Neuroinform (2014) 12:229–244
DOI 10.1007/s12021-013-9204-3

REVIEW

A Review of Feature Reduction Techniques in Neuroimaging

Benson Mwangi · Tian Siva Tian · Jair C. Soares

Published online: 8 September 2013

Abstract Machine learning techniques are increasingly being used in making relevant predictions and inferences on *individual* subjects neuroimaging scan data. Previous studies have mostly focused on categorical discrimination of patients and matched healthy controls and more recently, on prediction of *individual* continuous variables such as clinical scores or age. However, these studies are greatly hampered by the large number of predictor variables (voxels) and low observations (subjects) also known as the *curse-of-dimensionality* or *small-n-large-p* problem. As a result, feature reduction techniques such as feature subset selection and dimensionality reduction are used to remove redundant predictor variables and experimental noise, a process which mitigates the *curse-of-dimensionality* and *small-n-large-p* effects. Feature reduction is an essential step before training a machine learning model to avoid *overfitting* and therefore improving model prediction *accuracy* and generalization ability. In this review, we discuss feature reduction techniques used with machine learning in neuroimaging studies.

Keywords Feature reduction · Feature selection · Dimensionality reduction · Machine learning · Multivariate · Predictive Modeling · Neuroimaging

Introduction

Conventional *group-level* neuroimaging data analysis techniques (e.g. voxel-based morphometry) have been used in neuroimaging research for decades. However, these techniques have only been able to identify average *between-group* differences and unable to make predictions on *individual* subjects. The inability to make predictions from *individual* subjects neuroimaging scan data may have greatly hampered the ability to translate neuroimaging research results into clinical practice (Brammer 2009; Linden 2012). Recently though, machine learning (ML) techniques also variously known as multivariate pattern analysis (MVPA) or pattern recognition (PR) techniques, have been ucsed to decode or predict *individual* subjects' brain states using neuroimaging scan data. These studies have increased significantly and equally become successful in decoding brain states which raises the possibility of being deployed for potential clinical use and particularly in making *personalized* clinical decisions (Linden 2012; Mwangi et al. 2012a).

Machine learning applications in neuroimaging can be divided into two broad categories namely *classification* and *regression*. In classification, neuroimaging data with corresponding categorical labels (e.g. Healthy Controls vs Patients or Treatment responders vs Non-responders) are used to develop a predictive classifier. The resulting classifier is used to make predictions on subjects' data not present during the training stage. Previous classification studies include; Alzheimer's disease (AD) (Kloppel et al. 2008; Magnin et al. 2009; Zhang et al. 2011b), Major Depressive Disorder (MDD) (Mwangi et al. 2012a; Zeng et al. 2012), Autism Spectrum Disorder (ASD) (Ecker et al. 2010; Ingalhalikar et al. 2011),Schizophrenia (Koutsouleris et al. 2009), Mild Cognitive Impairment(MCI) (Haller et al. 2010b) and Attention Deficient Hyperactivity Disorder (ADHD) (Lim et al. 2013; Sato et al. 2012; Zhu et al. 2005). Conversely, in regression neuroimaging data with corresponding continuous targets (e.g. clinical scores or age) are used to 'train' a regression model. Similar to above, the model is used to make predictions on novel *individual* subjects' data not present during training. These techniques have recently been used to predict *individual* subjects' age (Brown et al. 2012;

B. Mwangi (✉) · J. C. Soares
UT Center of Excellence on Mood Disorders, Department of Psychiatry and Behavioral Sciences, UT Houston Medical School, Houston, TX 77054, USA
e-mail: benson.irungu@uth.tmc.edu

T. S. Tian
Department of Psychology, University of Houston, Houston, TX 77024, USA

Dosenbach et al. 2010; Franke et al. 2012; Franke et al. 2010; Mwangi et al. 2013) and clinical scores in MDD, AD and chronic pain (Marquand et al. 2010a; Mwangi et al. 2012b; Stonnington et al. 2010; Wang et al. 2010).

Notably, previous studies in both *classification* and *regression* have used neuroimaging scan data from multiple modalities such as T_1-weighted magnetic resonance imaging (MRI), functional MRI (fMRI), diffusion tensor imaging (DTI), positron emission tomography (PET) and single-photon emission computed tomography (SPECT). In this review, we generalize our discussions to all imaging modalities and assume that all scans have been subjected to standardized pre-processing routines (e.g. spatial normalisation, segmentation and or smoothing). Neuroimaging data pre-processing frameworks in different modalities are discussed elsewhere (Ashburner 2007; Ashburner 2009; Johansen-Berg and Behrens 2009; Penny et al. 2007). In this review, pre-processed neuroimaging scans' voxels are referred to as 'features' or 'predictor variables'. In addition, we generalize our feature reduction discussion to all machine learning techniques such as support vector machines (SVMs) (Mwangi et al. 2012a; Orru et al. 2012; Vapnik 1999), relevance vector machines (RVMs) (Mwangi et al. 2012a; Tipping 2001) and Gaussian process classifiers (GPCs) (Marquand et al. 2010b; Rasmussen and Williams 2006). Discussions of these methods with respect to neuroimaging data are also given elsewhere (Johnston et al. 2012; Misaki et al. 2010; Orru et al. 2012).

Rationale for Feature Reduction

In many neuroimaging studies, the sample size (number of subjects or observations) is often less than 1000. In comparison, pre-processed brain scans may contain (>100,000) non-zero voxels. As a result, the numbers of features (voxels) greatly outnumber the number of observations (sample size). This is a common problem in machine learning literature known as the *curse-of-dimensionality* (Bellman 1961) or *small-n-large-p* (Fort and Lambert-Lacroix 2005). Consequently, without preselecting the 'most relevant' features and effectively discarding redundant features plus noise, a predictive machine learning model has a marked risk of 'overfitting' (Guyon and Elisseeff 2003; Hua and Dougherty 2009). Overfitting implies that the model training process yields a machine learning model with poor generalization ability which is interpreted as inability to make accurate predictions on novel subjects' data (Guyon and Elisseeff 2003; Hua and Dougherty 2009; Kohavi and John 1997).

In view of the above, there is a general consensus from previous neuroimaging machine learning studies that feature reduction is a fundamental step before applying a predictive model (e.g. SVM or RVM) to neuroimaging data (Bray et al. 2009; Cheng et al. 2012; De Martino et al. 2008; Duchesnay et al. 2007; Duchesnay et al. 2004; Duff et al. 2011; Franke et al. 2010; Liu et al. 2012; Lohmann et al. 2007; Mitchell et al. 2004; Mwangi et al. 2012a; Norman et al. 2006; Pereira et al. 2009; Rizk-Jackson et al. 2011; Schrouff et al. 2013; Valente et al. 2011; Van De Ville and Lee 2012). Other than mitigating the *curse-of-dimensionality* effect as above, the feature reduction process may also facilitate a deeper understanding of the scientific question of interest. For example, in a highly accurate predictive classifier able to predict patient treatment responders against non-responders, brain regions identified during feature reduction may offer an insight into different neural substrates of non-responders. In this review we divide feature reduction techniques into two broad categories namely; *supervised* and *unsupervised* techniques respectively (Table 1).

Supervised Feature Reduction Techniques

Supervised feature reduction techniques use highly dimensional neuroimaging data and the required outcome labels (e.g. +1 treatment responders, -1 treatment non-responders) to select relevant features and discard redundant features plus noise. However, as in the wider machine learning literature, we further subdivide these techniques into three categories namely; 'filter', 'wrapper' and 'embedded' methods, respectively (Guyon and Elisseeff 2003; Saeys et al. 2007). A clear distinction between these categories exists. First, filter techniques such as t-tests, Anova and Pearson correlation coefficient use simple statistical measures (e.g. mean, variance, correlation coefficients) to rank features according to their relevance in detecting *group-level* differences. Second, wrapper techniques use an *objective* function from a classification or regression machine learning model to rank features

Table 1 A tabular summary of both supervised and unsupervised feature reduction techniques

Supervised	Unsupervised
a) Filter techniques	a) Data driven
- *T*-test, Anova, pearson correlation coefficient	- Principal component analysis
b) Wrapper techniques	- Independent component analysis
- Recursive feature elimination	
- Searchlight	
c) Embedded techniques	b) Domain Knowledge driven
-Least absolute shrinkage and selection operator	- coordinate-based meta-analysis
- Elastic Net	
- Partial least squares method	

according to their relevance to the model. Lastly, embedded methods select relevant features as 'part' of the machine learning process by enforcing certain 'penalties' on a machine learning model thus yielding a small subset of relevant features.

In passing, we note that relevant features are usually selected using training data only to avoid *double-dipping* which entails using both training and testing data partitions to select relevant features (Kriegeskorte et al. 2010; Kriegeskorte et al. 2009). In this review, we assume both (training and testing datasets) have been adequately separated using a *cross-validation* procedure such as hold-out or leave-one-out which are explored elsewhere (Johnston et al. 2012; Pereira et al. 2009; Strother et al. 2002; Theodoridis and Koutroumbas 2009). Figure 1 illustrates a feature reduction process without double dipping. Conversely, Fig. 2 illustrates a feature reduction procedure with double dipping – which should be avoided.

Filter Techniques

Pearson Correlation Coefficient

The Pearson correlation coefficient (PCC) ranks features by calculating linear correlations between individual features and class labels in classification or continuous targets in regression (Guyon and Elisseeff 2003). Here, we assume a two group classification problem with predictor variables x and diagnostic labels y_i, (e.g. Patients-1 vs Controls −2). The Pearson correlation coefficient between predictor variables and diagnostic labels is calculated as:

$$P_i = \frac{\sum_i \left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sqrt{\sum_i \left(x_i - \bar{x}\right)^2 \sum_i \left(y_i - \bar{y}\right)^2}} \quad (1)$$

Where x_i stands for the feature value of the i^{th} sample and $\bar{x}$ is the mean of these feature values. y_i are diagnostic labels and $\bar{y}$ is the mean of all y_i in the sample (Fan et al. 2007; Guyon and Elisseeff 2003). The relevance of a feature in separating both classes is evaluated by considering the absolute value of the correlation coefficient P_i with higher values indicating the feature's greater relevance in discriminating between classes. Lastly, relevant features above a *user-defined* threshold of correlation coefficients are selected for subsequent machine learning analyses. However, to select the optimal *user-defined* threshold, cross-validation procedures (e.g. k-fold or leave-one-out) are used to iteratively evaluate a range of threshold values and the threshold with a low generalization error selected.

Numerous studies have applied PCC filters to select relevant features. For example, in MCI classification (Wee et al. 2011), cocaine exposed individuals classification (Fan et al. 2007), AD classification (Dai et al. 2012; Davatzikos et al. 2008; Grana et al. 2011) and in gender (Men vs Women) classification (Fan et al. 2006).

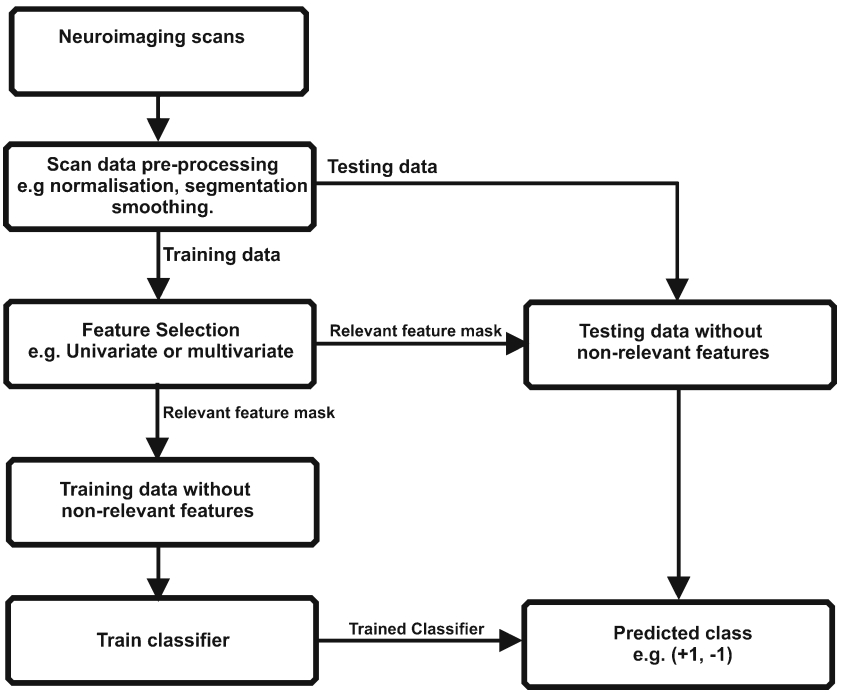

Fig. 1 Feature reduction without *double-dipping*. Training and testing datasets are separated before the feature reduction process

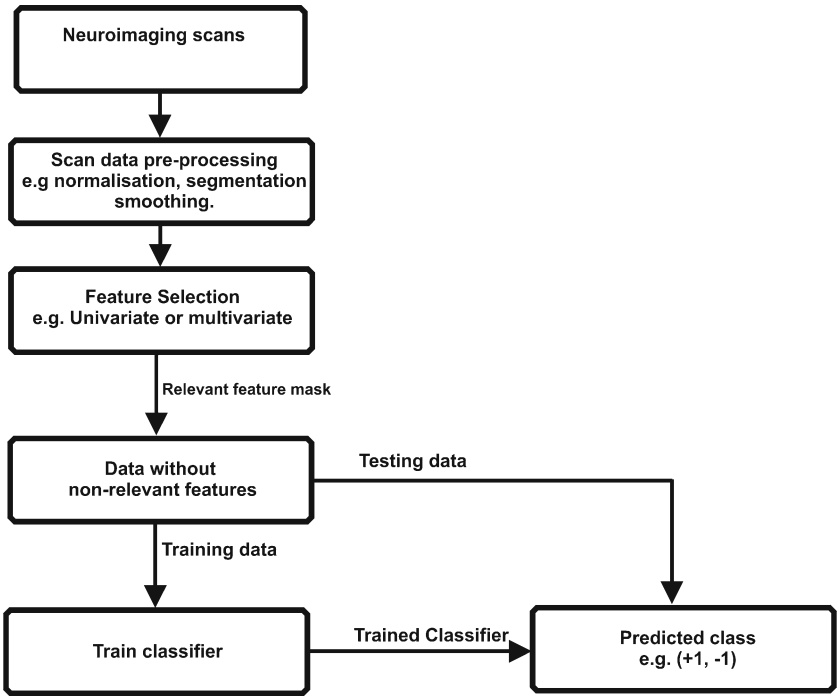

Fig. 2 Feature reduction with *double-dipping*. Features are selected from the same set of training and testing data

However, although PCC filters are applicable to both multi-group (>2) classification and regression tasks, they are only able to detect linear dependencies between features and corresponding targets (Guyon and Elisseeff 2003). This is a major setback for PCC filters especially in tasks involving selecting relevant features from high-dimensional data with multivariate relationships (Wang et al. 2010).

T-Tests and Analysis of Variance (ANOVA)

Statistical hypothesis testing techniques such as t-tests and ANOVA have extensively been used to detect *average* gro*up-level* differences in neuroimaging studies. Recently, though, these techniques have been used for supervised feature selection in neuroimaging machine learning studies with success. Here, we assume a two class classification task of categorising (treatment responders +1 vs. treatment Non-responders −1). The assumption that predictor variables are from pre-processed neuroimaging scans still holds. As a result, a two-sample t-test evaluates the null hypothesis that the two-group (Responders, Non-responders) sample means are equal given observed variance. Let $\overline{x_p}$ and $\overline{x_c}$ be voxel means for patient and control groups respectively and s_p, $s_c$denote corresponding standard deviations. The t-score of a feature v is calculated as (Sheskin 2004);

$$t = \frac{\left|\overline{x_p} - \overline{x_c}\right|}{\sqrt{\frac{(N_p-1)s_p^2 + (N_c-1)s_c^2}{N_p + N_c - 2} \cdot \left[\frac{1}{N_p} + \frac{1}{N_c}\right]}} \quad (2)$$

Where N_p and N_c denote number of subjects in each group. Once this calculation is performed, a threshold of significance (e.g. p-value) which represents the probability of obtaining a statistic greater in magnitude than t under the null hypothesis is defined (Ashburner and Friston 2000; Penny et al. 2007; Theodoridis and Koutroumbas 2009). Subsequently, an optimal *user-defined* threshold of significance (p-value) representing *relevant* features is selected through a cross-validation process and relevant features used for subsequent machine learning analyses (Mwangi et al. 2012a). Several classification studies have recently used t-tests to select relevant features for machine learning in neuroimaging. For example, AD (Chaves et al. 2009; Chu et al. 2012; Hinrichs et al. 2009; Zhang et al. 2011a), ASD (Duchesnay et al. 2011), MDD (Mwangi et al. 2012a), gender (Duchesnay et al. 2007), schizophrenia (Kovalev et al. 2003), MCI (Wee et al. 2012) and in other non-clinical studies (Balci et al. 2008; Haynes and Rees 2005). Applications of t-tests in feature selection are computationally fast and easy to implement and scale well to high dimensional data meaning they are able to select a small subset of *relevant* features from the original high-dimensional feature set (Hua and Dougherty 2009; Mwangi et al. 2012a; Saeys et al. 2007).

However, t-tests are also weighed down by a number of shortcomings. First, they are *univariate,* meaning they do not take into account interactions between multiple features and spatial patterns (non-multivariate). Second, t-tests are only able to detect two-group differences, although this shortcoming is compensated by the equivalent analysis of variance (ANOVA) technique.

The ANOVA technique is used to select *relevant* features in multiple (>2) groups by generalizing a t-test to more than two groups (Cohen 1998). However, similar to t-tests, ANOVAs are *univariate* but have also been used in selecting relevant features in neuroimaging classification tasks. For example, Costafreda and colleagues (Costafreda et al. 2011) recently used ANOVA for feature selection in classifying patients with bipolar disorder (BD), schizophrenia and healthy controls. Other neuroimaging classification studies using ANOVA for feature reduction include; ASD (Coutanche et al. 2011), MDD (Costafreda et al. 2009), schizophrenia (Yoon et al. 2008) and fMRI visual activation classification task (Cox and Savoy 2003). In passing, we note that ANOVAs offer the same benefit as t-tests, but an optimal threshold of relevant features is selected using a cross-validation process with the training data. Most notably, we note that the multivariate analysis of covariance (MANCOVA) is a multivariate extension of ANOVA and has recently been used in a number of neuroimaging feature reduction tasks (Friston et al. 1996; Calhoun et al. 2011; Allen et al. 2011). Lastly, Wilcoxon test (De Martino et al. 2008), signal-to-noise ratio (Ingalhalikar et al. 2011), Gini-contrast (Langs et al. 2011) and relief (Haller et al. 2010a) are other filter techniques, albeit with few applications in neuroimaging machine learning studies so far.

However, whilst these supervised filter techniques have extensively been used in mitigating the *small-n-large-p* or *curse-of-dimensionality* problems in neuroimaging, they collectively suffer from a significant setback by not considering interactions between multiple features or spatial patterns (not multivariate).

Multivariate Wrapper Techniques

Multivariate wrapper techniques use an *objective-function* from a multivariate machine learning model (e.g. classification or regression) to rank features according to their relevance or importance to the model (Guyon and Elisseeff 2003; Kohavi and John 1997). Multivariate wrapper approaches are further subdivided into two sub-categories namely; forward selection and backward elimination (Kohavi and John 1997). In forward selection, the search for *relevant* features begins with an empty set and features are iteratively added in 'small' pre-defined steps until an *optimal* number of features are found. On the contrary, in backward elimination, the search begins with all features in the training set which are iteratively removed in 'small' pre-defined steps until an *optimal* number of features are found. Recursive feature elimination (RFE) (Guyon et al. 2001), which is popularly used in neuroimaging studies is an example of a backward elimination technique.

Recursive Feature Elimination

We assume a two-group classification task with a set of features x_i, and corresponding labels y_i. Equally, we assume training data is now sub-divided into 'training' and 'evaluation' subsets respectively. A machine learning algorithm (e.g. linear support vector machine or linear relevance vector machine) is 'trained' resulting into observation weights a_i. As a result, feature or voxel relevance weights are calculated as:

$$W = \sum_{x_i \in cf} \alpha_i y_i x_i \tag{3}$$

Where cf stands for observations or subjects with non-zero weights (e.g. support vectors in SVM or analogous relevance vectors in RVM). Subsequently, absolute values of the weights W are ranked in order of importance (low-less relevant, high-most relevant) and a *user-defined* percentage (e.g. 2 %) of the lowest ranking features removed. In the next step, a new model is trained minus the non-relevant features and the new model's generalization error (or accuracy) on the evaluation subset is assessed. This process is repeated until a termination criterion is reached or until the feature set is empty. Lastly, features leading to the best generalization ability (high accuracy) are selected for training the final machine learning model and the rest of the features discarded. Figure 3 illustrates a recursive feature elimination process.

However, RFE requires definition of two parameters. 1) Backward-eliminations termination or stopping criteria. The first possibility is to remove low ranking features iteratively until the feature set is empty and the iteration resulting to the best generalization ability or high accuracy is selected. The second possibility is to terminate the procedure when the model performance of the current iteration is not significantly better than the previous step, as explored by De Martino and colleagues (De Martino et al. 2008). 2) Another *user-defined* parameter is the percentage of features removed at every backward-elimination step. Previous neuroimaging studies using RFE have variably used this parameter, for example 8 % (De Martino et al. 2008), 10 % (Craddock et al. 2009) and 2 % (Mwangi et al. 2013) . It is not currently well understood how the choice of this parameter affects the overall model performance and this remains an open research question. However, removing a very small percentage of features at every iteration is computationally expensive, whilst removing a higher percentage may result into inclusion of non-relevant features (Craddock et al. 2009; De Martino et al. 2008).

RFE offers several benefits by first, considering multivariate interactions between entire spatial patterns in the data. Second, the technique uses a predictive model to remove non-relevant or redundant features which may result in a better generalization ability (Guyon and Elisseeff 2003). However potential setbacks should be noted. First, this technique is computationally intensive as it performs a complete heuristic search of the feature input space (Saeys et al. 2007). Second, as cross-validation methods are used to avoid a biased feature selection

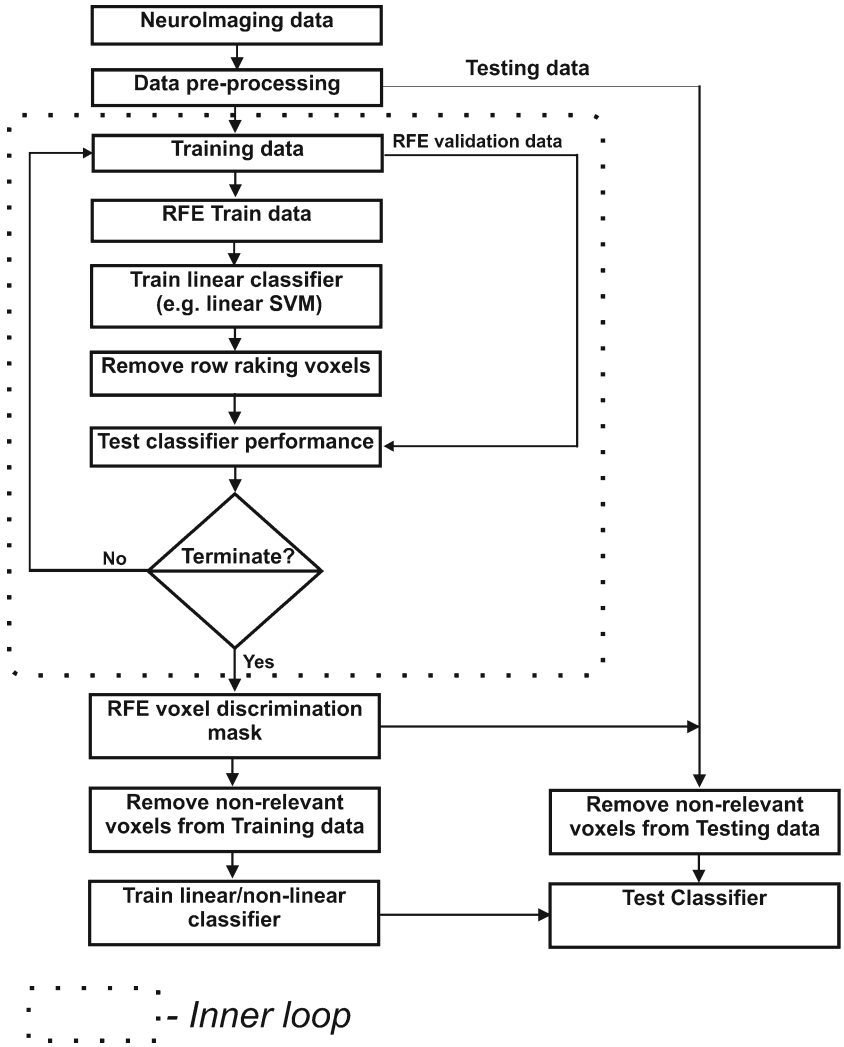

Fig. 3 Flow diagram illustrating the recursive feature elimination (RFE) process

process, this may result in different features being selected in every cross-validation iteration (Craddock et al. 2009). This problem has recently been addressed by Dosenbach and colleagues (Dosenbach et al. 2010) by recommending reporting of a *consensus* discrimination map which aggregates features selected in all cross-validation iterations.

Previous feature selection applications using RFE in neuroimaging classification tasks include; ASD (Calderoni et al. 2012; Duchesnay et al. 2011; Ecker et al. 2010; Ingalhalikar et al. 2011), MDD (Craddock et al. 2009), schizophrenia (Castro et al. 2011b), psychosis (Gothelf et al. 2011), object recognition in a fMRI task (Hanson and Halchenko 2008), fragile X syndrome (Hoeft et al. 2011), ADHD (Marquand et al. 2011), MCI (Nho et al. 2010), fMRI spatial patterns (De Martino et al. 2008), mood disorders (Mourao-Miranda et al. 2012) and AD (Davatzikos et al. 2008; Mesrob et al. 2008). An interesting variant of RFE, which involves backward-elimination of voxel clusters rather than individual features, has recently been explored (Deshpande et al. 2010). In passing, we note that although the majority of RFE applications in neuroimaging are largely in predictive classification, recently RFE has been used for *regression* tasks (Fan et al. 2010; He et al. 2008; Mwangi et al. 2013).

Remarkably, although previous RFE applications have largely utilized multivariate wrappers in *input* spaces (e.g. linear SVM), continuing efforts in solving the 'pre-image problem' (Kwok and Tsang 2004) have recently allowed extraction of *feature/voxel* weights in *non-linear* feature spaces (e.g. non-linear SVM) (Kjems et al. 2002; LaConte et al. 2005; Mwangi et al. 2013; Rasmussen et al. 2011).

Searchlight

The searchlight technique was introduced by Kriegeskorte and colleagues (Kriegeskorte et al. 2006) for multivariate feature reduction in neuroimaging data. This technique selects relevant features as follows. First, 3-dimensional (3D) spherical volumes of pre-defined radius or 'searchlight' (e.g. 4 mm) are centered at every voxel and populated through neuroimage volumes in the training data. Second, a machine learning

classifier (e.g. SVM) is trained at every searchlight volume using only voxels within the spherical searchlight volume and classifier accuracies from searchlight volumes centered at every voxel recorded. Through permutation methods (Kriegeskorte et al. 2006; Pereira et al. 2009), the searchlight accuracies at every voxel are converted into p-values, which are subsequently thresholded to remove non-relevant voxels. However, the searchlight technique requires a user to determine the spherical regions-of-interest *radius* (e.g. 4 mm), and according to Kriegeskorte and colleagues (Kriegeskorte et al. 2006), there should be a balance between the size of neuroimaging scans and the spherical searchlight volume.

Previous applications of feature reduction using searchlight in neuroimaging machine learning tasks include: episodic memory decoding (Chadwick et al. 2010), scene representation decoding (Bonnici et al. 2011) and attention shifting cortical activation decoding (Greenberg et al. 2010). We note that majority of these studies have largely been task-based fMRI decoding solutions.

Notably, the searchlight method assumes that any discriminative information lies within the searchlight radius and according to (Formisano et al. 2008), this technique may be unreliable in detecting 'distant pattern' differences (e.g. bilateral cerebral regions). However, we note that this is not the case with other multivariate wrapper methods such as RFE, which are able to detect multivariate interactions across 'distant patterns' as above.

Embedded Feature Reduction Techniques

In this section, we discuss three of the most popular embedded methods namely; least absolute shrinkage and selection operator (LASSO) (Tibshirani 1996; Tibshirani 2011), the Elastic Net (Zou and Hastie 2005) and the partial least square (PLS) method (McIntosh and Lobaugh 2004). The Elastic Net and LASSO techniques combine both machine learning and feature reduction steps by enlisting a regularization framework (e.g. L_1 and L_2 norm regularization) resulting to a reduced subset of relevant features (Zou and Hastie 2005). On the contrary, PLS selects relevant features by establishing or analyzing associations between the independent and dependent variables (e.g. brain activity or structure and behavior) (Krishnan et al. 2011).

Least Absolute Shrinkage and Selection Operator (LASSO)

We assume a two-group classification task with a set of features x_{ij}, and corresponding target labels y_i, where $i=1, 2, ..N$ represents observations (subjects) and $j=1, 2, \ldots, P$ represents the number of features (predictor variables). Additionally, we assume that the predictor variables are normalized by subtracting sample mean and dividing by the standard deviation. As a result, the LASSO computes model coefficients $\widehat{\beta}$ by minimizing the following function (Tibshirani 1996; Tibshirani 2011).

$$\sum_{i=1}^{N}\left(y_i - \sum_j x_{ij}\beta_j\right)^2 + \lambda\sum_{j=1}^{P}|\beta_j| \quad (4)$$

λ is a *user-defined* parameter which controls the balance between the model having few non-zero coefficients $\widehat{\beta}$ (sparsity) and high prediction accuracy or generalization ability. Interestingly, as λ approaches 1, the model becomes increasingly sparse meaning few 'relevant' features whilst as λ approaches 0, the model is less sparse meaning more 'relevant' features (Bunea et al. 2011). To arrive at an optimal value for λ, cross-validation procedures (e.g. k-fold or leave-one-out) are used to test a range of λ parameters with predefined steps and the parameter resulting to high model accuracies is selected. The LASSO function is solved using an optimization procedure such as the coordinate descent algorithm as explored elsewhere (Friedman et al. 2010; Tseng and Yun 2009). A solution to this optimization problem is provided by Friedman and colleagues (Friedman et al. 2010) in R and Matlab (The Mathworks, Inc) routines.

The benefits of this method in a feature reduction process are two fold. First, the LASSO yields a small set of model coefficients $\widehat{\beta}$ with majority of coefficients set to zero and corresponding features discarded from the subsequent machine learning process. Second, the LASSO is able to cope with situations where there are a large number of predictor variables (voxels) and fewer observations (subjects) as is the case in majority of neuroimaging studies (Bunea et al. 2011). Previous applications of feature selection using LASSO in neuroimaging machine learning tasks include: AD classification (Casanova et al. 2011; Kohannim et al. 2012b; Rao et al. 2011; Vounou et al. 2011; Yan et al. 2012), prediction of video stimulus scores in fMRI (Carroll et al. 2009), ASD classification (Duchesnay et al. 2011), prediction of brain characteristics using genetic data (Kohannim et al. 2012a; Kohannim et al. 2012b), prediction of pain stimuli in fMRI (Rish et al. 2010) and Gender classification (Casanova et al. 2012).

The LASSO technique has been successful in mitigating *small-n-large-p* or *curse-of-dimensionality* problems in neuroimaging albeit with several setbacks. First, if predictor variables in a group are highly correlated, the LASSO selects only one variable from the group and ignores the rest (Bunea et al. 2011; Zou and Hastie 2005). Second, the number of selected relevant features may not exceed the number of observations or samples before the model begins to saturate (Zou and Hastie 2005). To overcome these two limitations, Zou and Hastie (Zou and Hastie 2005) introduced the Elastic Net technique, which is relatively similar to LASSO, albeit with a few modifications, as explored below.

Elastic Net

The Elastic Net is formulated in a similar manner as the LASSO, but with an additional quadratic term (Zou and Hastie 2005). We assume a two-class classification problem with similar observations and predictor variables and a preceding normalization step, as above. As a result, the Elastic Net computes model coefficients $\widehat{\beta}$ by minimizing the following *objective* function (Bunea et al. 2011; Ogutu et al. 2012; Zou and Hastie 2005).

$$\sum_{i=1}^{N}\left(y_i-\sum_j x_{ij}\beta_j\right)^2+\lambda_1\sum_{j=1}^{P}|\beta_j|+\lambda_2\sum_{j=1}^{P}\beta_j^2 \tag{5}$$

Contrary to LASSO, the Elastic Net requires definition of two *user-defined* model regularization parameters (λ_1 and λ_2), which control the degree of penalization. The L_1 penalty $\sum_{j=1}^{P}\beta_j^2$ promotes sparsity in the solution, resulting in few features with non-zero weights, whilst L_1 penalty encourages stability in the solution and acts as a bound on the number of features selected (Bunea et al. 2011; Kohannim et al. 2012a; Ogutu et al. 2012; Zou and Hastie 2005). These parameters are often selected using an *objective* parameter grid-search process which evaluates a 'range' of parameters in two-dimensions (grid-search) and parameters giving the best performance selected. However, the grid-search process can be computationally intensive (Bunea et al. 2011). Similar to LASSO, the Elastic Net function is solved using an optimization procedure such as the coordinate descent algorithm also explored elsewhere (Friedman et al. 2010). An implementation of the Elastic Net solution is freely provided in both R and Matlab (The Mathworks, Inc) by Friedman and colleagues (Friedman et al. 2010).

Previous applications of feature reduction using Elastic Net in neuroimaging machine learning tasks include: AD classification (Rao et al. 2011; Shen et al. 2011; Wan et al. 2011) and treatment response predictions in ADHD (Marquand et al. 2012).

Partial Least Squares

The partial least square (PLS) feature reduction method is divided into two major categories namely; partial least squares correlation (PLSC; MacIntosh et al. 1996; Krishnan et al. 2011) and partial least squares regression (PLSR; Wold et al. 2001). Here, we discuss PLSC which is by far the most popular variant of the partial least squares method used in neuroimaging feature reduction tasks (Krishnan et al. 2011).

We recall our two-group classification task with a set of features x_{ij} stored in a matrix X and corresponding target labels y_i, stored in a matrix Y where $i=1, 2, ..N$ represents observations (subjects) and $j=1, 2, \ldots, P$ represents the number of features (predictor variables). Additionally, we assume that the predictor variables are normalized by subtracting sample mean and dividing by the standard deviation (z-scores). PLSC begins by computing the cross product of the predictor variables and target vectors as below;

$$M = Y^T X \tag{6}$$

The resulting matrix M is decomposed using singular value decomposition (SVD) (Krishnan et al. 2011; McIntosh and Misic 2013). The SVD matrix decomposition process is explored elsewhere (Bishop 2006; Krishnan et al. 2011). The SVD decomposition yields the following.

$$M = USV^T \tag{7}$$

The above decomposition results into a set of singular vectors (U-left, V-right) whilst S is a diagonal matrix representing the 'singular values' (McIntosh and Misic 2013). The left singular vectors U contains the weights or coefficients identifying the variables or voxels in matrix X making the highest contribution in explaining the relationship between the predictor variables (e.g. brain scans) and the targets (e.g. diagnostic labels or behavioural scores) (McIntosh and Misic 2013). Lastly, a set of latent variables for both the predictor variable matrix X and target labels Y are reconstructed by computing the dot product of the singular vectors and the original data as below and also explored elsewhere (Krishnan et al. 2011).

$$L_x = XV \tag{8}$$

Where L_x is a reduced set of latent variables representing the original predictor variables in X (Krishnan et al. 2011). On the contrary, L_y represents the latent variables for the target variables and computed as follows (Krishnan et al. 2011).

$$L_y = YU \tag{9}$$

At this point, the original high-dimensional predictor variables (e.g. brain scans) are now represented by a set of low dimensional latent variables. Lastly, bootstrap and permutation tests are used to identify the optimal number of latent variables as explored elsewhere (Krishnan et al. 2011; Ziegler et al. 2013; McIntosh and Misic. 2013). An implementation of the PLSC method for neuroimaging data analysis is freely provided in Matlab (The Mathworks, Inc) by Shen and colleagues at (research.baycrest.org/pls/source/). A rigorous introduction to the PLS method and accompanying pseudocodes are also provided elsewhere (Krishnan et al. 2011). Notably, as PLSC

correlates predictor variables (e.g. brain scans) and target variables (e.g. diagnosis or behavioral design), prediction tasks are handled by it's variant – partial least squares regression (PLSR) (Wold et al. 2001). The PLSR is explored elsewhere (Wold et al. 2001; Krishnan et al. 2011).

Previous applications of feature reduction using the partial least square method in neuroimaging machine learning tasks include: Age classification (Young vs Old) (Chen et al. 2009) and prediction of cognitive behavioral scores (Ziegler et al. 2013; Menzies et al. 2007; Nester et al. 2002). Notably, PLS has also been used in multimodal feature reduction tasks (Sui et al. 2012; Martinez-Montes et al. 2004).

Unsupervised Feature Reduction Techniques

Unsupervised feature reduction techniques also known as dimensionality reduction or feature extraction techniques *construct* relevant features through *linear* or *non-linear* combinations of the *original* predictor variables (features) (Lee and Verleysen 2007). In this review, we restrict our discussion to two of the most popular unsupervised dimensionality reduction techniques in neuroimaging namely, principal component analysis and independent component analysis.

Principal Component Analysis

Principal component analysis (PCA) constructs relevant features by linearly transforming correlated variables (e.g. raw voxels in a brain scan) into a smaller number of uncorrelated variables also known as principal components (Joliffe 2002). The resulting principal components are essentially linear combinations of the original data capturing most of the variance in the data (Joliffe 2002; Mourao-Miranda et al. 2005; Mourao-Miranda et al. 2006; Zhu et al. 2005). Construction of principal components from high-dimensional data begins by first, normalizing the original features by subtracting sample mean and dividing by the standard deviation. Second, Eigen decomposition of the covariance matrix from the standardized data is performed resulting in eigenvalues and eigenvectors of the covariance matrix. Third, eigenvalues are sorted in a decreasing order effectively representing decreasing variance in the data (Joliffe 2002). Lastly, principal components are constructed by multiplying the originally *normalized* data with the 'leading' eigenvectors whose exact number is a *user-defined* parameter The 'leading' eigenvectors explain most of the variance in the data. As a result, highly-dimensional neuroimaging scans with potentially many correlated voxels are now represented by relatively few uncorrelated principal components which are later used for machine learning analyses. A detailed and rigorous mathematical derivation of PCA is given elsewhere (Joliffe 2002).

PCA has been successful in extracting relevant features in neuroimaging classification studies. For example in Schizophrenia (Caprihan et al. 2008; Radulescu and Mujica-Parodi 2009; Yoon et al. 2007), AD (Lopez et al. 2011), Psychosis (Koutsouleris et al. 2009),MDD (Fu et al. 2008), ADHD (Zhu et al. 2008) and face recognition in fMRI (Mourao-Miranda et al. 2005). Most notably principal components have also been used in neuroimaging machine learning regression studies. For example, AD clinical scores prediction (Wang et al. 2010) and age predictions (Franke et al. 2010).

PCA offers two major benefits to dimensionality reduction in neuroimaging machine learning studies. First, the technique is easy to implement and computationally efficient. Second, the technique is *un-supervised-* meaning it does not require corresponding categorical or continuous labels or targets to extract relevant features. However, PCA also suffers from several setbacks. First, the user is required to define the number of principal components although Hansen and colleagues (Hansen et al. 1999) propose using the *generalization error* from a cross-validation process to select an *optimal* number of principal components. Second, as principal components are linear combinations of the original features, they may not be easily interpretable (Bunea et al. 2011). Third, whilst constructing relevant components, PCA may ignore the required outcome (e.g. discrimination of disease vs healthy) (Bunea et al. 2011). Lastly, as principal components are constructed through a linear transformation, this process may not adequately detect more complex non-linear feature interactions (Bishop 1995), such as those that may occur in neuroimaging scan data. This setback has led to the formulation of kernel-pca- which in brief is the application of PCA in a feature space created through a kernel function (Scholkopf and Smola 2002). Several studies have recently explored the application of kernel-pca to dimensionality reduction problems in neuroimaging (Rasmussen et al. 2012; Sidhu et al. 2012; Thirion and Faugeras 2003; Wang et al. 2011).

Independent Component Analysis

Independent component analysis (ICA) is a multivariate data-driven technique which belongs to the broader category of *blind-source separation* methods used to separate data into underlying *independent* information components (Stone 2004). ICA separates a set of 'mixed signals' (e.g. raw data from an fMRI scan) into a set of *independent* and relevant features (e.g. paradigm-related signals in fMRI). To achieve this goal, ICA assumes the source signals are statistically independent from an unknown but linear mixing process (Calhoun et al. 2009).

To elucidate this, first we consider a $i \times j$ matrix X where (i = number of fMRI scans in a study and j = number of voxels from the pre-processed scans). Second, we consider a $n \times j$ matrix S with rows representing (n) number of 'expected' independent components. Additionally, we represent A as a

$i \times n$ 'mixing' matrix with columns containing associated time-courses of the n components. Effectively, this becomes a 'blind-source separation' problem (Stone 2004) which can be represented by a linear model $X=AS$(Calhoun and Adali 2006). In addition, we consider the function $Y=WX$. Consequently, the goal of an ICA algorithm is to estimate the $j \times i$ 'unmixing' matrix W such that Y becomes a good approximation of the true signal sources S. ICA algorithms use high-order statistical methods to solve for independent components (ICs) and a gentle introduction to these methods is given elsewhere (Calhoun and Adali 2006; Stone 2004).

There are two broad variants of ICA applications in fMRI. The first variant depends on the dimension of priority (e.g. spatial dimension- spatial ICA, temporal dimension- temporal ICA). Majority of ICA dimensionality reduction studies in fMRI have mostly extracted *relevant* independent components from the spatial dimension (Calhoun et al. 2009) and a detailed evaluation of both spatial and temporal options is given elsewhere (Calhoun and Adali 2006). The second category of ICA further sub-divides the technique into either *individual-subject* ICA or *group-level* ICA (Calhoun et al. 2001; Calhoun and Adali 2006). In *individual-subject* ICA, each subject's data are entered into a separate ICA analysis while in *group-level* ICA one set of components for the groups are estimated and back-reconstructed from an aggregate mixing matrix to obtain *individual-subject* independent components and discriminative maps. The group-level ICA back-reconstruction step is explored elsewhere (Erhardt et al. 2011) . A Matlab (The Mathworks, Inc) implementation of group ICA for fMRI is made freely available elsewhere (Calhoun 2011).

Benefits of employing ICA in dimensionality reduction tasks should be noted. First, unlike *univariate* methods, ICA does not require an investigator to specify a regressor of interest- which may require prior knowledge and assumptions about the experiment (e.g. paradigm or model regressor in fMRI) (Calhoun and Adali 2006). Second, ICA has proved to be successful in disentangling otherwise mixed brain signals (e.g. separating physiological, motion and scanner related components) (Calhoun and Adali 2006). However, potential setbacks of ICA should also be noted. First, ICA algorithms are computationally intensive (Correa et al. 2007). Second, according to Birn and colleagues (Birn et al. 2008), current ICA algorithms may not adequately separate respiration and default mode network signals in fMRI.

Notable neuroimaging machine learning studies using ICA in dimensionality reduction include (Castro et al. 2011a; Chai et al. 2010; De Martino et al. 2007; Douglas et al. 2011; Duff et al. 2011; Ince et al. 2008; Sato et al. 2012; Tagliazucchi et al. 2012; Toussaint et al. 2012; Yang et al. 2010). A review of ICA applications in multi-modal feature reduction tasks is given elsewhere (Sui et al. 2012). Lastly we note that a significant methodological difference between PCA and ICA exists. In PCA a set of possibly correlated variables are converted into a set of uncorrelated features, whilst in ICA original variables are transformed into a set of nearly statistically independent features (Calhoun and Adali 2006). However, the main commonality between ICA and PCA and possibly a distinctive characteristic that separates them from filter, wrapper and embedded feature reduction techniques is that the former are *unsupervised* while the latter are *supervised.*

Coordinate-Based Meta-Analysis (CBMA) Techniques

In the previous sections, we have mostly focused on 'data-driven' feature reduction techniques. Here though, we discuss techniques which may rely on existing 'domain knowledge' for feature reduction. Meta-analysis techniques have previously been used to model, analyze and report brain activation or group-level differences across neuroimaging studies (Eickhoff et al. 2009). Popular meta-analysis techniques include; activation likelihood estimation (ALE) (Laird et al. 2005b), kernel-density estimation (Scott 1992) and multilevel kernel density estimation (Wager et al. 2007). A more detailed discussion of these techniques in relation to the wider meta-analysis reporting in neuroimaging is given elsewhere (Eickhoff et al. 2009; Laird et al. 2005a; Laird et al. 2005b).

Recently though, CBMA techniques have been used in modeling distributions of reported *foci* from fMRI activation studies and resulting regions-of-interest used as input features for machine learning analyses. For instance, Yarkoni and colleagues (Yarkoni et al. 2011) recently applied a CBMA technique to select relevant features in classifying working memory, emotion and pain using fMRI. The same group has recently developed a CBMA feature reduction framework known as *Neurosynth* (Mitchell 2011; Yarkoni et al. 2011). Neurosynth is a tool for automated synthesis of fMRI data as reported in published studies (Yarkoni et al. 2011). Durkat and colleagues (Dukart et al. 2012) have recently applied a CBMA technique in classifying AD subjects in multicenter studies with high generalization ability. Other studies exploring CBMA techniques in feature reduction have also been reported. For example, Dosenbach and colleagues (Dosenbach et al. 2010) used regions-of-interest (ROIs) derived from a fMRI meta-analysis as relevant features for predicting individual subject's age. Doyle and colleagues (Doyle et al. 2013) used existing domain knowledge to select the thalamus, anterior cingulate cortex and occipital cortex as relevant regions to predict the effect of a drug (ketamine) on brain activity.

Recently, Chu and colleagues (Chu et al. 2012), reported that features from ROIs selected using a priori domain knowledge (e.g. hippocampal degeneration in Alzheimer's disease) resulted to better generalization ability as compared to features selected from a data-driven approach such as t-test or RFE.

Notably, as CBMA techniques pool activation *foci* coordinates from numerous neuroimaging studies, this approach

may improve a-posteriori certainty, increase statistical power and therefore make neuroimaging studies less susceptible to type II errors (Wager et al. 2007). The latter is a common problem in neuroimaging studies with relatively small sample sizes. However, CBMA techniques may also suffer from information loss as they represent data in published studies with a high degree of sparseness (Salimi-Khorshidi et al. 2009).

Summary and Discussion

There is general consensus within the neuroimaging machine-learning community that feature reduction is an important process before training a machine learning model. The main benefits of this process are two fold. First, to remove any redundant features (voxels) plus noise a process which may improve prediction accuracy or generalization ability as well as support interpretability of study results. The latter may in turn help in generating *post-hoc* inferences.

In this review, we have noticeably distinguished feature reduction methods into two major categories namely; *supervised and unsupervised* techniques respectively. We have further subdivided supervised techniques into three subcategories namely; filter, wrapper and embedded techniques. In summary, filter techniques use statistical feature ranking criterions (e.g. labels vs feature correlations in PCC) to discard redundant features. However, these techniques have two common setbacks. First, they are not multivariate and as such they do not take into account interactions between multiple features and spatial patterns. Second, these techniques need a user to define an optimal relevant feature threshold value (e.g. absolute correlation coefficient in PCC or p-value in t-test), although previous studies have prevailed over this impediment by optimizing this parameter through cross-validation procedures. On the other hand, wrapper techniques are multivariate although computationally intensive. Notably, De Martino and colleagues recently attempted a *hybrid* combination of both filter and wrapper techniques (t-test and RFE) which had superior prediction accuracy as compared to both techniques alone. Lastly, we note that the performance of *embedded* feature reduction methods (LASSO and Elastic Net) strongly depends on the choice of penalization parameters which should be chosen through a cross-validation process (Bunea et al. 2011; Shi et al. 2007).

We note that a significant difference between supervised and unsupervised techniques exists. Unsupervised techniques *construct* features independent of the outcome of interest whilst supervised techniques choose relevant features based on their ability to detect group-level differences.

Pertinent issues that in general may determine the performance of many feature reduction techniques at a high-level should be noted. First, the need to set a threshold of optimal number of features. Second, In the event one is using a cross-validation process to train and test the model (e.g. leave-one-out), relevant features may differ from fold-to-fold. Notably, although feature reduction methods ultimately help in removing redundant data and noise, equally important and relevant features may be inadvertently removed during feature selection (Guyon and Elisseeff 2003).

In conclusion, feature reduction techniques are frequently being used in neuroimaging machine learning studies to mitigate the curse-of-dimensionality or small-n-large-p problems and maximize prediction accuracies. Lastly, whilst there are a number of studies empirically comparing different feature reduction approaches (Craddock et al. 2009; De Martino et al. 2008; Ryali et al. 2010) in neuroimaging studies, none of these studies recommend any technique as the best in all neuroimaging machine learning tasks.

Information Sharing Statement

Public repositories and software routines explicitly mentioned in this review were available in the public domain at the time of publication of this review.

Acknowledgments This research was funded by NIMH R01085667 and Pat Rutherford, Jr. Endowed Chair in Psychiatry (UT Medical School) grants to J.C.S.

Conflict of Interest J.C.S has participated in research funded by Forest, Merck, BMS and GSK. He has been a speaker for Pfizer and Abbot.

References

Allen, E. A., Erhardt, E. B., Damaraju, E., Gruner, W., Segall, J. M., Silva, R. F., Havlicek, M., Rachakonda, S., Fries, J., Kalyanam, R., Michael, A. M., Caprihan, A., Turner, J. A., Eichele, T., Adelsheim, S., Bryan, A. D., Bustillo, J., Clark, V. P., Ewing, S. W. F., Filbey, F., Ford, C. C., Hutchison, K., Jung, R. E., Kiehl, K. A., Kodituwakku, P., Komesu, Y. M., Mayer, A. R., Pearlson, G. D., Phillips, J. R., Sadek, J. R., Michael, S., Teuscher, U., Thoma, R. J., & Calhoun, V. D. (2011). A baseline for the multivariate comparison of resting-state networks. *Frontiers in systems neuroscience, 5*, 2.

Ashburner, J. (2007). A fast diffeomorphic image registration algorithm. *NeuroImage, 38*, 95–113.

Ashburner, J. (2009). Computational anatomy with the SPM software. *Magnetic Resonance Imaging, 27*, 1163–1174.

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

Balci, S., Sabuncu, M., Yoo, J., Gosh, S., Gabrieli, W., Gabrieli, J., & Golland, P. (2008). Prediction of successful memory encoding from fMRI data. *Med Image Comput Comput Assist Interv, 11*, 97–104.

Bellman, R. (1961). Adaptive control process: A guided tour:, *Princenton University*.

Birn, R. M., Murphy, K., & Bandettini, P. A. (2008). The effect of respiration variations on independent component analysis results of resting state functional connectivity. *Human Brain Mapping, 29*, 740–750.

Bishop, C. (1995). *Neural networks for pattern recognition*. New York: Oxford University Press.

Bishop, C. (2006). *Pattern recongition and machine learning*. New York: Springer.

Bonnici, H. M., Kumaran, D., Chadwick, M. J., Weiskopf, N., Hassabis, D., & Maguire, E. A. (2011). Decoding representations of scenes in the medial temporal lobes. *Hippocampus, 22*, 1143–1153.

Brammer, M. (2009). The role of neuroimaging in diagnosis and personalized medicine–current position and likely future directions. *Dialogues In Clinical Neuroscience, 11*, 389–396.

Bray, S., Chang, C., & Hoeft, F. (2009). Applications of multivariate pattern classification analyses in developmental neuroimaging of healthy and clinical populations. *Frontiers In Human Neuroscience, 3*, 32.

Brown, T. T., Kuperman, J. M., Chung, Y., Erhart, M., McCabe, C., Hagler, D. J., Venkatraman, V. K., Akshoomoff, N., Amaral, D. G., and Bloss, C. S. (2012). Neuroanatomical assessment of biological maturity. *Current Biology, 22*(18), 1693–1698.

Bunea, F., She, Y., Ombao, H., Gongvatana, A., Devlin, K., & Cohen, R. (2011). Penalized least squares regression methods and applications to neuroimaging. *NeuroImage, 55*, 1519–1527.

Calderoni, S., Retico, A., Biagi, L., Tancredi, R., Muratori, F., & Tosetti, M. (2012). Female children with autism spectrum disorder: An insight from mass-univariate and pattern classification analyses. *NeuroImage, 59*, 1013–1022.

Calhoun, V. D. (2011). Group ICA Of fMRI Toolbox(GIFT) (http://mialab.mrn.org/software/gift/), pp. http://mialab.mrn.org/software/gift/.

Calhoun, V. D., & Adali, T. L. (2006). Unmixing fMRI with independent component analysis. *IEEE Engineering In Medicine And Biology Magazine: The Quarterly Magazine Of The Engineering In Medicine & Biology Society, 25*, 79–90.

Calhoun, V. D., Adali, T., Pearlson, G. D., & Pekar, J. J. (2001). A method for making group inferences from functional MRI data using independent component analysis. *Human Brain Mapping, 14*, 140–151.

Calhoun, V. D., Liu, J., & Adali, T. L. (2009). A review of group ICA for fMRI data and ICA for joint inference of imaging, genetic, and ERP data. *NeuroImage, 45*, S163–S172.

Calhoun, V. D., Sui, J., Kiehl, K., Turner, J., Allen, E., & Pearlson, G. (2011). Exploring the psychosis functional connectome: aberrant intrinsic networks in schizophrenia and bipolar disorder. *Frontiers in Psychiatry, 2*, 75.

Caprihan, A., Pearlson, G. D., & Calhoun, V. D. (2008). Application of principal component analysis to distinguish patients with schizophrenia from healthy controls based on fractional anisotropy measurements. *NeuroImage, 42*, 675–682.

Carroll, M. K., Cecchi, G. A., Rish, I., Garg, R., & Rao, A. R. (2009). Prediction and interpretation of distributed neural activity with sparse models. *NeuroImage, 44*, 112–122.

Casanova, R., Whitlow, C. T., Wagner, B., Williamson, J., Shumaker, S. A., Maldjian, J. A., and Espeland, M. A. (2011). High dimensional classification of structural MRI Alzheimer's disease data based on large scale regularization. *Frontiers in Neuroinformatics, 5*. doi:10.3389/fninf.2011.0002.

Casanova, R., Whitlow, C., Wagner, B., Espeland, M., & Maldjian, J. (2012). Combining graph and machine learning methods to analyze differences in functional connectivity across sex. *The Open Neuroimaging Journal, 6*, 1.

Castro, E., Martanez-Ramon, M., Pearlson, G., Sui, J., & Calhoun, V. D. (2011a). Characterization of groups using composite kernels and multi-source fMRI analysis data: Application to schizophrenia. *NeuroImage, 58*, 526–536.

Castro, E., Martinez-Raman, M., Pearlson, G., Sui, J., & Calhoun, V. D. (2011b). Characterization of groups using composite kernels and multi-source fMRI analysis data: application to schizophrenia. *NeuroImage, 58*, 526–536.

Chadwick, M. J., Hassabis, D., Weiskopf, N., & Maguire, E. A. (2010). Decoding individual episodic memory traces in the human hippocampus. *Current Biology, 20*, 544–547.

Chai, J.-W., Chi-Chang Chen, C., Chiang, C.-M., Ho, Y.-J., Chen, H.-M., Ouyang, Y.-C., Yang, C.-W., Lee, S.-K., & Chang, C.-I. (2010). Quantitative analysis in clinical applications of brain MRI using independent component analysis coupled with support vector machine. *Journal Of Magnetic Resonance Imaging: JMRI, 32*, 24–34.

Chaves, R., Ramarez, J., Garriz, J. M., Lopez, M., Salas-Gonzalez, D., Alvarez, I., & Segovia, F. (2009). SVM-based computer-aided diagnosis of the Alzheimer's disease using t-test NMSE feature selection with feature correlation weighting. *Neuroscience Letters, 461*, 293–297.

Chen, K., Reiman, E. M., Huan, Z., Caselli, R. J., Bandy, D., Ayutyanont, N., & Alexander, G. E. (2009). Linking functional and structural brain images with multivariate network analyses: A novel application of the partial least square method. *NeuroImage, 47*, 602–610.

Cheng, W., Ji, X., Zhang, J., & Feng, J. (2012). Individual classification of ADHD patients by integrating multiscale neuroimaging markers and advanced pattern recognition techniques. *Frontiers In Systems Neuroscience, 6*.

Chu, C., Hsu, A.-L., Chou, K.-H., Bandettini, P., & Lin, C. (2012). Does feature selection improve classification accuracy? Impact of sample size and feature selection on classification using anatomical magnetic resonance images. *NeuroImage, 60*, 59–70.

Cohen, J. (1998). *Statistical power analysis for the behavioural sciences-2nd edition*. New Jersey: Lawrence Erbaum Associates.

Correa, N., AdalÄ, T., & Calhoun, V. D. (2007). Performance of blind source separation algorithms for fMRI analysis using a group ICA method. *Magnetic Resonance Imaging, 25*, 684–694.

Costafreda, S. G., Chu, C., Ashburner, J., & Fu, C. H. Y. (2009). Prognostic and diagnostic potential of the structural neuroanatomy of depression. *Plos One, 4*, e6353.

Costafreda, S., Fu, C., Picchioni, M., Toulopoulou, T., McDonald, C., Kravariti, E., Walsge, M., Prata, D., Murray, R., & McGuire, P. (2011). Pattern of neural responses to verbal fluency shows diagnostic specificity for schizophrenia and bipolar disorder. *BMC Psychiatry, 11*, 18.

Coutanche, M. N., Thompson-Schill, S. L., & Schultz, R. T. (2011). Multi-voxel pattern analysis of fMRI data predicts clinical symptom severity. *NeuroImage, 57*, 113–123.

Cox, D. D., & Savoy, R. L. (2003). Functional magnetic resonance imaging (fMRI) "brain reading"•: detecting and classifying distributed patterns of fMRI activity in human visual cortex. *NeuroImage, 19*, 261–270.

Craddock, R. C., Holtzheimer, P. E., 3rd, Hu, X. P., & Mayberg, H. S. (2009). Disease state prediction from resting state functional connectivity. *Magnetic Resonance In Medicine: Official Journal Of The Society Of Magnetic Resonance In Medicine/Society Of Magnetic Resonance In Medicine, 62*, 1619–1628.

Dai, Z., Yan, C., Wang, Z., Wang, J., Xia, M., Li, K., & He, Y. (2012). Discriminative analysis of early Alzheimer's disease using multimodal imaging and multi-level characterization with multi-classifier (M3). *NeuroImage, 59*, 2187–2195.

Davatzikos, C., Fan, Y., Wu, X., Shen, D., & Resnick, S. M. (2008). Detection of prodromal Alzheimer's disease via pattern classification of magnetic resonance imaging. *Neurobiology of Aging, 29*, 514–523.

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*, 177–194.

De Martino, F., Valente, G., Staeren, N. L., Ashburner, J., Goebel, R., & Formisano, E. (2008). Combining multivariate voxel selection and support vector machines for mapping and classification of fMRI spatial patterns. *NeuroImage, 43*, 44–58.

Deshpande, G., Li, Z., Santhanam, P., Coles, C. D., Lynch, M. E., Hamann, S., & Hu, X. (2010). Recursive cluster elimination based support vector machine for disease state prediction using resting state functional and effective brain connectivity. *Plos One, 5*, e14277.

Dosenbach, N. U. F., Nardos, B., Cohen, A. L., Fair, D. A., Power, J. D., Church, J. A., Nelson, S. M., Wig, G. S., Vogel, A. C., Lessov-Schlaggar, C. N., et al. (2010). Prediction of individual brain maturity using fMRI. *Science, 329*, 1358–1361.

Douglas, P. K., Harris, S., Yuille, A., & Cohen, M. S. (2011). Performance comparison of machine learning algorithms and number of independent components used in fMRI decoding of belief vs. disbelief. *NeuroImage, 56*, 544–553.

Doyle, O. M., Ashburner, J., Zelaya, F. O., Williams, S. C. R., Mehta, M. A., & Marquand, A. F. (2013). Multivariate decoding of brain images using ordinal regression. *NeuroImage, 81*, 347–357.

Duchesnay, E., Roche, A., Riviere, D., Papadopoulos, D., Cointepas, Y., and Mangin, J.-F. (2004). Population classification based on structural morphometry of cortical sulci. Paper presented at: *Biomedical Imaging: Nano to Macro*, 2004.

Duchesnay, E., Cachia, A., Roche, A., Riviere, D., Cointepas, Y., Papadopoulos-Orfanos, D., Zilbovicius, M., Martinot, J.-L., Regis, J., & Mangin, J.-F. (2007). Classification based on cortical folding patterns. *Medical Imaging, IEEE Transactions, 26*, 553–565.

Duchesnay, E., Cachia, A., Boddaert, N., Chabane, N., Mangin, J.-F., Martinot, J.-L., Brunelle, F., & Zilbovicius, M. (2011). Feature selection and classification of imbalanced datasets: Application to PET images of children with autistic spectrum disorders. *NeuroImage, 57*, 1003–1014.

Duff, E. P., Trachtenberg, A. J., Mackay, C. E., Howard, M. A., Wilson, F., Smith, S. M., and Woolrich, M. W. (2011). Task-driven ICA feature generation for accurate and interpretable prediction using fMRI. *NeuroImage, 6*(1), 189–203.

Dukart, J., Mueller, K., Barthel, H., Villringer, A., Sabri, O., and Schroeter, M. L. (2012). Meta-analysis based SVM classification enables accurate detection of Alzheimer's disease across different clinical centers using FDG-PET and MRI. *Psychiatry Research: Neuroimaging, 212*(3), 230–236.

Ecker, C., Rocha-Rego, V., Johnston, P., Mourao-Miranda, J., Marquand, A., Daly, E. M., Brammer, M. J., Murphy, C., & Murphy, D. G. (2010). Investigating the predictive value of whole-brain structural MR scans in autism: a pattern classification approach. *NeuroImage, 49*, 44–56.

Eickhoff, S., Laird, A., Grefkes, C., Wang, L., Zilles, K., & Fox, P. (2009). Coordinate-based activation likelihood estimation meta-analysis of neuroimaging data: a random-effects approach based on empirical estimates of spatial uncertainty. *Human Brain Mapping, 30*, 2907–2926.

Erhardt, E. B., Rachakonda, S., Bedrick, E. J., Allen, E. A., Adali, T., & Calhoun, V. D. (2011). Comparison of multi-subject ICA methods for analysis of fMRI data. *Human Brain Mapping, 32*, 2075–2095.

Fan, Y., Shen, D., Gur, R. C., Gur, R. E., & Davatzikos, C. (2006). COMPARE: Classification of Morphological Patterns Using Adaptive Regional Elements. *Medical Imaging, IEEE Transactions, 26*, 93–105.

Fan, Y., Rao, H., Hurt, H., Giannetta, J., Korczykowski, M., Shera, D., Avants, B., & Gee, J. (2007). Multivariate examination of brain abnormality using both structural and functional MRI. *NeuroImage, 36*, 1189–1199.

Fan, Y., Kaufer, D., and Shen, D. (2010). Joint estimation of multiple clinical variables of neurological diseases from imaging patterns. Paper presented at: *Biomedical Imaging: From Nano to Macro*, 2010 I.E. International Symposium on (IEEE).

Formisano, E., De Martino, F., & Valente, G. (2008). Multivariate analysis of fMRI time series: classification and regression of brain responses using machine learning. *Magnetic Resonance Imaging, 26*, 921–934.

Fort, G., & Lambert-Lacroix, S. (2005). Classification using partial least squares with penalized logistic regression. *Bioinformatics, 21*, 1104–1111.

Franke, K., Ziegler, G., Kloppel, S., & Gaser, C. (2010). Estimating the age of healthy subjects from T1-weighted MRI scans using kernel methods: exploring the influence of various parameters. *NeuroImage, 50*, 883–892.

Franke, K., Luders, E., May, A., Wilke, M., & Gaser, C. (2012). Brain maturation: Predicting individual BrainAGE in children and adolescents using structural MRI. *NeuroImage, 63*, 1305–1312.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software, 33*, 1.

Friston, K. J., Poline, J. B., Holmes, A. P., Frith, C. D., & Frakowiack, R. S. (1996). A multivariate analysis of PET activation studies. *Human Brain Mapping, 4*, 140–151.

Fu, C. H. Y., Mourao-Miranda, J., Costafreda, S. G., Khanna, A., Marquand, A. F., Williams, S. C. R., & Brammer, M. J. (2008). Pattern classification of sad facial processing: toward the development of neurobiological markers in depression. *Biological Psychiatry, 63*, 656–662.

Gothelf, D., Hoeft, F., Ueno, T., Sugiura, L., Lee, A. D., Thompson, P., & Reiss, A. L. (2011). Developmental changes in multivariate neuroanatomical patterns that predict risk for psychosis in 22q11.2 deletion syndrome. *Journal Of Psychiatric Research, 45*, 322–331.

Grana, M., Termenon, M., Savio, A., Gonzalez-Pinto, A., Echeveste, J., Perez, J. M., & Besga, A. (2011). Computer aided diagnosis system for Alzheimer disease using brain diffusion tensor imaging features selected by pearson's correlation. *Neuroscience Letters, 502*, 225–229.

Greenberg, A. S., Esterman, M., Wilson, D., Serences, J. T., & Yantis, S. (2010). Control of spatial and feature-based attention in frontoparietal cortex. *The Journal of Neuroscience, 30*, 14330–14339.

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning, 7*(8), 1157–1182.

Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2001). Gene selection for cancer classification using support vector machines. *Machine Learning, 46*, 389–422.

Haller, S., Bartsch, A., Nguyen, D., Rodriguez, C., Emch, J., Gold, G., Lovblad, K. O., & Giannakopoulos, P. (2010a). Cerebral microhemorrhage and iron deposition in mild cognitive impairment: susceptibility-weighted MR imaging assessment. *Radiology, 257*, 764–773.

Haller, S., Nguyen, D., Rodriguez, C., Emch, J., Gold, G., Bartsch, A., Lovblad, K. O., & Giannakopoulos, P. (2010b). Individual prediction of cognitive decline in mild cognitive impairment using support vector machine-based analysis of diffusion tensor imaging data. *Journal of Alzheimer's disease, 22*, 315–327.

Hansen, L. K., Larsen, J., Nielsen, F. Ã., Strother, S. C., Rostrup, E., Savoy, R., Lange, N., Sidtis, J., Svarer, C., & Paulson, O. B. (1999). Generalizable patterns in neuroimaging: How many principal components? *NeuroImage, 9*, 534–544.

Hanson, S. J., & Halchenko, Y. O. (2008). Brain reading using full brain support vector machines for object recognition: there is no "face" identification area. *Neural Computation, 20*, 486–503.

Haynes, J.-D., & Rees, G. (2005). Predicting the orientation of invisible stimuli from activity in human primary visual cortex. *Nature Neuroscience, 8*, 686–691.

He, W., Wang, Z., & Jiang, H. (2008). Model optimizing and feature selecting for support vector regression in time series forecasting. *Neurocomputing Machine Learning for Signal Processing (MLSP 2006)/Life System Modelling, Simulation, and Bio-inspired Computing (LSMS 2007), 72*, 600–611.

Hinrichs, C., Singh, V., Mukherjee, L., Xu, G., Chung, M. K., & Johnson, S. C. (2009). Spatially augmented LPboosting for AD classification with evaluations on the ADNI dataset. *NeuroImage, 48*, 138–149.

Hoeft, F., Walter, E., Lightbody, A. A., Hazlett, H. C., Chang, C., Piven, J., & Reiss, A. L. (2011). Neuroanatomical differences in toddler boys with fragile x syndrome and idiopathic autism. *Archives Of General Psychiatry, 68*, 295–305.

Hua, T. W., & Dougherty, E. (2009). Performance of feature-selection methods in the classification of high-dimension data. *Pattern Recognition, 42*, 409–424.

Ince, N. F., Goksu, F., Pellizzer, G., Tewfik, A., and Stephane, M. (2008). Selection of spectro-temporal patterns in multichannel MEG with support vector machines for schizophrenia classification. Paper presented at: *Engineering in Medicine and Biology Society*, 2008.

Ingalhalikar, M., Parker, D., Bloy, L., Roberts, T. P. L., & Verma, R. (2011). Diffusion based abnormality markers of pathology: toward learned diagnostic prediction of ASD. *NeuroImage, 57*, 918–927.

Johansen-Berg, H., & Behrens, T. (2009). *Diffusion MRI - from quantitative measurements to in-vivo neuroanatomy*. London: UK, Academic Press.

Johnston, B., Mwangi, B., Matthews, K., Coghill, D., and Steele, J. (2012). Predictive classification of individual magnetic resonance imaging scans from children and adolescents. *European Child & Adolescent Psychiatry (Springer Berlin/Heidelberg)*, pp. 1–12.

Joliffe, I. (2002). *Principle component analysis*. New York: Springer.

Kjems, U., Hansen, L. K., Anderson, J., Frutiger, S., Muley, S., Sidtis, J., Rottenberg, D., & Strother, S. (2002). The quantitative evaluation of functional neuroimaging experiments: Mutual information learning curves. *NeuroImage, 15*, 772–786.

Kloppel, S., Stonnington, C., Chu, C., Draganski, B., Scahill, R., Rohrer, J., Fox, N., Jack, C., Jr., Ashburner, J., & Frackowiak, R. (2008). Automatic classification of MR scans in Alzheimer's disease. *Brain: A Journal of Neurology, 131*, 681–689.

Kohannim, O., Hibar, D., Jahanshad, N., Stein, J., Hua, X., Toga, A., Jack, C., Weinen, M., and Thompson, P. (2012a). Predicting temporal lobe volume on MRI from genotypes using L1-L 2regularized regression. Paper presented at: *Biomedical Imaging (ISBI)*, 2012 9th IEEE International Symposium on (IEEE).

Kohannim, O., Hibar, D. P., Stein, J. L., Jahanshad, N., Hua, X., Rajagopalan, P., Toga, A. W., Jack, C. R., Jr., Weiner, M. W., & de Zubicaray, G. I. (2012b). Discovery and replication of gene influences on brain structure using LASSO regression. *Frontiers in Neuroscience, 6*.

Kohavi, R., & John, G. (1997). Wrappers for feature subset selection. *Artificial Intelligence, 97*, 1–2.

Koutsouleris, N., Meisenzahl, E. M., Davatzikos, C., Bottlender, R., Frodl, T., Scheuerecker, J., Schmitt, G., Zetzsche, T., Decker, P., Reiser, M., et al. (2009). Use of neuroanatomical pattern classification to identify subjects in at-risk mental states of psychosis and predict disease transition. *Archives of General Psychiatry, 66*, 700–712.

Kovalev, V. A., Petrou, M., & Suckling, J. (2003). Detection of structural differences between the brains of schizophrenic patients and controls. *Psychiatry Research: Neuroimaging, 124*, 177–189.

Kriegeskorte, N., Goebel, R., & Bandettini, P. (2006). Information-based functional brain mapping. *Proceedings of the National Academy of Sciences of the United States of America, 103*, 3863–3868.

Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S. F., & Baker, C. I. (2009). Circular analysis in systems neuroscience: the dangers of double dipping. *Nature Neuroscience, 12*, 535–540.

Kriegeskorte, N., Lindquist, M. A., Nichols, T. E., Poldrack, R. A., & Vul, E. (2010). Everything you never wanted to know about circular analysis, but were afraid to ask. *Journal of Cerebral Blood Flow & Metabolism, 30*, 1551–1557.

Krishnan, A., Williams, L. J., McIntosh, A. R., & Abdi, H. (2011). Partial Least Squares (PLS) methods for neuroimaging: A tutorial and review. *NeuroImage, 56*, 455–475.

Kwok, J. T. Y., & Tsang, I. W. H. (2004). The pre-image problem in kernel methods. *Neural Networks, IEEE Transactions, 15*, 1517–1525.

LaConte, S., Strother, S., Cherkassky, V., Anderson, J., & Hu, X. (2005). Support vector machines for temporal classification of block design fMRI data. *NeuroImage, 26*, 317.

Laird, A., Lancaster, J., & Fox, P. (2005a). BrainMap: the social evolution of a functional neuroimaging database. *Neuroinformatics, 3*, 65–78.

Laird, A. R., McMillan, K. M., Lancaster, J. L., Kochunov, P., Turkeltaub, P. E., Pardo, J. V., & Fox, P. T. (2005b). A comparison of label-based review and ALE meta-analysis in the Stroop task. *Human Brain Mapping, 25*, 6–21.

Langs, G., Menze, B. H., Lashkari, D., & Golland, P. (2011). Detecting stable distributed patterns of brain activation using Gini contrast. *NeuroImage, 56*, 497–507.

Lee, J., & Verleysen, M. (2007). *Nonlinear Dimensionality Reduction*. New York: USA, Springer Publishing Co.

Lim, L., Marquand, A., Cubillo, A. A., Smith, A. B., Chantiluke, K., Simmons, A., Mehta, M., & Rubia, K. (2013). Disorder-specific predictive classification of adolescents with Attention Deficit Hyperactivity Disorder (ADHD) relative to autism using structural magnetic resonance imaging. *PLOS ONE, 8*, e63660.

Linden, D. E. J. (2012). The challenges and promise of neuroimaging in psychiatry. *Neuron, 73*, 8–22.

Liu, M., Zhang, D., & Shen, D. (2012). Ensemble sparse classification of Alzheimer's disease. *NeuroImage, 60*, 1106–1116.

Lohmann, G., Volz, K. G., & Ullsperger, M. (2007). Using non-negative matrix factorization for single-trial analysis of fMRI data. *NeuroImage, 37*, 1148–1160.

Lopez, M., Ramirez, J., Garriz, J., Alvarez, I., Salas-Gonzalez, D., Segovia, F., Chaves, R., Padilla, P., & Gomez-Rao, M. (2011). Principal component analysis-based techniques and supervised classification schemes for the early detection of Alzheimer's disease. *Neurocomputing, 74*, 1260–1271.

MacIntosh, A. R., Bookstei, F., Haxby, J., & Grady, C. (1996). Spatial pattern analysis of functional brain images using partial least squares. *NeuroImage, 3*, 143–157.

Magnin, B., Mesrob, L., Kinkingnahun, S., Palagrini-Issac, M., Colliot, O., Sarazin, M., Dubois, B., Leharicy, S., & Benali, H. (2009). Support vector machine-based classification of Alzheimer's disease from whole-brain anatomical MRI. *Neuroradiology, 51*, 73–83.

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

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

Marquand, A. F., De Simoni, S., O'Daly, O. G., Williams, S. C. R., Mourao-Miranda, J., & Mehta, M. A. (2011). Pattern classification of working memory networks reveals differential effects of methylphenidate, atomoxetine, and placebo in healthy volunteers. *Neuropsychopharmacology, 36*, 1237–1247.

Marquand, A. F., O'Daly, O. G., De Simoni, S., Alsop, D. C., Maguire, R. P., Williams, S. C. R., Zelaya, F. O., & Mehta, M. A. (2012). Dissociable effects of methylphenidate, atomoxetine and placebo on regional cerebral blood flow in healthy volunteers at rest: A multi-class pattern recognition approach. *NeuroImage, 60*, 1015–1024.

Martinez-Montes, E., Valdes-Sosa, P. A., Miwakeichi, F., Goldman, R. I., & Cohen, M. S. (2004). EEG/fMRI analysis by multiway partial least squares. *NeuroImage, 22*, 1023–34.

McIntosh, A. R., & Lobaugh, N. J. (2004). Partial least squares analysis of neuroimaging data: applications and advances. *NeuroImage, 23*, 250–263.

McIntosh, A. R., & Misic, B. (2013). Multivariate statistical analyses for neuroimaging data. *Annual Review of Psychology, 64*, 499–525.

Menzies, L., Achard, S., Chamberlain, S. R., Fineberg, N., Chen, C.-H., delCampo, N., Sahakian, B. J., Robbins, T. W., & Bullmore, E. (2007). Neurocognitive endophenotypes of obsessive-compulsive disorder. *Brain, 130*, 3223–3236.

Mesrob, L., Magnin, B., Colliot, O., Sarazin, M., Hahn-Barma, V., Dubois, B., Gallinari, P., Leharicy, S., Kinkingnhun, S., and Benali, H. (2008). Identification of Atrophy Patterns in Alzheimers Disease Based on SVM Feature Selection and Anatomical Parcellation. *Medical Imaging and Augmented Reality, 5128*, 124–132.

Misaki, M., Kim, Y., Bandettini, P. A., & Kriegeskorte, N. (2010). Comparison of multivariate classifiers and response normalizations for pattern-information fMRI. *NeuroImage, 53*, 103–118.

Mitchell, T. M. (2011). From journal articles to computational models: a new automated tool. *Nature methods, 8*, 627.

Mitchell, T. M., Hutchinson, R., Niculescu, R. S., Pereira, F., Wang, X., Just, M., & Newman, S. (2004). Learning to decode cognitive states from brain images. *Machine Learning, 57*, 145–175.

Mourao-Miranda, J., Bokde, A. L. W., Born, C., Hampel, H., & Stetter, M. (2005). Classifying brain states and determining the discriminating activation patterns: Support Vector Machine on functional MRI data. *NeuroImage, 28*, 980–995.

Mourao-Miranda, J., Reynaud, E., McGlone, F., Calvert, G., & Brammer, M. (2006). The impact of temporal compression and space selection on SVM analysis of single-subject and multi-subject fMRI data. *NeuroImage, 33*, 1055–1065.

Mourao-Miranda, J., Oliveira, L., Ladouceur, C. D., Marquand, A., Brammer, M., Birmaher, B., Axelson, D., & Phillips, M. L. (2012). Pattern recognition and functional neuroimaging help to discriminate healthy adolescents at risk for mood disorders from low risk adolescents. *Plos One, 7*, e29482.

Mwangi, B., Ebmeier, K., Matthews, K., & Douglas Steele, J. (2012a). Multi-centre diagnostic classification of individual structural neuroimaging scans from patients with major depressive disorder. *Brain: A Journal of Neurology, 135*, 1508–1521.

Mwangi, B., Matthews, K., & Steele, J. (2012b). Prediction of illness severity in patients with major depression using structural MR brain scans. *Journal of Magnetic Resonance Imaging, 35*, 64–71.

Mwangi, B., Hasan, K. M., and Soares, J. C. (2013). Prediction of individual subject's age across the human lifespan using diffusion tensor imaging: A machine learning approach. *NeuroImage, 75*, 58–67.

Nester, P. G., O'Donnell, B. F., Mccarley, R. W., Niznikeiwicz, M., Barnard, J., Shen, Z. J., Bookstein, F. L., & Shenton, M. E. (2002). A new statistical method for testing hypotheses of neuropsychological/MRI relationships in schizophrenia: partial least squares analysis. *Schizophrenia Research, 53*, 57–66.

Nho, K., Shen, L., Kim, S., Risacher, S. L., West, J. D., Foroud, T., Jack, C. R., Weiner, M. W., & Saykin, A. J. (2010). Automatic prediction of conversion from mild cognitive impairment to probable alzheimer's disease using structural magnetic resonance imaging. *AMIA Annual Symposium Proceedings/AMIA Symposium AMIA Symposium, 2010*, 542–546.

Norman, K. A., Polyn, S. M., Detre, G. J., & Haxby, J. V. (2006). Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends in Cognitive Sciences, 10*, 424–430.

Ogutu, J. O., Schulz-Streeck, T., and Piepho, H. P. (2012). Genomic selection using regularized linear regression models: ridge regression, lasso, elastic net and their extensions. Paper presented at: *BMC proceedings (Springer)*.

Orru, G., Pettersson-Yeo, W., Marquand, A. F., Sartori, G., & Mechelli, A. (2012). Using support vector machine to identify imaging biomarkers of neurological and psychiatric disease: A critical review. *Neuroscience And Biobehavioral Reviews, 36*, 1140–1152.

Penny, W., Friston, K., Ashburner, J., and Nicols, T. (2007). *Statistical parametric mapping: The Analysis of functional Brain images.* London: Academic Press

Pereira, F., Mitchell, T., & Botvinick, M. (2009). Machine learning classifiers and fMRI: A tutorial overview. *NeuroImage Mathematics in Brain Imaging, 45*, S199–S209.

Radulescu, A. R., & Mujica-Parodi, L. R. (2009). A principal component network analysis of prefrontal-limbic functional magnetic resonance imaging time series in schizophrenia patients and healthy controls. *Psychiatry Research, 174*, 184–194.

Rao, A., Lee, Y., Gass, A., and Monsch, A. (2011). Classification of Alzheimer's Disease from structural MRI using sparse logistic regression with optional spatial regularization. Paper presented at: *Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE (IEEE)*.

Rasmussen, C., and Williams, C. (2006). Gaussian Processes for Machine Learning *MIT Press*

Rasmussen, P. M., Madsen, K. H., Lund, T. E., & Hansen, L. K. (2011). Visualization of nonlinear kernel models in neuroimaging by sensitivity maps. *NeuroImage, 55*, 1120–1131.

Rasmussen, P. M., Abrahamsen, T. J., Madsen, K. H., & Hansen, L. K. (2012). Nonlinear denoising and analysis of neuroimages with kernel principal component analysis and pre-image estimation. *NeuroImage, 60*, 1807–1818.

Rish, I., Cecchi, G., Baliki, M., & Apkarian, A. (2010). Sparse regression models of pain perception. *Brain Informatics, 212–223*.

Rizk-Jackson, A., Stoffers, D., Sheldon, S., Kuperman, J., Dale, A., Goldstein, J., Corey-Bloom, J., Poldrack, R. A., & Aron, A. R. (2011). Evaluating imaging biomarkers for neurodegeneration in pre-symptomatic Huntington's disease using machine learning techniques. *NeuroImage, 56*, 788–796.

Ryali, S., Supekar, K., Abrams, D. A., & Menon, V. (2010). Sparse logistic regression for whole brain classification of fMRI data. *NeuroImage, 51*, 752.

Saeys, Y., Inza, I., and Larranaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics, 23*(19), 2507–2517.

Salimi-Khorshidi, G., Smith, S. M., Keltner, J. R., Wager, T. D., & Nichols, T. E. (2009). Meta-analysis of neuroimaging data: a comparison of image-based and coordinate-based pooling of studies. *NeuroImage, 45*, 810–823.

Sato, J. R., Hoexter, M. Q., Fujita, A., and Rohde, L. A. (2012). Evaluation of pattern recognition and feature extraction methods in ADHD prediction. *Frontiers in Systems Neuroscience, 6*(68). doi:10.3389/fnsys.2012.00068.

Scholkopf, B., & Smola, A. (2002). *Learning with Kernels: Support vector machines, regularization, optimization, and beyond*. Cambridge: MIT Press.

Schrouff, J., Rosa, M. J., Rondina, J., Marquand, A., Chu, C., Ashburner, J., Phillips, C., Richiardi, J., and Mourao-Miranda, J. (2013). PRoNTo: Pattern Recognition for Neuroimaging Toolbox. *Neuroinformatics, 11*(3), 319–339. doi:10.1007/s12021-013-9178-1.

Scott, D. (1992). *Multivariate density estimation: Theory, practice and visualization*. New York: Wiley.

Shen, L., Kim, S., Qi, Y., Inlow, M., Swaminathan, S., Nho, K., Wan, J., Risacher, S., Shaw, L., and Trojanowski, J. (2011). Identifying neuroimaging and proteomic biomarkers for MCI and AD via the elastic net. *Multimodal Brain Image Analysis, 7012*, 27–34.

Sheskin, D. (2004). *Handbook of parametric and nonparametric statistical procedures*. Florida: Chapman & Hall.

Shi, W., Lee, K. E., and Wahba, G. (2007). Detecting disease-causing genes by LASSO-Patternsearch algorithm. Paper presented at: *BMC proceedings*, (BioMed Central Ltd).

Sidhu, G. S., Asgarian, N., Greiner, R., & Brown, M. R. G. (2012). Kernel principal component analysis for dimensionality reduction in fMRI-based diagnosis of ADHD. *Frontiers in Systems Neuroscience, 6*, 74.

Stone, J. (2004). *Independent component analysis*. Cambridge: MIT Press.

Stonnington, C. M., Chu, C., Kloppel, S., Jack, C. R., Jr., Ashburner, J., & Frackowiak, R. S. J. (2010). Predicting clinical scores from

magnetic resonance scans in Alzheimer's disease. *NeuroImage, 51*, 1405–1413.

Strother, S. C., Anderson, J., Hansen, L. K., Kjems, U., Kustra, R., Sidtis, J., Frutiger, S., Muley, S., LaConte, S., & Rottenberg, D. (2002). The quantitative evaluation of functional neuroimaging experiments: The NPAIRS data analysis framework. *NeuroImage, 15*, 747–771.

Sui, J., Adali, T., Yu, Q., Chen, J., & Calhoun, V. D. (2012). A review of multivariate methods for multimodal fusion of brain imaging data. *Journal of Neuroscience Methods, 204*, 68–81.

Tagliazucchi, E., von Wegner, F., Morzelewski, A., Borisov, S., Jahnke, K., and Laufs, H. (2012). Automatic sleep staging using fMRI functional connectivity data. *Neuroimage, 63*(1) 63–72.

Theodoridis, S., & Koutroumbas, K. (2009). *Pattern Recognition* (4th ed.). California: Elseiver.

Thirion, B., & Faugeras, O. (2003). Dynamical components analysis of fMRI data through kernel PCA. *NeuroImage, 20*, 34–49.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B (Methodological), 267–288*.

Tibshirani, R. (2011). Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 73*, 273–282.

Tipping, M. (2001). Sparse Bayesian Learning and the relevance vector machine. *J ournal of Machine Learning Research, 1*(211), 244.

Toussaint, P. J., Perlbarg, V., Bellec, P., Desarnaud, S., Lacomblez, L., Doyon, J., Habert, M. O., and Benali, H. (2012). Resting state FDG-PET functional connectivity as an early biomarker of Alzheimer's disease using conjoint univariate and independent component analyses. *NeuroImage, 63*(2) 936–946.

Tseng, P., & Yun, S. (2009). Block-coordinate gradient descent method for linearly constrained nonsmooth separable optimization. *Journal of optimization theory and applications, 140*, 513–535.

Valente, G., De Martino, F., Esposito, F., Goebel, R., & Formisano, E. (2011). Predicting subject-driven actions and sensory experience in a virtual world with Relevance Vector Machine Regression of fMRI data. *NeuroImage, 56*, 651–661.

Van De Ville, D., & Lee, S.-W. (2012). Brain decoding: Opportunities and challenges for pattern recognition. *Pattern Recognition Brain Decoding, 45*, 2033–2034.

Vapnik, V. (1999). *The nature of statistical learning theory-2nd edition*. New York: Springer.

Vounou, M., Janousova, E., Wolz, R., Stein, J. L., Thompson, P. M., Rueckert, D., and Montana, G. (2011). Sparse reduced-rank regression detects genetic associations with voxel-wise longitudinal phenotypes in Alzheimer's disease. *Neuroimage, 60*(1) 700–716.

Wager, T. D., Lindquist, M., & Kaplan, L. (2007). Meta-analysis of functional neuroimaging data: current and future directions. *Social Cognitive and Affective Neuroscience, 2*, 150–158.

Wan, J., Kim, S., Inlow, M., Nho, K., Swaminathan, S., Risacher, S., Fang, S., Weiner, M., Beg, M., & Wang, L. (2011). Hippocampal surface mapping of genetic risk factors in AD via sparse learning models. *Medical Image Computing and Computer-Assisted Intervention, MICCAI, 2011*, 376–383.

Wang, Y., Fan, Y., Bhatt, P., & Davatzikos, C. (2010). High-dimensional pattern regression using machine learning: from medical images to continuous clinical variables. *NeuroImage, 50*, 1519–1535.

Wang, X., Jiao, Y., and Lu, Z. (2011). Discriminative analysis of resting-state brain functional connectivity patterns of Attention-Deficit Hyperactivity Disorder using Kernel Principal Component Analysis. Paper presented at: Fuzzy Systems and Knowledge Discovery (FSKD), 2011 Eighth International Conference on (IEEE).

Wee, C.-Y., Yap, P.-T., Li, W., Denny, K., Browndyke, J. N., Potter, G. G., Welsh-Bohmer, K. A., Wang, L., & Shen, D. (2011). Enriched white matter connectivity networks for accurate identification of MCI patients. *NeuroImage, 54*, 1812–1822.

Wee, C.-Y., Yap, P.-T., Zhang, D., Denny, K., Browndyke, J. N., Potter, G. G., Welsh-Bohmer, K. A., Wang, L., & Shen, D. (2012). Identification of MCI individuals using structural and functional connectivity networks. *NeuroImage, 59*, 2045–2056.

Wold, S., Sjostrom, M., & Eriksson, L. (2001). PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems, 58*, 109–130.

Yan, J., Risacher, S., Kim, S., Simon, J., Li, T., Wan, J., Wang, H., Huang, H., Saykin, A., and Shen, L. (2012). Multimodal Neuroimaging Predictors for Cognitive Performance Using Structured Sparse Learning. *Multimodal Brain Image Analysis, 7509*, 1–17.

Yang, H., Liu, J., Sui, J., Pearlson, G., and Calhoun, V. D. (2010). A hybrid machine learning method for fusing fMRI and genetic data: combining both improves classification of schizophrenia. *Frontiers in human neuroscience*, *4*. doi:10.3389/fnhum.2010.00192.

Yarkoni, T., Poldrack, R. A., Nichols, T. E., Van Essen, D. C., & Wager, T. D. (2011). Large-scale automated synthesis of human functional neuroimaging data. *Nature methods, 8*, 665–670.

Yoon, U., Lee, J. M., Im, K., Shin, Y. W., Cho, B. H., Kim, I. Y., Kwon, J. S., & Kim, S. I. (2007). Pattern classification using principal components of cortical thickness and its discriminative pattern in schizophrenia. *NeuroImage, 34*, 1405–1415.

Yoon, J. H., Tamir, D., Minzenberg, M. J., Ragland, J. D., Ursu, S., and Carter, C. S. (2008). Multivariate Pattern Analysis of Functional Magnetic Resonance Imaging Data Reveals Deficits in Distributed Representations in Schizophrenia. *Biological Psychiatry, 64*(12) 1035–1041.

Zeng, L.-L., Shen, H., Liu, L., Wang, L., Li, B., Fang, P., Zhou, Z., Li, Y., & Hu, D. (2012). Identifying major depression using whole-brain functional connectivity: a multivariate pattern analysis. *Brain, 135*, 1498–1507. doi:10.1093/brain/aws059.

Zhang, D., Wang, Y., Zhou, L., Yuan, H., & Shen, D. (2011a). Multimodal classification of Alzheimer's disease and mild cognitive impairment. *NeuroImage, 55*, 856–867.

Zhang, D., Wang, Y., Zhou, L., Yuan, H., & Shen, D. (2011b). Multimodal classification of Alzheimer's disease and mild cognitive impairment. *NeuroImage, 55*, 856–867.

Zhu, C. Z., Zang, Y. F., Liang, M., Tian, L. X., He, Y., Li, X. B., Sui, M. Q., Wang, Y. F., & Jiang, T. Z. (2005). Discriminative analysis of brain function at resting-state for attention-deficit/hyperactivity disorder. *Medical Image Computing and Computer-Assisted Intervention: MICCAI International Conference On Medical Image Computing And Computer-Assisted Intervention, 8*, 468–475.

Zhu, C. Z., Zang, Y. F., Cao, Q. J., Yan, C. G., He, Y., Jiang, T. Z., Sui, M. Q., & Wang, Y. F. (2008). Fisher discriminative analysis of resting-state brain function for attention-deficit/hyperactivity disorder. *NeuroImage, 40*, 110–120.

Ziegler, G., Dahnke, R., Winkler, A. D., & Gaser, C. (2013). Partial least squares correlation of multivariate cognitive abilities and local brain structure in children and adolescents. *NeuroImage, 82*, 284–294.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67*, 301–320.