anhedonia and increased anxiety only in major depressive disorder, or whether it is also applicable to other disorders in which these symptoms commonly arise, such as bipolar disorder, schizophrenia and chronic pain. The correlation between objective SVM weighting factors and subjective illness severity ratings strongly supports the suggestion that major depressive disorder illness severity is 'encoded' within a distributed network of brain regions. Notably, these brain regions (identified using feature selection) are very similar to those reported using conventional VBM group level analyses. With feature selection the

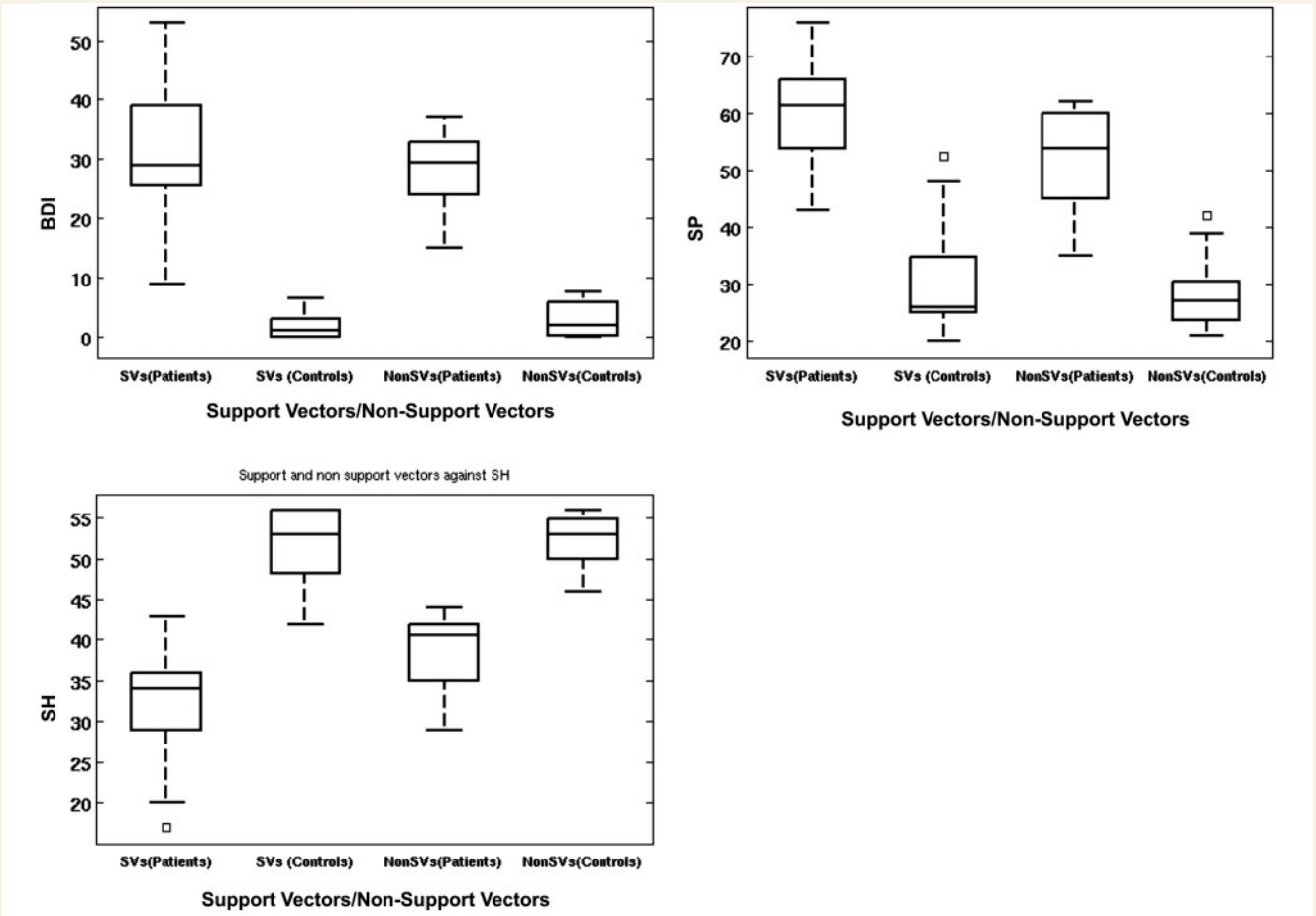

Figure 9 Box plot showing the clinical characteristics of subjects categorized as providing support vector image data (SV) versus those categorized as providing non-support vector data (Non-SV). No significant difference between groups was observed. SH = Snaith–Hamilton Pleasure Scale; SP = Spielberger state-trait anxiety inventory.

P-value threshold was iteratively 'tuned' on the basis of predictive accuracy, so could have been quite different from the VBM threshold of significance.

There are various potential limitations to the present study. The use of a VBM stage for feature selection may have excluded voxels with individually low discriminative power, but when taken in conjunction with other voxels, have significant combined power. This would have been a concern if we reported low accuracy, but we instead report high accuracy classification, compared to when an alternative multivariate wrapper feature selection method was used (Supplementary material). The current technique is limited to two-group discrimination studies but multi-group discrimination is also of clinical relevance. It should be possible to extend the model. We studied patients with major depressive disorder with chronic relapsing illness as such patients are common in UK secondary care psychiatric services. Further studies would be needed to determine whether it is possible to classify less ill patient populations with a similarly high degree of accuracy. We excluded comorbidity to facilitate interpretation of results. Further work on patients with multiple comorbidities would be required to determine the applicability of the method. While we report high accuracy predictive classification, applying the technique to a low prevalence population could, of course, result in a high false positive rate. The two scanners used in the present study were very similar and it is possible that if other scanners were used, this could have produced different results. Almost all of our participants with major depressive disorder were taking antidepressant medications whereas the controls were unmedicated. Grey matter reductions are unlikely to be due to medication as there is little evidence for antidepressants decreasing grey matter volume and evidence instead for antidepressant-linked neurogenesis (Sapolsky, 2001). It is most unlikely our results reflect medication effects as only grey matter reductions were identified, plus the SVM (and to some extent RVM) weighting factors correlated strongly with illness severity ratings. Some major depressive disorder illness descriptors were not available, such as numbers of previous illness episodes, durations of episodes of illness and detailed information on medication doses and compliance over the years. It is possible that there may be a relationship between such variables and predictive accuracy of classification, but the available data did not allow such an investigation.

In summary, we report relatively high predictive accuracy of classification of major depressive disorder using data collected from two study centres. RVM has an advantage over SVM in

that a probability of classification is obtained, which may be useful from a clinical perspective. However, SVM was best for calculating individual subject weighting factors for all participants and these may also be useful clinically. Although further work is required before this technique may be applied routinely in a clinical setting, it merits further study. Estimation of SVM weighting factors could represent an objective, brain structure derived, biomarker of major depressive disorder illness severity. If it becomes possible to not only improve diagnostic accuracy, but also to begin to prospectively define important clinical sub-phenotypes (e.g. antidepressant responsive versus unresponsive patients) on the basis of a relatively quick, hazard-free and inexpensive magnetic resonance scan, the benefits for improved patient care and enhanced targeting of therapies—'personalized medicine'—may be realized.

Funding

Gordon Small Charitable Trust, the Miller McKenzie Trust and by a SiNAPSE PhD studentship (to B.M.).

Supplementary material

Supplementary material is available at *Brain* online.

References

Ashburner J. A fast diffeomorphic image registration algirithm. Neuroimage 2007; 38: 95–113.

Ashburner J. Computational anatomy with the SPM software. Magn Reson Imaging 2009; 27: 1163–74.

Ashburner J, Friston K. Voxel-based morphometry—the methods. Neuroimage 2000; 11: 805–21.

Austin MP, Mitchell P, Goodwin GM. Cognitive deficits in depression: possible implications for functional neuropathology. Br J Psychiatry 2001; 178: 200–6.

Beck A, Ward H, Mendelson M. An inventory for measuring depression. Arch Gen Psychiatry 1961; 4: 561–71.

Belzeaux R, Formisano-Tréziny C, Loundou A, Boyer L, Gabert J, Samuelian J-C, et al. Clinical variations modulate patterns of gene expression and define blood biomarkers in major depression. J Psychiatr Res 2010; 44: 1205–13.

Bishop C. Pattern recognition and machine learning. New York, USA: Springer-Verlag; 2007.

Castellani U, Rossato E, Murino V, Bellani M, Rambaldelli G, Perlini C, et al. Classification of schizophrenia using feature-based morphometry. J Neural Transm 2012; 119: 395–404.

Costafreda S, Chu C, Ashburner J, Fu C. Prognostic and diagnostic potential of the structural neuroanatomy of depression. PLos One 2009; 4: e6353.

Davatzikos C, Resnick S, Wu X, Parmpi P, Clark C. Individual patient diagnosis of AD and FTD via high-dimensional pattern classification of MRI. Neuroimage 2008; 41: 1220–7.

De Martino F, Valente G, Staeren G, Ashburner J, Goebel R, Formisano E. Combining multivariate voxel selection and support vector machines for mapping and classification of functional MRI spatial patterns. Neuroimage 2008; 43: 44–58.

Ebmeier K, Donaghey C, Steele J. Recent developments and current controversies in depression. Lancet 2006; 367: 153–67.

Fava M, Kendler K. Major depressive disorder. Neuron 2000; 28: 335–41.

Fei-Fei L, Pietro P. A Bayesian hierarchical model for learning natural scene categories. Proc IEEE Comput Vision Pattern Recognit 2005; 2: 524–31.

Fu C, Mourao-Miranda J, Costafreda S, Khanna A, Marquand A, Williams S, et al. Pattern classification of sad facial processing: toward the development of neurobiological markers in depression. Biol Psychiatry 2008; 63: 656–62.

Godin O, Dufoull C, Maillard P, Delcroix N, Mazoyer B, Crivello F, et al. White matter lesions as a predictor of depression in the elderly: the 3C-Dijon Study. Biol Psychiatry 2008; 63: 663–9.

Gong Q, Wu Q, Scarpazza C, Lui S, Jia Z, Marquand A, et al. Prognostic prediction of therapeutic response in depression using high-field MR imaging. Neuroimage 2011; 55: 1497–503.

Gradin V, Gountouna V, Waiter G, Ahearn T, Brennan D, Condon B, et al. Between- and within-scanner variability in the CaliBrain study n-back cognitive task. Psychiatry Res 2010; 184: 86–95.

Guyon I, Elisseeff A. An Introduction to variable and feature selection. J Mach Learn Res 2003; 3: 1157–82.

Hamilton M. Rating scale for depression. J Neurol Neurosurg Psychiatry 1960; 23: 56–62.

Hardoon D, Mourão-Miranda J, Brammer M, Shawe-Taylor J. Unsupervised analysis of functional MRI data using kernel canonical correlation. Neuroimage 2007; 37: 1250–9.

Hua J, Tembe W, Dougherty E. Performance of feature-selection methods in the classification of high-dimension data. Pattern Recognit 2009; 42: 409–24.

Ji S, Li Y, Zhou Z, Kumar S, Ye J. A bag-of-words approach for *Drosophila* gene expression pattern annotation. BMC Informatics 2009; 10: 119.

Kloppel S, Chu C, Tan G, Draganski B, Johnson H, Paulsen J, et al. Automatic detection of preclinical neurodegeneration: presymptomatic Huntington disease. Neurology 2009; 72: 426–31.

Kloppel S, Stonnington C, Chu C, Draganski B, Scahill R, Rohrer J, et al. Automatic classification of MR scans in Alzheimer's disease. Brain 2008; 131: 681–9.

Koolschijn P, van Haren N, Lensvelt-Mulders G, Hulshoff Pol H, Kahn R. Brain volume abnormalities in major depressive disorder: a meta-analysis of magnetic resonance imaging studies. Hum Brain Mapp 2009; 30: 3719–35.

Koutsouleris N, Meisenzahl E, Davatzikos C, Bottlender R, Frodl T, Scheuerecker J, et al. Use of neuroanatomical pattern classification to identify subjects in at-risk states of psychosis and predict disease transition. Arch Gen Psychiatry 2009; 66: 700–12.

Kriegeskorte N, Simmons K, Bellgowan P, Baker C. Circular analysis in systems neuroscience: the dangers of double dipping. Nat Neurosci 2009; 12: 535–40.

Kumar P, Waiter G, Ahearn T, Milders M, Reid I, Steele J. Abnormal temporal difference reward-learning signals in major depression. Brain 2008; 131: 2084–93.

Kvaal K, Ulstein I, Nordhus I, Engedal K. The Spielberger State-Trait Anxiety Inventory (STAI): the state scale in detecting mental disorders in geriatric patients. Int J Geriatr Psychiatry 2005; 20: 629–34.

Lao Z, Shen D, Xue Z, Karacali B, Resnick SM, Davatzikos C. Morphological classification of brains via high-dimensional shape transformations and machine learning methods. Neuroimage 2004; 21: 46–57.

Lloyd S. Least squares quantisation in PCM. IEEE Transactions on Information Theory 1982; IT-28: 129–37.

Lorenzetti V, Allen N, Fornito A, Yucel M. Structural brain abnormalities in major depressive disorder: a selective review of recent MRI studies. J Affect Disord 2009; 117: 1–17.

Mackay C. The evidence framework applied to classification networks. Neural Comput 1992; 4: 720–36.

Magnin B, Mesrob L, Kinkingnehun S, Pelegrini-Issac M, Colliot O, Sarazin M, et al. Support vector machine-based classification of Alzheimer's disease from whole-brain anatomical MRI. Neuroradiology 2009; 51: 73–83.

Martino F, Valente G, Staere N, Ashburner J, Goebel R, Formisano E. Combining multivariate voxel selection and support vector machine for

mapping and classification of functional MRI spatial patterns. Neuroimage 2008; 43: 44–58.
Miller AH, Maletic V, Raison CL. Inflammation and its discontents: the role of cytokines in the pathophysiology of major depression. Biol Psychiatry 2009; 65: 732–41.
Moorhead T, Gountouna V, Job D, McIntosh A, Romaniuk L, Lymer G, et al. Prospective multi-centre voxel based morphometry study employing scanner specific segmentations: procedure development using CaliBrain structural MRI data. BMC Med Imaging 2009; 9: 8.
Nakonezny PA, Carmody TJ, Morris DW, Kurian BT, Trivedi MH. Psychometric evaluation of the Snaith–Hamilton pleasure scale in adult outpatients with major depressive disorder. Int Clin Psychopharmacol 2010; 25: 328–33.
Nelson H, Willison J. The revised national adult reading test (NART)-test manual. Windsor, UK: NFER-Nelson; 1991.
Nielsen F, Balsleva D, Hansenb K. Mining the posterior cingulate: segregation between memory and pain components. Neuroimage 2005; 27: 520–32.
Rasmussen C, Williams K. Gaussian processes for machine learning. Cambridge, Massachusetts: The MIT Press; 2006.
Saeys Y, Inza I, Larranaga P. A review of feature selection techniques in bioinformatics. Bioinformatics 2007; 23: 2507–17.
Sapolsky R. Depression, antidepressants, and the shrinking hippocampus. Proc Natl Acad Sci USA 2001; 98: 12320–2.
Schmidt HD, Shelton RC, Duman RS. Functional biomarkers of depression: diagnosis, treatment, and pathophysiology. Neuropsychopharmacology 2011; 36: 2375–94.
Schneider B, Prvulovic D, Oertel-Knöchel V, Knöchel C, Reinke B, Grexa M, et al. Biomarkers for major depression and its delineation from neurodegenerative disorders. Prog Neurobiol 2011; 95: 703–17.
Schwaighofer A. SVM Toolbox, 2001.
Scovanner P, Ali S, Shah M. A 3-Dimensional SIFT Descriptor. In: ACM Multimedia, 'MM07'. Augustburg, Germany; 2007.
Shah P, Ebmeier K, Glabus M, Goodwin G. Cortical grey matter reductions associated with treament-resistant unipolar depression. Controlled magnetic resonance imaging study. Br J Psychiatry 1998; 172: 527–32.
Snaith R, Hamilton M, Morley S, Humayan A, Hargreaves D, Trigwell P. A scale for the assessment of hedonic tone the Snaith–Hamilton Pleasure Scale. Br J Psychiatry 1995; 167: 99–103.
Sobin C, Sackeim H. Psychomotor symptoms of depression. Am J Psychiatry 1997; 154: 154–61.
Steele J, Kumar P, Ebmeier K. Blunted response to feedback information in depressive illness. Brain 2007; 130: 2367–74.
Suckling J, Barnes A, Job D, Brenan D, Lymer K, Dazzan P, et al. Power calculations for multicenter imaging studies controlled by the false discovery rate. Hum Brain Mapp 2010; 31: 1183–95.
Tekin S, Cummings J. Frontal-subcortical neuronal circuits and clinical neuropsychiatry: an update. J Psychosom Res 2002; 53: 647–54.
Theodoridis S, Koutroumbas K. Pattern recognition. 3rd edn. London: Academic Press; 2006.
Tipping M. Sparse Bayesian learning and the relevance vector machine. J Mach Learn Res 2001; 1: 211–44.
Toews M, Wells W III, Collins DL, Arbel T. Feature-based morphometry: discovering group-related anatomical patterns. Neuroimage 2010; 49: 2318–27.
Toews M, Wells WM III, Collins DL, Arbel T. Feature-based morphometry. Medical Image Computing And Computer-Assisted Intervention: MICCAI. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Vol. 12. London, United Kingdom; 2009. p. 109–16.
Tommasi T, Orabona F, Caputo B. Discriminative cue integration for medical image annotation. Pattern Recognit Lett 2008; 29: 1996–2002.
Vapnik V. Statistical Learning Theory. New York, USA: Wiley; 1998.
World Health Organisation. Mental health: new understanding, new hope. Geneva, Switzerland: World Health Organisation; 2001.

Appendix I

Machine learning

Given a supervised learning problem $\{x_i, y_i\}_{i=1}^{N}$ where $y_i \in \{-1, 1\}$ represents class labels (major depressive disorder versus controls) and x_i represents feature vectors during training, both SVM and RVM share a similar linear structure

$$y(x) = \sum_{i=1}^{M} w_i k(x, x_i) + b. \quad \text{(A1)}$$

where b represents a bias parameter. The objective of SVM/RVM training is to estimate optimal weights $W = (w_1, w_2, \ldots, w_M)^T$ and identify support (SVM) or relevant (RVM) vectors corresponding to these weights. Notably, each subject was associated with, at most, one imaging data support vector and one associated weighting factor. Equation (A1) is also used to make predictions about novel imaging data during the testing stage. SVM and RVM use different detailed formulations as described below.

Support vector machine

Given a supervised learning problem with two labelled classes (e.g. major depressive disorder versus control), the SVM algorithm estimates a hyperplane boundary (Fig. 1) which best separates the two classes. This boundary $w.x - b = 0$ was obtained by minimizing classification errors and simultaneously maximizing the margin 'd' (Fig. 1) between the closest support vectors from both classes. This is equivalent to minimizing

$$\arg\min_{w,b} \frac{1}{2}||w||^2 + C\sum_{i}^{N} \xi_i, \quad \text{(A2)}$$

subject to the constraint; $y_i(w.x_i - b) \geq 1 - \xi_i$, $\xi_i \geq 0$, $i = 1, 2, \ldots N$ where w is a normal vector and ξ_i are slack variables. To solve the above constraint as well as minimize $||w||$, Lagrange multipliers were introduced leading to the following formulation which was maximized subject to $0 \leq \alpha_i \leq C$ and $\sum \alpha_i y_i = 0$

$$\arg\max_{\alpha} w(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum \alpha_i \alpha_j y_i y_j x_i^T x_j \quad \text{(A3)}$$

$C(>0)$ is a 'regularization parameter' controlling the trade-off between regularization and constraint violation. The advantage of the above is that it is possible to make predictions by computing dot products with the support vectors. Having obtained optimal α values, the hyperplane (Fig. 1) decision function was defined

$$f(x) = \sum_{i=1}^{M} \alpha_i y_i x^T x_i + b \quad \text{(A4)}$$

where $\{x\}_{i=1}^{M}$ are a set of support vectors (defining a margin 'd', Fig. 1), b is a model bias term and y are class labels (major depressive disorder versus control).

Relevance vector machine

Similar to above, the classification problem was represented as $\{X_n, t_n\}_{n=1}^{N}$, where X_n represents training stage data and

t_n-corresponding target labels which can either be continuous values (e.g. clinical rating scores) for a regression problem, or binary classification values (e.g. major depressive disorder versus controls) for a classification problem.

The above formulation can be represented as a standard linear model

$$t_n = y(x_n; W) + \varepsilon_n \tag{A5}$$

generalized to

$$t_n = \sum_{n=1}^{N} \omega_n k(x, x_n) + b + \varepsilon_n \tag{A6}$$

where $W = (\omega_1, \omega_2, \ldots \omega_N)^T$ is a weighting vector, b is a model bias identified during the training process, x_n is a feature vector and ε_n represents measurement noise (assumed zero mean Gaussian distribution and variance σ^2). $k(x, x_n)$ is a non-linear Gaussian kernel mapping function $k(x_1, x_2) = \exp(-\eta||x_1 - x_2||)$, $\eta > 0$, $\eta = \frac{1}{2\sigma^2}$. As discussed in the main text and Supplementary material, the kernel width parameter σ was estimated using a non-biased cross-validation procedure during the training stage.

The objective of training the model was to estimate the best 'weighting' values W given input neuroimaging feature vector data. Feature vectors with corresponding weighting values were then used for making predictions during the testing stage.

The RVM algorithm employs a Bayesian formulation aimed at estimating the posterior distribution of weighting values:

$$\text{Posterior distribution} = \frac{\text{Likelihood} \times \text{prior}}{\text{marginal distribution (normalization factor)}} \tag{A7}$$

The likelihood of a given dataset can be expressed as

$$p(t|W, \sigma^2) = (2\pi\sigma^2)^{-N/2} \exp\left\{-\frac{1}{2\sigma^2}||t - \Phi W||^2\right\} \tag{A8}$$

where $t = (t_1 \ldots t_N)^T$, Φ is a $N \times (N+1)$ design matrix defined as $\Phi = [\varphi(x_1), \varphi(x_2), \ldots \varphi(x_N)]^T$ and $\varphi(x_n) = [1, K(x_n, x_1), K(x_n, x_2), \ldots, K(x_n, x_N)]^T$.

A standard approach to avoid over-fitting (Tipping, 2001) is to introduce a zero-mean Gaussian prior distribution of the parameters defined as

$$p(W/\alpha) = \prod_{n=0}^{N} \mathrm{N}(\omega_n|0, \alpha_n^{-1}) \tag{A9}$$

where α is a vector of $N + 1$ hyperparameters, each with a weighting parameter (ω_n) paired with individual hyperparameters.

The process of defining the priors of the hyperparameters α (hyper-priors) and noise variance σ^2 is described in detail elsewhere (Tipping, 2001; Rasmussen and Williams, 2006; Bishop, 2007). When the evidence for the model was maximized with respect to the hyperparameters, a number of hyperparameters tend towards infinity constraining the corresponding parameters to zero. These parameters were pruned using the automatic relevance determination method (Mackay, 1992; Tipping, 2001) resulting in a sparse model.

Since the likelihood (A8) and prior (A9) are defined this allows the Bayesian formula (A7) to be re-expressed as

$$p(W|t, \alpha, \sigma^2) = \frac{p(t|W, \sigma^2)p(W|\alpha)}{p(t|\alpha, \sigma^2)} \tag{A10}$$

The likelihood and prior are Gaussian, so the posterior distribution is also Gaussian with mean μ and covariance Σ

$$\begin{aligned} p(W|t, \alpha, \sigma^2) &= \mathrm{N}(\mu, \textstyle\sum) \\ \mu &= \sigma^{-2} \textstyle\sum \Phi^T t \\ \textstyle\sum &= \left(\sigma^{-2}\Phi^T\Phi + A\right)^{-1} \\ A &= \mathrm{diag}(\alpha_0, \alpha_1, \ldots, \alpha_N) \end{aligned} \tag{A11}$$

where A is a diagonal matrix depicting the inverse of the noise for each weight. The parameters α and σ were estimated using a type-II maximum likelihood method (Tipping, 2001; Bishop, 2007).

Predictions made by weighting the basis functions (A8) by the posterior mean weights can be used for regression problems. For classification problems, as in the present study, a modified likelihood function was used. The objective was to predict the posterior probability of membership of one of the two classes (major depressive disorder versus controls) given training data. The linear model introduced earlier $t_n = y(x_n; W) + \varepsilon_n$ was transformed using a logistic sigmoid function

$$\sigma(y) = \frac{1}{(1 + e^{-\gamma})} \text{to} y(x, w) = \sigma\{W^T, \varphi(x)\} \quad = \frac{1}{1 + \exp\{W^T\varphi(X)\}}$$

By adopting a Bernoulli distribution it is possible to derive:

$$p(t|w) = \prod_{n=1}^{N} \sigma\{y(x_n; W)\}^{t_n}[1 - \sigma\{y(x_n; W)\}]^{1-t_n} \tag{A12}$$

Finally, the posterior distribution of the weights was estimated using a Laplace method (Mackay, 1992; Tipping, 2001). These weights were used for making RVM predictions (Tipping, 2001; Rasmussen and Williams, 2006; Bishop, 2007).

Feature-based morphometry: 3D-SIFT and 'Bag of Words'

For FBM, local anatomical brain regions were identified by feature selection (VBM t-test) during the training stage and used for input to the 3D-SIFT calculation. In the 3D-SIFT calculation, gradient magnitude $m_{3D}(x, y, z)$ and orientations $\theta(x, y, z)$, $\phi(x, y, z)$ within a $4 \times 4 \times 4$ voxel 3D neighbourhood were calculated as follows:

$$m_{3D}(x, y, z) = \sqrt{L_x^2 + L_y^2 + L_z^2} \tag{A13}$$

$$\theta(x, y, z) = \tan^{-1}\left(\frac{L_y}{L_x}\right) \tag{A14}$$

$$\phi(x, y, z) = \tan^{-1}\left(\frac{L_z}{\sqrt{L_x^2 + L_y^2}}\right) \tag{A15}$$

where L_x, L_y and L_z are grey matter gradients in the x, y and z directions calculated using voxel differencing:

$$L_x = L(x+1,y,z) - L(x-1,y,z) \quad \text{(A16)}$$

$$L_y = L(x,y+1,z) - L(x,y-1,z) \quad \text{(A17)}$$

$$L_z = L(x,y,z+1) - L(x,y,z-1) \quad \text{(A18)}$$

Each (θ, ϕ) pair represents the orientation of the gradient (rate of change) of grey matter probability for each voxel within a given 3D anatomical neighbourhood. Tessellation binning (Scovanner *et al*., 2007) was used to calculate histograms, the peaks representing the dominant rates of change of grey matter probability. The maximal peak was used to create the 3D-SIFT descriptor. As will be described, the Bag-of-Words method combined the multiple 3D-SIFT descriptors for each subject into a single feature vector descriptor per subject (scan).

To implement the 'Bag of Words' method, k-means clustering (Lloyd, 1982) was used to cluster all 3D-SIFT descriptors for a given subject during training into K classes or cluster centres. A single feature vector representing a single scan was calculated by binning minimal Euclidean distances between 3D-SIFT features extracted from the scan and cluster centres. The number of K clusters had to be specified in advance and various k parameters were tested ($K = 100$, 300, 500, 1000) and results from a validation dataset from the training set indicated $K = 100$ was best (Supplementary material). Therefore, a single T_1-weighted scan from a single subject was represented by a 100D anatomical feature vector. These feature vectors with corresponding diagnostic labels (major depressive disorder or control) were used as training data for machine learning (SVM or RVM). Further details of the 3D-SIFT process and 'Bag of Words' procedure are provided in the Supplementary material.

Inferred support vector machine weights: L_1 norm distance calculation

The inferred SVM weights were calculated using the L_1 norm distance, by calculating the pairwise distances between a given non-support vector subject data and all support vector subject data. The closest distance was used to identify the inferred weight as follows:

$$d(a,b) = ||a-b|| = \sum_{i=1}^{n} |a-b| \quad \text{(A19)}$$

where a and b represent non-support vector subject data and support vector subject data respectively and n is the number of non-support vector subject data.