Learning Tensor-Based Features for Whole-Brain fMRI Classification

Xiaonan Song, Lingnan Meng, Qiquan Shi, and Haiping Lu*

Department of Computer Science, Hong Kong Baptist University, Hong Kong
haiping@hkbu.edu.hk

Abstract. This paper presents a novel tensor-based feature learning approach for whole-brain fMRI classification. Whole-brain fMRI data have high exploratory power, but they are challenging to deal with due to large numbers of voxels. A critical step for fMRI classification is dimensionality reduction, via *feature selection* or *feature extraction*. Most current approaches perform voxel selection based on *feature selection* methods. In contrast, *feature extraction* methods, such as principal component analysis (PCA), have limited usage on whole brain due to the small sample size problem and limited interpretability. To address these issues, we propose to directly extract features from natural tensor (rather than vector) representations of whole-brain fMRI using multilinear PCA (MPCA), and map MPCA bases to voxels for interpretability. Specifically, we extract low-dimensional tensors by MPCA, and then select a number of MPCA features according to the captured variance or mutual information as the input to SVM. To provide interpretability, we construct a mapping from the selected MPCA bases to raw voxels for localizing discriminating regions. Quantitative evaluations on challenging multiclass tasks demonstrate the superior performance of our proposed methods against the state-of-the-art, while qualitative analysis on localized discriminating regions shows the spatial coherence and interpretability of our mapping.

1 Introduction

Over the past decades, functional Magnetic Resonance Imaging (fMRI) has emerged as a powerful instrument to collect vast quantities of data for measuring brain activities. It becomes a popular tool in applications such as brain state encoding/decoding and brain disease detection, including Alzheimer's disease, Mild Cognitive Impairment, and Autism Spectrum Disorder [2,14]. Most existing studies on fMRI classification restrict the analysis to specific brain *regions of interest* (ROIs). However, ROI analysis is labor-intensive, subject to human error, and requires the assumption that a functionally active brain region will be within an anatomically standardized index [10]. In contrast, *whole-brain* fMRI data have higher exploratory power and lower bias (with no prior user-dependent hypothesis/selection of spatial voxels) [5,17], and recent works reported promising results on whole-brain-based classification [1,16,17]. Inspired by these works, this paper focuses on whole-brain fMRI classification.

* Corresponding author.

N. Navab et al. (Eds.): MICCAI 2015, Part I, LNCS 9349, pp. 613–620, 2015.
DOI: 10.1007/978-3-319-24553-9_75

It is challenging to analyze all voxels in the whole brain. The number of whole-brain voxels usually far exceeds the number of observations available in practice, leading to *overfitting* [17]. We need to perform dimensionality reduction first, through either *feature selection* or *feature extraction* [12].

Feature selection methods are more popular for fMRI classification, partly due to their good interpretability. There are two main approaches: univariate and multivariate feature selection [14]. For the *univariate* approach, mutual information (MI) is a popular choice [3,15], e.g., Chou et al. [3] select informative fMRI voxels with high MI values individually for brain state decoding and report good improvement in classification accuracy. In contrast, the *multivariate* methods consider interactions between multiple features, e.g., Ryali et al. [17] and Kampa et al. [7] present sparse optimization frameworks for whole-brain fMRI feature selection and demonstrate the effectiveness of logistic regression (LR) with the elastic net penalty, which outperforms LR with ℓ_1-norm regularization and recursive feature elimination [16], and serves as the *state-of-the-art*.

The other dimensionality reduction approach is feature extraction. Principal component analysis (PCA) is arguably the most popular *linear* feature extraction method. To apply PCA to whole-brain fMRI, we need to concatenate all voxels into a very high-dimensional vector, making the small sample size problem more severe. Moreover, though individual PCA bases can be well interpreted [6], a group of PCA bases are seldom interpreted together effectively [12,13]. On the other hand, *multilinear* feature extraction methods, such as the multilinear PCA (MPCA) [9], are getting popular recently. They represent *multidimensional data* as tensors rather than vectors, with *three key benefits*: preserved multidimensional structure, lower computational demand, and less parameters to estimate. For example, for 3D $128 \times 128 \times 64$ volumes, a PCA basis needs $128 \times 128 \times 64 = 1,048,576$ parameters, while an MPCA basis needs only $128 + 128 + 64 = 320$ parameters [8]. fMRI data are *multidimensional* so it is more intuitive to analyze them using tensor representations [1,8].

In this paper, we propose a novel tensor-based feature learning approach via MPCA for whole-brain fMRI classification and a new mapping scheme to localize discriminating regions based on MPCA features. We perform evaluations on a challenging multiclass fMRI dataset [11]. Our contributions are twofold:

- Our methods directly extract features from tensor representations of fMRI using MPCA for the three key benefits mentioned above. The extracted MPCA features are then selected according to variance or mutual information to be fed into the Support Vector Machine (SVM). Superior performance on both binary and multiclass tasks is achieved without requiring vectorization or a priori identification of localized ROIs.
- Our mapping scheme localizes discriminating regions in the voxel space via MPCA bases for interpretability. It is different from the scheme in [12] of mapping the coefficients of the optimal hyperplane in linear SVM, which tends to be noisy and fragmented, as pointed out in [4]. We can obtain spatial maps with good spatial coherence and good interpretability for neuroscience.

2 Methods

Our proposed methods use the fMRI data represented by the mean percent signal change (PSC) over the time dimension [11] as input features and model them directly as third-order tensors (3D data). We use MPCA to learn multilinear bases from these tensorial input to obtain low-dimensional tensorial MPCA features. We then select the most informative MPCA features to form feature vectors for the SVM classifier. We present the key steps of our methods in detail below.

Notations and Basic Operations. Following [8], we denote vectors by lowercase boldface letters, e.g., $\mathbf{x}$; matrices by uppercase boldface, e.g., $\mathbf{X}$; and tensors by calligraphic letters, e.g., $\mathcal{X}$. An index is denoted with a lowercase letter, spanning the range from 1 to the uppercase letter of the index, e.g., $i = 1, \ldots, I$. We denote an Nth-order tensor as $\mathcal{A} \in \mathbb{R}^{I_1 \times \cdots \times I_N}$ and their elements with indices in parentheses $\mathcal{A}(i_1, \ldots, i_N)$. The n-mode index is denoted with i_n, $n = 1, \ldots, N$. The n-mode product of a tensor $\mathcal{A}$ by a matrix $\mathbf{U} \in \mathbb{R}^{J_n \times I_n}$, is written as $\mathcal{B} = \mathcal{A} \times_n \mathbf{U}$, with its entries obtained as [8]:

$$\mathcal{B}(i_1, \ldots, i_{n-1}, j_n, i_{n+1}, \ldots, i_N) = \sum_{i_n} \mathcal{A}(i_1, \ldots, i_N)\mathbf{U}(j_n, i_n), j_n = 1, \ldots, J_n. \tag{1}$$

The scalar product of two tensors $\langle \mathcal{A}, \mathcal{B} \rangle \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$ is defined as:

$$\langle \mathcal{A}, \mathcal{B} \rangle = \sum_{i_1} \cdots \sum_{i_N} \mathcal{A}(i_1, \ldots, i_N)\mathcal{B}(i_1, \ldots, i_N). \tag{2}$$

A rank-one tensor $\mathcal{U}$ equals to the outer product of N vectors [8]:

$$\mathcal{U} = \mathbf{u}^{(1)} \circ \cdots \circ \mathbf{u}^{(N)}, \text{ where } \mathcal{U}(i_1, \ldots, i_N) = \mathbf{u}^{(1)}(i_1) \cdots \mathbf{u}^{(N)}(i_N). \tag{3}$$

MPCA Feature Extraction. MPCA [9] is an unsupervised learning method to learn features directly from tensorial representations of multidimensional data. Thus, we represent our M training fMRI samples as third-order tensors $\{\mathcal{X}_1, \ldots, \mathcal{X}_M \in \mathbb{R}^{I_1 \times I_2 \times I_3}\}$ as input to MPCA. MPCA then extracts low-dimensional tensor features $\{\mathcal{Y}_1, \ldots, \mathcal{Y}_M \in \mathbb{R}^{P_1 \times P_2 \times P_3}\}$ through three ($N = 3$) projection matrices $\{\mathbf{U}^{(n)} \in \mathbb{R}^{I_n \times P_n}, n = 1, 2, 3\}$ as follows:

$$\mathcal{Y}_m = \mathcal{X}_m \times_1 \mathbf{U}^{(1)^T} \times_2 \mathbf{U}^{(2)^T} \times_3 \mathbf{U}^{(3)^T}, m = 1, \ldots, M, \tag{4}$$

where $P_n < I_n$. In this way, the tensor dimensions are reduced from $I_1 \times I_2 \times I_3$ to $P_1 \times P_2 \times P_3$. The solutions for the projection matrices $\{\mathbf{U}^{(n)}\}$ are obtained via maximizing the total tensor scatter $\Psi_{\mathcal{Y}} = \sum_{m=1}^{M} \parallel \mathcal{Y}_m - \bar{\mathcal{Y}} \parallel_F^2$, where $\bar{\mathcal{Y}} = \frac{1}{M}\sum_{m=1}^{M} \mathcal{Y}_m$ is the *mean tensor feature* and $\parallel \cdot \parallel_F$ is the Frobenius norm [9]. This problem is solved through an iterative alternating projection method in [9]. Each iteration involves N modewise eigendecompositions to get the n-mode eigenvalues and eigenvectors. We denote the i_nth n-mode eigenvalue as $\lambda_{i_n}^{(n)}$.

There are two parameters to set in MPCA. One is Q for determining the tensor subspace dimensions $\{P_1, P_2, P_3\}$. Specifically, the first P_n eigenvectors

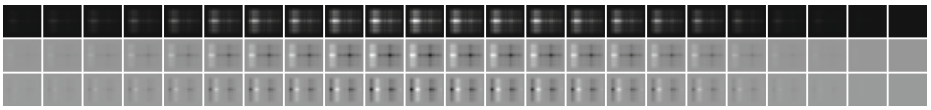


Fig. 1. Illustration of three selected eigentensors (MPCA features), where each row corresponds to an eigentensor with the third (depth) dimension concatenated

are kept in the n-mode so that the same (or similar) amount of variances is kept in each mode: $Q^{(1)} = Q^{(2)} = Q^{(3)} = Q$, where $Q^{(n)}$ is the ratio of variances kept in the n-mode defined as $Q^{(n)} = \sum_{i_n=1}^{P_n} \lambda_{i_n}^{(n)*} / \sum_{i_n=1}^{I_n} \lambda_{i_n}^{(n)*}$, and $\lambda_{i_n}^{(n)*}$ is the i_nth n-mode eigenvalue in the full projection [9]. The second parameter is the maximum number of iterations K, which can be safely set to 1 following [9].

MPCA Feature Selection. The MPCA projection matrices $\{\mathbf{U}^{(n)}, n = 1, 2, 3\}$ can be viewed as $P_1 \times P_2 \times P_3$ *eigentensors* [9] using (3):

$$\mathcal{U}_{p_1 p_2 p_3} = \mathbf{u}_{p_1}^{(1)} \circ \mathbf{u}_{p_2}^{(2)} \circ \mathbf{u}_{p_3}^{(3)} \in \mathbb{R}^{I_1 \times I_2 \times I_3},\ p_n = 1, \ldots, P_n, \tag{5}$$

where $\mathbf{u}_{p_n}^{(n)}$ is the p_nth column of $\mathbf{U}^{(n)}$. Each eigentensor $\mathcal{U}_{p_1 p_2 p_3}$ is an MPCA feature and it can be mapped to the voxel space. Figure 1 illustrates three eigentensors capturing the most variance of the fMRI data studied in this paper. Each eigentensor is shown in a row by concatenating the third dimension. It is rich in structure because it is a rank-one tensor. Since our objective is whole-brain fMRI classification, it will be beneficial to select the P most informative (rather than all) features to be fed into a classifier such as the SVM [9].

Therefore, we further perform feature selection based on an importance score using either the variance or the MI criterion. We arrange the entries in $\{\mathcal{Y}_m\}$ into feature vectors $\{\mathbf{y}_m\}$ according to the importance score in descending order. Only the first P entries of $\{\mathbf{y}_m\}$ are selected as SVM input. We can determine the optimal value for P via cross-validation. For convenience, we denote the eigentensor corresponding to the pth selected feature as $\mathcal{U}_p$ so the pth feature y_p can be written as $y_p = \langle \mathcal{X}, \mathcal{U}_p \rangle$ using (2).

The **variance** is an *unsupervised* criterion. We obtain the variance $S_{p_1 p_2 p_3}$ captured by the eigentensor $\mathcal{U}_{p_1 p_2 p_3}$ using a scatter measure as

$$S_{p_1 p_2 p_3} = \sum_{m=1}^{M} \left[\mathcal{Y}_m(p_1, p_2, p_3) - \bar{\mathcal{Y}}(p_1, p_2, p_3) \right]^2. \tag{6}$$

The **MI** is a criterion to quantify statistical dependence between two discrete random variables A and B (for example) as [3]:

$$\mathrm{MI}(A; B) = \sum_{a \in A} \sum_{b \in B} p(a, b) \log \frac{p(a, b)}{p(a) p(b)}, \tag{7}$$

where $p(a, b)$ is the joint probability distribution, and $p(a)$ and $p(b)$ are the marginal probability distribution. In feature selection, we can use MI in a *supervised* way to measure the relevancy between a feature $\mathcal{Y}(p_1, p_2, p_3)$ and the class label c. A higher MI indicates a greater dependency or relevancy between them.

Table 1. Details of the four classification tasks for experimental evaluation

#Class	#Sample	Semantic categories (classes)
2	120	Animals (animal+insect) / tools (tool+furniture) [7]
4	120	Animal/insect/tool/vegetable [7]
6	180	Animal/insect/tool/vegetable/building/vehicle
8	240	Animal/insect/tool/vegetable/building/vehicle/buildingpart/clothing

Mapping for Interpretability. It is often useful to localize regions in the original voxel space of the brain for interpretation. Good features for classification are expected to be closely related to discriminating regions. Since $y_p = \langle \mathcal{X}, \mathcal{U}_p \rangle$, we can view y_p as a weighted summation of the voxels in $\mathcal{X}$, where the weights are contained in $\mathcal{U}_p$. Therefore, we propose a scheme to map the selected MPCA features (the eigentensors) to the voxel space. We perform a weighted aggregation of the selected eigentensors first and then determine the D most informative voxels to produce a spatial map $\mathcal{M}$ by choosing an appropriate threshold T (depending on D): $\mathcal{M} = \sum_{p=1}^{P} w_p \, |\mathcal{U}_p| > T$, where w_p is the weight for the pth eigentensor, and $|\cdot|$ denotes the absolute value (magnitude). Note that $\mathcal{M}$ is actually a low-rank tensor (rank P) since it is a summation of P rank-one tensors $\{\mathcal{U}_p\}$ [8].

3 Experiments and Discussions

Data. We choose a challenging multiclass dataset, the CMU Science 2008 fMRI data (CMU2008) [11]. It aims to predict brain activities associated with the meanings of nouns. The data acquisition experiments had 9 subjects viewing 60 different word-picture stimuli from 12 semantic categories, with 5 exemplars per category and 6 runs per stimulus. The acquisition matrix was 51×61 with 23 slices, with the numbers of brain voxels for 9 subjects ranging from 19,750 to 21,764. The mean PSC values over time are extracted as input fMRI features.[1]

Multiclass Tasks. We study four classification tasks: the binary (2-class) and 4-class tasks with the same settings as [7], and two additional, more challenging, 6-class and 8-class tasks. Table 1 summarizes the details.

Algorithms.[2] We evaluate seven feature selection/extraction algorithms in Table 2 on the four tasks above: the MI-based univariate feature selection (MI) [3], variance-based univariate feature selection (Var), LR with the elastic net penalty (LR+ENet) for multivariate feature selection [5,7]; PCA-based feature extraction followed by MI-based and variance-based feature selection (PCA-MI and PCA-Var), and the proposed methods of MPCA with MI-based and variance-based feature selection (MPCA-MI and MPCA-Var).

[1] `http://www.cs.cmu.edu/afs/cs/project/theo-73/www/science2008/data.html`

[2] We have also tested MPCA alone and MPCA with Lasso-based feature selection. They have similar accuracy as MPCA-MI/Var, but using much more features. In addition, replacing PCA with its popular extension, kernel PCA, gives only slightly better results than PCA-MI/Var.

Table 2. The classification accuracy in percentage (Acc) and average numbers of features selected (#Features) by five competing methods and the proposed two methods for the four tasks. The top two results in accuracy are highlighted in bold font.

Task	2-Class		4-Class		6-Class		8-Class	
Method	Acc	#Features	Acc	#Features	Acc	#Features	Acc	#Features
MI	72.78	2276	42.59	3472	37.65	4157	33.94	4208
Var	73.43	5120	39.26	5795	35.31	5405	33.52	5662
LR+ENet	72.13	629	42.78	3058	**40.19**	1441	34.58	3136
PCA-MI	71.30	65	36.85	59	34.14	78	30.19	102
PCA-Var	71.39	58	37.31	56	34.20	78	31.30	98
MPCA-MI	**75.83**	701	**44.35**	995	**40.86**	968	**36.06**	1088
MPCA-Var	**77.41**	910	**44.81**	973	40.06	1165	**34.68**	1034

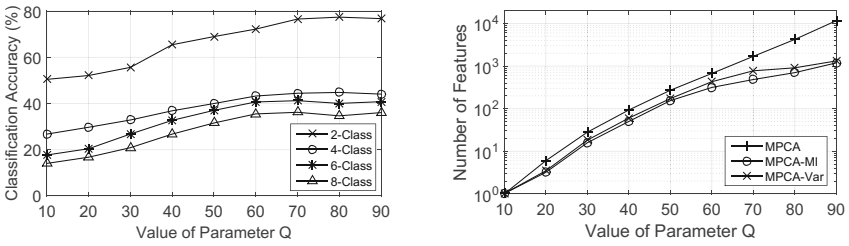
Experimental Settings. We follow [7] to arrange testing, validation, and training sets in the format of $(1:1:4)$ for the six runs in all the experiments. Following [3], we use the SVM classifier with the linear kernel to classify selected features for all methods except LR+ENet which serves as a classifier itself [7,17]. We use the average classification accuracy as the evaluation metric.

Algorithm Settings. Parameters for LR+ENet are set following [7] and the number of selected features is determined according to the weight matrix in LR [17]. Other methods use the validation set to determine the number of selected features with the same steps as in [3]. For our MPCA-MI and MPCA-Var methods, we set the parameter $Q = 80$ in MPCA to report the results. Empirical studies to be shown in Fig. 2(a) show that the classification performance is not sensitive to Q as long as it is not too small (e.g., for $Q \geq 70$).

Classification Accuracy. As shown in Table 2, our proposed methods, MPCA-MI and MPCA-Var achieve the top two overall accuracy (highlighted in bold font), with MPCA-Var achieving the best results on 2-class and 4-class tasks and MPCA-MI achieving the best results on 6-class and 8-class tasks. MPCA-Var and MPCA-MI outperform the state-of-the-art (LR+ENet) by an average of 1.82% and 1.86%, respectively. Though inferior to our methods, LR+ENet indeed outperforms the other existing methods on the whole. In particular, LR+ENet achieves the second best result on 6-class task, slightly better than MPCA-Var. PCA gives the worst results.

Number of Selected Features. The number of features selected varies for different methods in Table 2. It fluctuates drastically (from 629 to 3,136) for LR+ENet. It is stable for the Var method but exceeds 5,000. It increases monotonically with the class number for the MI method. PCA can extract only $(M-1)$ features, e.g., $M = 150$ for the 6-class task. MPCA-Var and MPCA-MI use fewer features in general than MI, Var, and LR+ENet, and the number is relatively stable in contrast.

Parameter Sensitivity. Figure 2(a) plots the average accuracy against the Q values for each task. The four curves share a similar trend. The accuracy increases monotonically with Q till $Q = 70$ and then remains almost constant for $Q \geq 70$.



(a) (b)

Fig. 2. Sensitivity against Q: (a) average accuracy of MPCA-Var (MPCA-MI has similar trends), and (b) average number of extracted (for MPCA) and selected (for MPCA-MI/MPCA-Var) features on the 2-class task (other tasks share similar trends).

(a) Subject 1 (b) Subject 5

Fig. 3. Discriminating regions localized by MPCA-MI for a 2-class task

Thus we choose $Q = 80$ to report the results. In addition, as shown in Fig. 2(b), the Q value has a greater effect on the number of features, which affects the efficiency in turn. The number of features extracted by MPCA increases almost exponentially with Q, while that by MPCA-MI or MPCA-Var increases with Q at a much slower rate for $Q > 60$.

Mapping and Interpretation. Since raw fMRI data are not provided in the CMU2008, we overlay the regions localized by our mapping scheme on a properly scaled and cropped version of the MRI template in "mri.mat" (Matlab R2013b). We set the weight $w_p = 1/p$ to give higher weights to MPCA features with higher importance scores. The regions localized with $D = 2000$ voxels are highlighted in red in Fig. 3 for the 9th-11th slices of two subjects for MPCA-MI on the 2-class task. The localized regions are spatially coherent and largely consistent between different subjects. Moreover, the localized discriminating regions of Subject 5 have significant overlap with the interpretable regions of the same subject depicted in Fig. 3(B) of [11], indicating good interpretability.

4 Conclusion

In this paper, we propose to learn features directly from tensor representations of whole-brain fMRI data via MPCA for classification. We use a variance-based or an MI-based criterion to select the most informative MPCA features for SVM classification. In addition, we propose a novel scheme to localize discriminating regions by mapping the selected MPCA features to the raw voxel space. Experimental results on challenging multiclass tasks show that our methods outperform the state-of-the-art methods. Furthermore, the proposed mapping scheme can

localize discriminating regions that are spatially coherent and consistent cross subjects, with good potential for neuroscience interpretation.

Acknowledgments. We thank the support of Hong Kong Research Grants Council (under Grant 22200014 and the Hong Kong PhD Fellowship Scheme).

References

1. Batmanghelich, N., Dong, A., Taskar, B., Davatzikos, C.: Regularized tensor factorization for multi-modality medical image classification. In: Fichtinger, G., Martel, A., Peters, T. (eds.) MICCAI 2011, Part III. LNCS, vol. 6893, pp. 17–24. Springer, Heidelberg (2011)
2. Chen, M., et al.: Survey of encoding and decoding of visual stimulus via fMRI: An image analysis perspective. Brain Imaging and Behavior 8(1), 7–23 (2014)
3. Chou, C.A., et al.: Voxel selection framework in multi-voxel pattern analysis of fMRI data for prediction of neural response to visual stimuli. IEEE Transactions on Medical Imaging 33(4), 925–934 (2014)
4. Cuingnet, R., Rosso, C., Lehéricy, S., Dormont, D., Benali, H., Samson, Y., Colliot, O.: Spatially regularized SVM for the detection of brain areas associated with stroke outcome. In: Jiang, T., Navab, N., Pluim, J.P.W., Viergever, M.A. (eds.) MICCAI 2010, Part I. LNCS, vol. 6361, pp. 316–323. Springer, Heidelberg (2010)
5. Ecker, C., et al.: Investigating the predictive value of whole-brain structural MR scans in autism: A pattern classification approach. NeuroImage 49(1), 44–56 (2010)
6. Irimia, A., Van Horn, J.D.: Systematic network lesioning reveals the core white matter scaffold of the human brain. Frontiers in Human Neuroscience 8, 1–14 (2014)
7. Kampa, K., Mehta, S., et al.: Sparse optimization in feature selection: application in neuroimaging. Journal of Global Optimization 59(2-3), 439–457 (2014)
8. Lu, H., Plataniotis, K.N., Venetsanopoulos, A.: Multilinear Subspace Learning: Dimensionality Reduction of Multidimensional Data. CRC Press (2013)
9. Lu, H., Plataniotis, K.N., et al.: MPCA: Multilinear principal component analysis of tensor objects. IEEE Trans. Neural Networks 19(1), 18–39 (2008)
10. McKeown, M.J., et al.: Local linear discriminant analysis (LLDA) for group and region of interest (ROI)-based fMRI analysis. NeuroImage 37(3), 855–865 (2007)
11. Mitchell, T.M., Shinkareva, S.V., et al.: Predicting human brain activity associated with the meanings of nouns. Science 320(5880), 1191–1195 (2008)
12. Mourão-Miranda, J., et al.: Classifying brain states and determining the discriminating activation patterns: Support vector machine on functional MRI data. NeuroImage 28(4), 980–995 (2005)
13. Mourão-Miranda, J., et al.: The impact of temporal compression and space selection on SVM analysis of single-subject and multi-subject fMRI data. NeuroImage 33(4), 1055–1065 (2006)
14. Mwangi, B., Tian, T.S., Soares, J.C.: A review of feature reduction techniques in neuroimaging. Neuroinformatics 12(2), 229–244 (2013)
15. Rasmussen, P.M., et al.: Model sparsity and brain pattern interpretation of classification models in neuroimaging. Pattern Recognition 45(6), 2085–2100 (2012)
16. Retico, A., Bosco, P., et al.: Predictive models based on support vector machines: Whole-brain versus regional analysis of structural MRI in the alzheimer's disease. Journal of Neuroimaging, 1–12 (2014)
17. Ryali, S., Supekar, K., Abrams, D.A., Menon, V.: Sparse logistic regression for whole-brain classification of fMRI data. NeuroImage 51(2), 752–764 (2010)