



Supervised, Multivariate, Whole-Brain Reduction Did Not Help to Achieve High Classification Performance in Schizophrenia Research

Eva Janousova^{1*}, Giovanni Montana², Tomas Kasperek^{3,4} and Daniel Schwarz¹

¹ Institute of Biostatistics and Analyses, Faculty of Medicine, Masaryk University, Brno, Czech Republic, ² Department of Biomedical Engineering, King's College London, London, UK, ³ Behavioural and Social Neuroscience Group, CEITEC - Central European Institute of Technology, Masaryk University, Brno, Czech Republic, ⁴ Department of Psychiatry, University Hospital Brno and Masaryk University, Brno, Czech Republic

OPEN ACCESS

Edited by:

Xi-Nian Zuo,
Chinese Academy of Sciences, China

Reviewed by:

Hugo Schnack,
Utrecht University, Netherlands
Zhengjia Dai,
Sun Yat-Sen University, China

*Correspondence:

Eva Janousova
janousova@iba.muni.cz

Specialty section:

This article was submitted to
Brain Imaging Methods,
a section of the journal
Frontiers in Neuroscience

Received: 20 January 2016

Accepted: 10 August 2016

Published: 25 August 2016

Citation:

Janousova E, Montana G, Kasperek T and Schwarz D (2016) Supervised, Multivariate, Whole-Brain Reduction Did Not Help to Achieve High Classification Performance in Schizophrenia Research. *Front. Neurosci.* 10:392. doi: 10.3389/fnins.2016.00392

We examined how penalized linear discriminant analysis with resampling, which is a supervised, multivariate, whole-brain reduction technique, can help schizophrenia diagnostics and research. In an experiment with magnetic resonance brain images of 52 first-episode schizophrenia patients and 52 healthy controls, this method allowed us to select brain areas relevant to schizophrenia, such as the left prefrontal cortex, the anterior cingulum, the right anterior insula, the thalamus, and the hippocampus. Nevertheless, the classification performance based on such reduced data was not significantly better than the classification of data reduced by mass univariate selection using a *t*-test or unsupervised multivariate reduction using principal component analysis. Moreover, we found no important influence of the type of imaging features, namely local deformations or gray matter volumes, and the classification method, specifically linear discriminant analysis or linear support vector machines, on the classification results. However, we ascertained significant effect of a cross-validation setting on classification performance as classification results were overestimated even though the resampling was performed during the selection of brain imaging features. Therefore, it is critically important to perform cross-validation in all steps of the analysis (not only during classification) in case there is no external validation set to avoid optimistically biasing the results of classification studies.

Keywords: computational neuroanatomy, pattern recognition, classification, penalized linear discriminant analysis, support vector machines, cross-validation, magnetic resonance imaging, schizophrenia

INTRODUCTION

Schizophrenia is a disabling chronic psychiatric disease characterized by symptoms such as hallucinations, delusions, thought disturbances, or poverty of speech (Liddle, 1987). It affects young people in their productive years (i.e., typically in their late teens or early twenties; Andreasen, 1995); the lifetime prevalence of schizophrenia is ~0.48% (Simeone et al., 2015). Thus, schizophrenia is one of the leading contributors to the global disease burden. Despite decades of schizophrenia etiology and manifestation research, its diagnostics is still based on clinical interviews that are subjective and sometimes inaccurate. There are accordingly endeavors to employ computer-assisted

methods based on brain imaging data for schizophrenia diagnostics in order to reduce subjectivity and also increase the speed of diagnostic assessments (Klöppel et al., 2012; Wolfers et al., 2015).

Modern neuroimaging techniques have enabled the identification of brain regions in which schizophrenia patients show significant group differences compared with healthy subjects (Wright et al., 2000; Shenton et al., 2001; Niznikiewicz et al., 2003; Honea et al., 2005; Haijma et al., 2013). On individual levels, however, brain-imaging measurements in schizophrenics exhibit a wide overlap with the normal range (Sun et al., 2009), which makes the diagnosis of schizophrenia based on individual brain-image data quite demanding. The first studies that sought to distinguish schizophrenia patients from healthy controls (HC) based on magnetic resonance imaging (MRI) data were based on brain anatomical measures of a few selected regions of interest (Leonard et al., 1999; Nakamura et al., 2004). Later studies attempted to use whole-brain data for schizophrenia classification since the neural mechanisms of schizophrenia are not restricted to a certain region of the brain but are instead distributed (Michael et al., 2008; Melonakos et al., 2011). For example, some studies utilized voxel-based morphometry or other univariate whole-brain methods to select imaging features discriminating strongly between schizophrenia subjects and HC and use them for classification (Kovalev et al., 2003; Diaz et al., 2010). Such feature selection makes it possible to increase the signal-to-noise ratio and also avoid the so-called “small sample size problem” (Lemm et al., 2011) when the number of subjects is considerably smaller than the number of features, which often causes unstable classification performance (Demirci et al., 2008). Selecting imaging features based on feature-by-feature tests does not allow utilizing information about feature correlation, however. As a result, multivariate, whole-brain data reduction techniques are preferred. The most commonly used multivariate method in schizophrenia research is principal component analysis (PCA; Yoon et al., 2007; Karageorgiou et al., 2011). Nevertheless, PCA also has a few disadvantages: It is sometimes ineffective at capturing complex relationships in high-dimensional spaces since it is a linear method (Fan et al., 2005), and it is not guaranteed that the selected principal components corresponding to the largest eigenvalues are those discriminating the best between patients and HC since PCA does not take in account information about the patient and control group labels (Bunea et al., 2011; Janousova et al., 2015). To overcome these shortcomings, recent studies have used deep learning methods (Plis et al., 2014), normative modeling (Marquand et al., 2016), or supervised, multivariate, whole-brain reduction techniques that utilize group labels during data reduction, such as sparse multinomial logistic regression, a ν -multiple kernel learning approach and sparse network-based models used for MRI or functional MRI classification in schizophrenia research (Sun et al., 2009; Castro et al., 2014; Rosa et al., 2015). In other fields of study, supervised, multivariate, whole-brain reduction methods such as penalized least squares regression or sparse logistic regression have been used to classify HIV-infected individuals and music vs. speech stimuli in healthy subjects, respectively (Ryali et al., 2010; Bunea

et al., 2011). In our previous imaging-genetics study of patients with Alzheimer’s disease, we successfully used penalized linear discriminant analysis (pLDA) with resampling to pre-select image phenotypes (Vounou et al., 2012).

Here, we use pLDA with resampling to select the most discriminative features from MRI data to differentiate first-episode schizophrenia (FES) patients from HC. We investigate if this supervised, multivariate, whole-brain reduction method is capable of efficient data reduction that results in significantly better classification performance than mass univariate selection of features using a t -test or unsupervised, multivariate reduction using PCA. Moreover, we also examine the influence of the type of features extracted from the MRI data, namely gray matter (GM) volumes and local deformations, and the type of classifier on the classification performance. To enable comparison with other studies, we use two of the most commonly applied classifiers in schizophrenia research: linear discriminant analysis (Leonard et al., 1999; Nakamura et al., 2004; Takayanagi et al., 2010; Karageorgiou et al., 2011; Ota et al., 2012) and linear support vector machines (SVM; Yoon et al., 2007; Pohl and Sabuncu, 2009; Mourao-Miranda et al., 2012; Nieuwenhuis et al., 2012; Dluhos et al., 2014). Furthermore, we explore the effect of leave-one-out cross-validation (LOOCV) performed during classification solely or during data reduction and classification. The algorithm employs resampling during data reduction and therefore the brain imaging features are not selected from all of the input imaging data but rather from a number of random data subsets. Such an approach might enable to avoid optimistically biased classification results that occur in cases in which cross-validation is executed only during data classification.

The remainder of the paper is organized as follows. In Section Materials and Methods, we describe a data set of FES patients and HC used in our study, the preprocessing of their MRI data leading to the extraction of two sets of imaging features, the reduction of the feature sets using pLDA with resampling, and the subsequent classification using two classifiers including cross-validation. Section Results is dedicated to a summary of our classification results. Section Discussion discusses our results, compares them with classification performance based on commonly used data reduction techniques, and concludes the paper.

MATERIALS AND METHODS

Subjects

Our patient group included 52 males (median age: 23 years, range: 17–40 years) hospitalized in the all-male unit of the Department of Psychiatry, Masaryk University in Brno, Czech Republic; this cohort constituted the FES group. The diagnosis was established during a clinical interview performed by a trained psychiatrist. The interview was conducted in compliance with the International Statistical Classification of Disease and Related Health Problems (ICD-10) research criteria and was focused on information about family and personal history, somatic conditions, pharmacological history and current treatment, substance abuse, previous psychiatric conditions, the current clinical manifestation, symptoms and their duration, and functional impact. Where possible,

information from relatives was collected as well. After the interview, the patients were physically examined, inclusive of a urine analysis (biochemistry, toxicology) and laboratory blood tests (hematology, biochemistry). Abnormal findings were subjected to additional tests and examinations. A fully trained senior psychiatrist reviewed all information, established the diagnosis and suggested the case for inclusion in the study. The exclusion criteria included substance dependence (detected via urine toxicology tests and clinical evaluation), neurological, or systemic disease with known relationships to brain alteration (detected by clinical evaluation, neurological and physical examinations, serum and urine chemistry, and blood count, serological examination for neurotropic agents and clinical evaluation of MRI scans), and contraindications for MRI examination.

The control group consisted of 52 HC matched for age (median age: 23 years, range: 18–38 years) and sex (all males) who were recruited from the community, medical staff, and students. Exclusion criteria included substance dependence, a personal or family history of mental illness, somatic conditions affecting the function or structure of the brain as assessed during clinical interviews performed by a trained psychiatrist, and contraindications for MRI.

The study was approved by the local ethics committee, and all subjects signed an informed consent form prior to their participation in the study.

MRI Data Acquisition and Preprocessing

All 104 subjects were scanned with a 1.5T Siemens Symphony machine. We acquired whole-head T1-weighted images using three-dimensional acquisition with IR/GR sequence, TR 1700 ms, TE 3.93 ms, TI 1100 ms, flip angle 15° . The sagittal tomographic plane thickness was 1.17 mm, the in-plane resolution was 0.48×0.48 mm, the 3-D field of view contained $160 \times 512 \times 512$ voxels. We performed the MRI examination during the first episode, which means that the duration of treatment with antipsychotics was only from 3 to 14 weeks at the time of the MRI.

Prior to the analysis, we checked all of the images for morphological abnormalities. Since no abnormalities were found, we preprocessed all 104 images using Statistical Parametric Mapping 8 (SPM8: <http://www.fil.ion.ucl.ac.uk/spm/>). Specifically, the images were corrected for bias-field inhomogeneity, spatially normalized (i.e., transformed into stereotactic space), resampled to $1.5 \times 1.5 \times 1.5$ mm, and segmented into GM, white matter and cerebrospinal fluid. The spatially normalized images were oriented axially and consisted of $170 \times 256 \times 256$ voxels at 1.5 mm isotropic resolution. We modulated the GM tissue segment with the determinant of Jacobian matrices of the deformations to account for registration-related changes in local volumes, and we smoothed the data using an 8-mm full width at half maximum Gaussian kernel according to the optimized voxel-based morphometry (oVBM) procedure (Good et al., 2001). The GM volumes that we obtained were then used as the first type of features in the analysis.

The second type of features included local deformations that revealed how the brain anatomy of a diagnosed subject differs from the normal template anatomy in terms of

local volume expansions and contractions. These data were acquired using deformation-based morphometry (DBM). The spatial normalization steps for DBM included the same affine registration algorithm as for oVBM apart from the resolution ($1 \times 1 \times 1$ mm). After transforming all bias-corrected images into stereotactic space, we used our original high-dimensional nonlinear registration technique (Schwarz et al., 2007) to compute vector displacement fields that maximized the normalized mutual information between the images and the high-resolution digital brain atlas. The resulting three-dimensional displacement vector fields were converted into scalar fields (referred to as “local deformations”) by computing the logarithms of their respective determinants of Jacobian matrices at each voxel in stereotactic space.

The final preprocessing step removed features corresponding to non-brain tissues and image background from the two feature sets. Each feature set from every subject was then transformed into one-dimensional vectors and arranged in an $(n \times p)$ matrix \mathbf{X} , where n is the number of individuals in a data set and p is the number of features. The feature matrix of the GM volumes had a size of $(104 \times 748, 931)$; the dimensions of the feature matrix of the local deformations were $(104 \times 1, 924, 670)$.

Imaging Feature Selection Using pLDA with Resampling

Since the number of features in the data set was very high even after the removal of non-brain and background features, both feature sets were reduced prior to classification to avoid the so-called “small sample size problem,” to increase the signal-to-noise ratio and possibly prevent for unstable classification results with poor generalizability (Demirci et al., 2008; Lemm et al., 2011). We used pLDA with resampling to identify the most discriminative features. The pLDA is described in detail elsewhere (Vounou et al., 2012) and is briefly summarized below.

In pLDA, a lasso penalty is imposed on the l_1 norm of the direction vector \mathbf{v} (Witten and Tibshirani, 2011) that best discriminates two classes within a data sample by maximizing the between-class variance and simultaneously minimizing the within-class variance. The direction vector \mathbf{v} is a solution of the following optimization problem in the case of the common Fisher’s LDA:

$$\max_{\mathbf{v}} \left\{ \mathbf{v}^T \mathbf{S}_B \mathbf{v} \right\} \text{ subject to } \mathbf{v}^T \mathbf{S}_W \mathbf{v} = 1, \quad (1)$$

where \mathbf{S}_B and \mathbf{S}_W are the between-class scatter matrix and the within-class scatter matrix, respectively, which can be calculated as:

$$\begin{aligned} \mathbf{S}_B &= (\mathbf{m}_H - \mathbf{m}_D)^T (\mathbf{m}_H - \mathbf{m}_D), \\ \mathbf{S}_W &= \sum_{i=1}^{n_H} (\mathbf{x}_i - \mathbf{m}_H)^T (\mathbf{x}_i - \mathbf{m}_H) \\ &\quad + \sum_{i=1}^{n_D} (\mathbf{x}_i - \mathbf{m}_D)^T (\mathbf{x}_i - \mathbf{m}_D), \end{aligned} \quad (2)$$

where $\mathbf{m}_H = 1/n_H \cdot \sum_{i=1}^{n_H} \mathbf{x}_i$ is the mean vector of class H (healthy controls), $\mathbf{m}_D = 1/n_D \cdot \sum_{i=1}^{n_D} \mathbf{x}_i$ is the mean vector of class D (diseased individuals), n_H and n_D are the numbers of

subjects in classes H and D , respectively, and $\mathbf{x}_i, i = 1, \dots, n$, are the rows of the feature matrix \mathbf{X} .

In pLDA, imposing the lasso penalty leads to setting the coefficients $v_j, j = 1, \dots, p$, of the least discriminative features to zero. Therefore, the optimization problem, which is solved by the minorization-maximization algorithm, changes to:

$$\max_{\mathbf{v}} \left\{ \mathbf{v}^T \mathbf{S}_B \mathbf{v} - \lambda \sum_{j=1}^p s_j |v_j| \right\} \text{ subject to } \mathbf{v}^T \mathbf{S}_W^* \mathbf{v} = 1, \quad (3)$$

where \mathbf{S}_W^* is the diagonal estimate of \mathbf{S}_W , $\text{diag}(\mathbf{S}_W^*) = (s_1^2, \dots, s_p^2)$, to avoid problems with possible singularity of \mathbf{S}_W , and λ is a regularization parameter that controls the degree of sparsity in the model. Specifically, when λ is exactly zero, no penalty is imposed and all p features contribute in the direction vector \mathbf{v} . As λ increases above zero, fewer features contribute in \mathbf{v} . At its maximum value, all coefficients of \mathbf{v} are set to zero.

For the particular data, the optimal value of λ is not known a priori. Tuning λ commonly involves cross-validating the prediction error for a grid of values of λ and selecting the value of λ that leads to the smallest cross-validated error. However, this approach may be prone to sampling errors. Therefore, we combined the pLDA with a resampling method proposed in Meinshausen and Bühlmann (2010) for sparse predictive modeling. This procedure aims to calculate the selection probabilities $P_j(\lambda)$ for each feature by repeatedly fitting the pLDA model on random subsets of the data set while keeping track of the features associated with the non-zero coefficients of \mathbf{v} . Specifically, an indicator variable $\mathbf{c}^{(k)}(\lambda) = (c_1^{(k)}(\lambda), \dots, c_p^{(k)}(\lambda))$ is calculated in each iteration $k = 1, \dots, K$, where $c_j^{(k)}(\lambda), j = 1, \dots, p$, is equal to 1 if the j th feature is selected (i.e., its respective coefficient v_j is non-zero) and 0 otherwise. Using all K random subsets, the selection probability for each feature can then be calculated as:

$$P_j(\lambda) = \frac{1}{K} \sum_{k=1}^K c_j^{(k)}(\lambda), j = 1, \dots, p. \quad (4)$$

The final set of the most discriminative features consists of those features with selection probabilities above a predefined threshold, e.g., 0.99.

To compare the results with common data reduction techniques, we also performed feature extraction using PCA (all $n - 1$ principal components corresponding to non-zero eigenvalues were used for classification) and feature selection using the Student's t -test with three different thresholds ($p < 0.01$, $p < 0.005$, and $p < 0.001$).

Imaging Feature Classification

The final set of features selected using pLDA with resampling was then used for classifying individuals into the D or H class using two most commonly employed classifiers (i.e., LDA and linear SVM).

Since there was no external testing data set available, we used the LOOCV procedure to avoid overestimating the classification

results. This procedure contains a loop in which one subject is chosen as a testing one and the remaining $n - 1$ subjects are employed for training the classifier. The testing subject is then classified as belonging to the patient or HC class, and the resulting class label is compared with the true classification label. This procedure is repeated n times using all subjects as the testing subject sequentially. Afterwards, the overall classification performance measures of accuracy, sensitivity, and specificity are calculated, which are defined as:

$$\text{accuracy} = \frac{TP + TN}{TP + TN + FP + FN}, \quad (5)$$

$$\text{sensitivity} = \frac{TP}{TP + FN}, \quad (6)$$

$$\text{specificity} = \frac{TN}{TN + FP}, \quad (7)$$

where TP and TN refer to numbers of true positive and true negative results, respectively, and FP and FN denote the numbers of false positive and false negative results, respectively.

It is well-known that data classification as well as data reduction should not use all of the input data simultaneously to avoid overoptimistic classification performance (Guyon and Elisseeff, 2003). Since resampling is performed during data reduction here, and therefore the brain imaging features are selected based on random data subsets, it appears sufficient to perform LOOCV only for feature classification [denoted by LOOCV^(c)]. To verify this hypothesis, we also performed analysis when we repeatedly split the data into training and testing sets prior to data reduction [marked as LOOCV^(rc) and illustrated in **Figure 1**]. In that case, only the data of $n - 1$ subjects were input into the pLDA with resampling in each LOOCV^(rc) iteration, and the data of the testing subject were only used during classification.

All algorithms described in the Materials and Methods Section were implemented in “Penalised Reduction & Classification Toolbox” for MATLAB (the toolbox and documentation are freely available at <http://www.iba.muni.cz/index-en.php?pg=research--data-analysis-tools--plda>).

RESULTS

The classification results based on local deformations and GM volumes reduced using pLDA with resampling based on 100 random data subsets and various values of λ , PCA, t -test with three thresholds, and the imaging data without reduction are given in **Tables 1, 2**. In the case of local deformations and GM volumes, the parameter λ ranged from 0.6 to 2.0 and from 0.2 to 1.2, respectively. The pLDA algorithm did not converge for $\lambda > 2$ and $\lambda > 1.2$, respectively; whereas for $\lambda < 0.6$ and $\lambda < 0.2$, respectively, it did not reduce the data to less than hundreds of thousands of features. In the case of t -test, three thresholds $p < 0.01$, $p < 0.005$, and $p < 0.001$ were used. When we applied a false discovery rate correction, no local deformations remained statistically significant, and only 2000 GM volumes remained significant. On the other hand, a threshold of $p < 0.05$ resulted

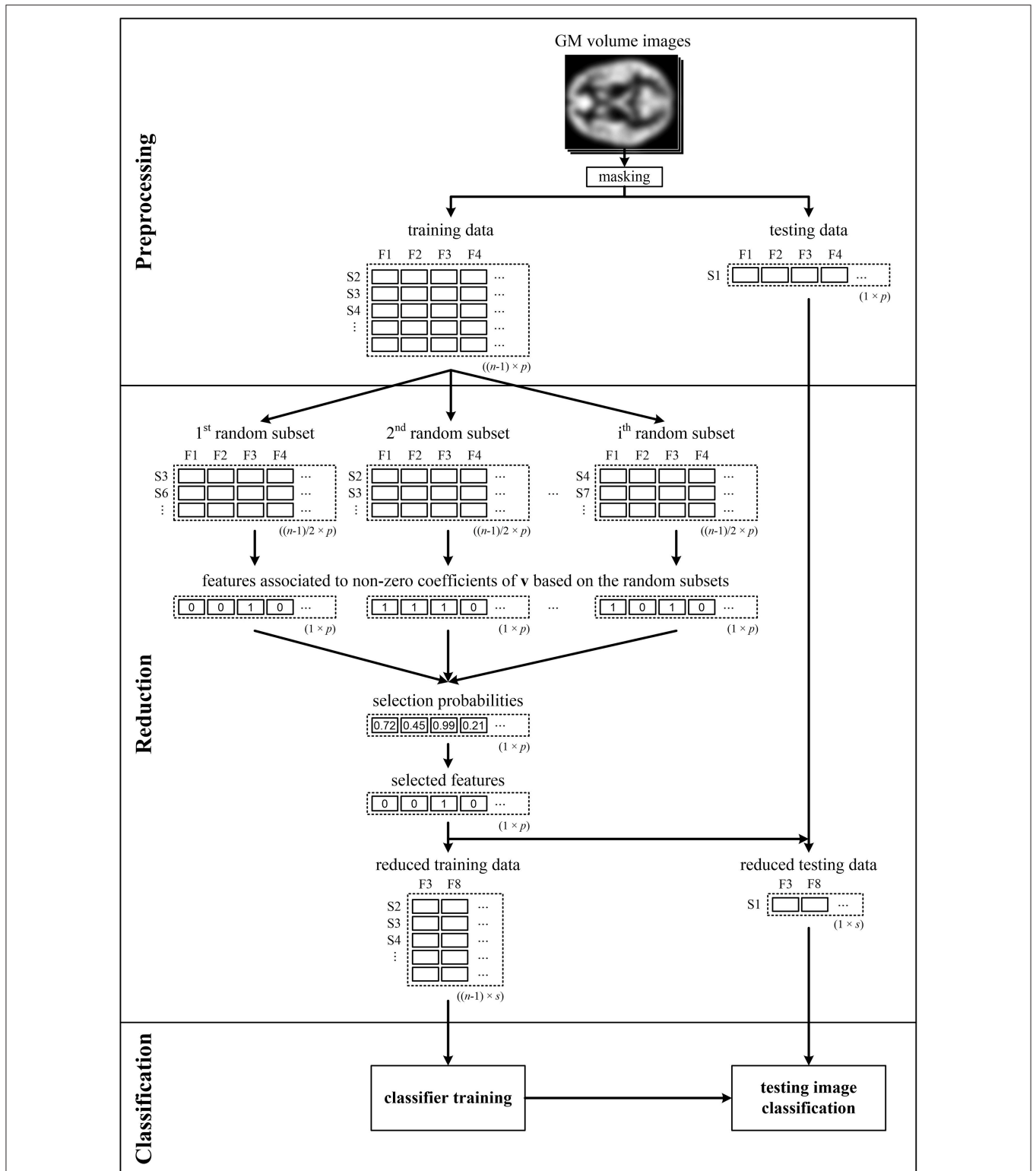


FIGURE 1 | Illustrative scheme of recognition of patients from controls comprising LOOCV during data reduction and classification—LOOCV^(rc). Brain images (e.g., GM volume images), from which the background and non-brain areas are removed using a binary mask, are split into training and testing sets. The training data are used to calculate the selection probabilities of features using pLDA based on random subsets. The training and testing data are then reduced by selecting only the features with selection probabilities ≥ 0.99 . Next, the reduced training data are used to train the classifier, which makes it possible to classify the testing data into a group of schizophrenia patients or HC. The entire process is repeated n times as each image is sequentially chosen as the testing image.

TABLE 1 | Results of LDA and linear SVM (linSVM) classification experiments based on local deformations reduced by pLDA with resampling using various values of λ , PCA, t-test with three thresholds or unreduced.

Reduction	λ	Classifier	LOOCV during classification—LOOCV(c)				LOOCV during reduction and classification—LOOCV(c)					
			# features	# iterations (mean±SD)	Acc (%)	Sens (%)	Spec (%)	# features (mean±SD)	# iterations (mean±SD)	Acc (%)	Sens (%)	Spec (%)
pLDA	0.6	LDA	98,320	3.0 ± 0.0	84.6	86.5	82.7	104,184.2 ± 8527.3	3.0 ± 0.0	65.4	57.7	73.1
		linSVM			86.5	88.5	84.6			69.2	67.3	71.2
pLDA	0.7	LDA	83,193	3.0 ± 0.0	83.7	84.6	82.7	89,008.5 ± 7796.1	3.0 ± 0.0	65.4	57.7	73.1
		linSVM			86.5	88.5	84.6			70.2	67.3	73.1
pLDA	0.8	LDA	71,245	3.0 ± 0.1	82.7	82.7	82.7	76,965.1 ± 7170.8	3.0 ± 0.0	65.4	57.7	73.1
		linSVM			86.5	88.5	84.6			70.2	65.4	75.0
pLDA	0.9	LDA	61,398	2.9 ± 0.3	82.7	82.7	82.7	66,372.6 ± 5736.2	3.0 ± 0.2	66.3	59.6	73.1
		linSVM			86.5	88.5	84.6			64.4	63.5	65.4
pLDA	1.0	LDA	52,927	2.9 ± 0.3	81.7	80.8	82.7	58,355.1 ± 5250.0	2.9 ± 0.3	67.3	61.5	73.1
		linSVM			85.6	86.5	84.6			68.3	65.4	71.2
pLDA	1.1	LDA	45,557	3.1 ± 0.5	82.7	80.8	84.6	50,510.9 ± 4896.3	3.0 ± 0.4	66.3	59.6	73.1
		linSVM			84.6	82.7	86.5			63.5	61.5	65.4
pLDA	1.2	LDA	39,008	3.2 ± 0.6	83.7	82.7	84.6	44,624.4 ± 5221.4	3.0 ± 0.6	66.3	57.7	75.0
		linSVM			84.6	82.7	86.5			66.3	65.4	67.3
pLDA	1.3	LDA	33,082	3.6 ± 0.7	84.6	82.7	86.5	38,712.8 ± 4816.2	3.3 ± 0.6	65.4	55.8	75.0
		linSVM			84.6	82.7	86.5			67.3	65.4	69.2
pLDA	1.4	LDA	27,807	3.9 ± 0.8	84.6	82.7	86.5	33,401.8 ± 4425.2	3.7 ± 0.6	65.4	57.7	73.1
		linSVM			82.7	80.8	84.6			66.3	65.4	67.3
pLDA	1.5	LDA	23,116	4.2 ± 0.8	85.6	82.7	88.5	28,275.0 ± 3403.0	4.1 ± 0.7	68.3	63.5	73.1
		linSVM			81.7	80.8	82.7			63.5	61.5	65.4
pLDA	1.6	LDA	18,917	4.7 ± 0.9	85.6	82.7	88.5	24,092.3 ± 3688.9	4.4 ± 0.7	67.3	61.5	73.1
		linSVM			81.7	80.8	82.7			68.3	67.3	69.2
pLDA	1.7	LDA	15,203	5.0 ± 1.1	84.6	80.8	88.5	19,575.5 ± 3001.0	4.6 ± 0.9	67.3	61.5	73.1
		linSVM			81.7	80.8	82.7			63.5	63.5	63.5
pLDA	1.8	LDA	11,736	5.6 ± 1.5	85.6	80.8	90.4	16,241.5 ± 2996.4	5.0 ± 1.0	63.5	59.6	67.3
		linSVM			80.8	78.8	82.7			63.5	65.4	61.5
pLDA	1.9	LDA	9555	7.1 ± 9.5	83.7	78.8	88.5	12,483.7 ± 2421.6	5.5 ± 1.2	65.4	61.5	69.2
		linSVM			80.8	80.8	80.8			55.8	57.7	53.8
pLDA	2.0	LDA	7306	8.7 ± 13.2	83.7	78.8	88.5	9,910.8 ± 2196.8	6.2 ± 1.4	62.5	59.6	65.4
		linSVM			80.8	78.8	82.7			59.6	59.6	59.6
PCA	-	LDA	-	-	60.6	55.8	65.4	-	-	66.3	65.4	67.3
		linSVM			67.3	61.5	73.1			67.3	65.4	69.2
t-test, $p < 0.01$	-	LDA	62,644	-	82.7	82.7	82.7	62,110.1 ± 3241.1	-	66.3	59.6	73.1
		linSVM			84.6	84.6	84.6			70.2	69.2	71.2
t-test, $p < 0.005$	-	LDA	38,630	-	84.6	82.7	86.5	38,184.7 ± 2389.9	-	65.4	57.7	73.1

(Continued)

TABLE 1 | Continued

Reduction	λ	Classifier	LOOCV during classification—LOOCV ^(c)				LOOCV during reduction and classification—LOOCV ^(rc)					
			# features	# iterations (mean±SD)	Acc (%)	Sens (%)	Spec (%)	# features (mean±SD)	# iterations (mean±SD)	Acc (%)	Sens (%)	Spec (%)
t-test, $p < 0.001$	-	linSVM			84.6	82.7	86.5			66.3	65.4	67.3
		LDA	10,860	-	83.7	78.8	88.5	10,671.2 ± 1110.8	-	60.6	57.7	63.5
		linSVM			78.8	76.9	80.8			59.6	63.5	55.8
no	-	LDA	1,924,670	-	67.3	63.5	71.2	-	-	-	-	-
		linSVM			64.4	63.5	65.4			-	-	-

In columns: the number of selected features, the number of iterations of the pLDA algorithm and cross-validated classification performance measures in percentage—accuracy (Acc), sensitivity (Sens), and specificity (Spec).

in too many features (more than 200,000 local deformations and 110,000 GM volumes).

The **Tables 1, 2** show that increasing λ led to a decreased number of selected discriminating features and an increased number of iterations of the minorization-maximization algorithm within pLDA. The classification performance differed considerably according to type of LOOCV that we performed. The median accuracy of the classification of local deformations using LDA was 83.7% (range: 81.7–85.6%) in the case of LOOCV^(c) compared with 65.4% (range: 62.5–68.3%) in the case of LOOCV^(rc). The linear SVM classification of local deformations led to similar median accuracies of 84.6% (range: 80.8–86.5%) and 66.3% (range: 55.8–70.2%), respectively. The classification of GM volumes based on LOOCV^(c) led to slightly less overestimated results compared with local deformations; the median accuracy of classification was 76.9% (range: 74.0–83.7%) and 77.9% (range: 75.0–80.8%) using LDA and linear SVM, respectively, compared with LOOCV^(rc) with median accuracies of 66.3% (range: 65.4–68.3%) and 64.4% (range: 58.7–67.3%), respectively. For better illustration, classification accuracy achieved in the experiments vs. the values of the parameter λ was visualized (**Figure 2**).

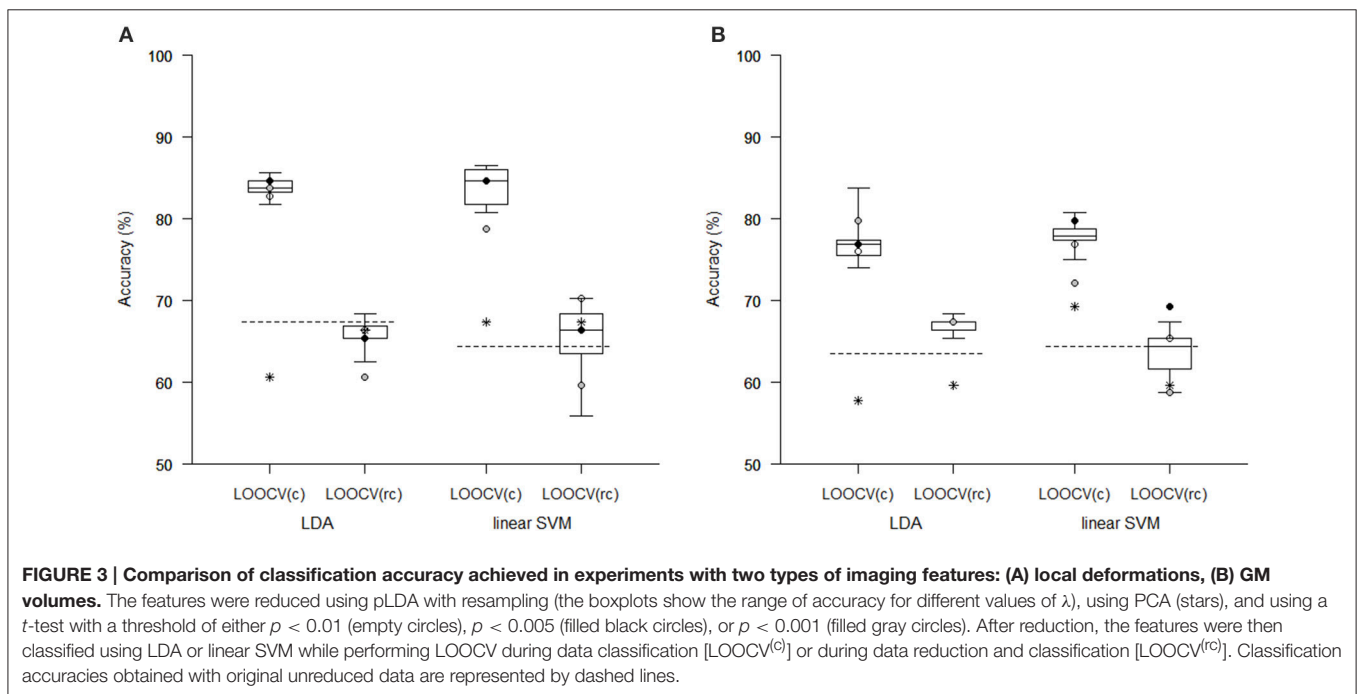
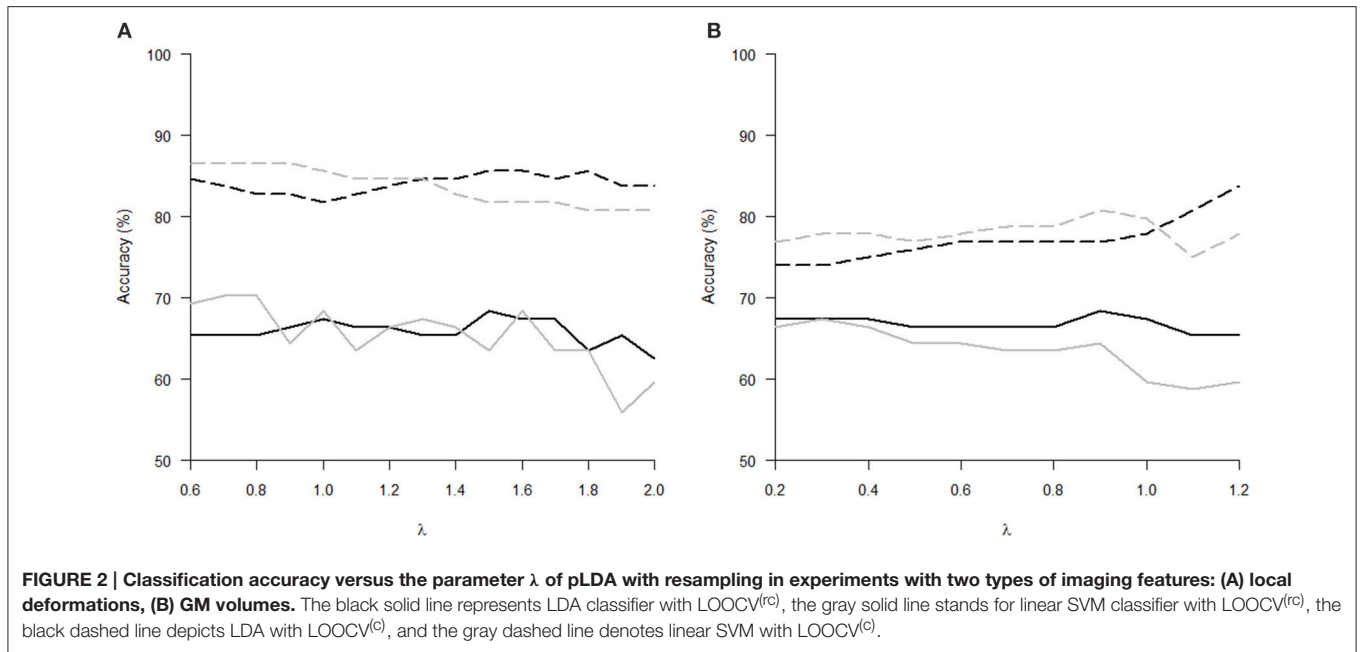
Comparison of the reduction methods showed that the maximum accuracy achieved in the classification of data reduced using pLDA with resampling was higher than the accuracy of the classification of data reduced by PCA or a t -test in six out of eight of the classification experiments (**Tables 1, 2, Figure 3**). However, the difference in accuracy was quite small and statistically insignificant (based on McNemar's test) apart from the cases of reduction using PCA in the case of LOOCV^(c), classification of GM volumes reduced by a t -test with $p < 0.01$ and $p < 0.005$ using LDA with LOOCV^(c); and t -test with $p < 0.001$ for the second, third, fourth, and seventh experiment (the order corresponds to the **Figure 3**) showing that too strict threshold of t -test leading to low number of selected voxels results in low classification performance. In the case of linear SVM with LOOCV^(rc), the classification of data reduced by pLDA yielded a lower accuracy than the classification of GM volumes reduced with a t -test with $p < 0.005$ and same accuracy as the classification of local deformations reduced with t -test with $p < 0.01$. When comparison with unreduced data was performed, the maximum accuracy achieved in the classification of data reduced using pLDA with resampling was higher than the accuracy obtained in the classification of unreduced data in all eight experiments. However, the differences were statistically significant only in the case of LOOCV^(c). Finally, comparison of the achieved classification accuracies with classification by chance was also performed (using one-sample binomial test), showing that classification results obtained by PCA, t -test and maximum classification performance achieved by pLDA was statistically significantly better than chance in all experiments except for PCA on GM volumes with LOOCV^(c) and t -test with $p < 0.001$ on GM volumes with LOOCV^(rc).

The aim of the pLDA algorithm is to identify the features that best discriminate between schizophrenia patients and HC. **Figure 4** shows a visualization of the selected strongly discriminating local deformations and GM volume features. It

TABLE 2 | Results of LDA and linear SVM (linSVM) classification experiments based on GM volumes reduced by pLDA with resampling using various values of λ , PCA, t-test with three thresholds or unreduced.

Reduction	λ	Classifier	LOOCV during classification—LOOCV(c)				LOOCV during reduction and classification—LOOCV(c)					
			# features	# iterations (mean±SD)	Acc (%)	Sens (%)	Spec (%)	# features (mean±SD)	# iterations (mean±SD)	Acc (%)	Sens (%)	Spec (%)
pLDA	0.2	LDA	104,565	3.0 ± 0.0	74.0	73.1	75.0	110,082.0 ± 9659.9	3.0 ± 0.0	67.3	65.4	69.2
		linSVM			76.9	76.9	76.9			66.3	63.5	69.2
pLDA	0.3	LDA	75,620	3.1 ± 0.3	74.0	73.1	75.0	80,612.8 ± 7792.0	3.1 ± 0.2	67.3	65.4	69.2
		linSVM			77.9	76.9	78.8			67.3	65.4	69.2
pLDA	0.4	LDA	58,840	3.5 ± 0.5	75.0	73.1	76.9	63,736.4 ± 6748.8	3.3 ± 0.5	67.3	65.4	69.2
		linSVM			77.9	75.0	80.8			66.3	65.4	67.3
pLDA	0.5	LDA	47,535	3.8 ± 0.4	76.0	75.0	76.9	51,974.5 ± 5957.1	3.7 ± 0.5	66.3	65.4	67.3
		linSVM			76.9	73.1	80.8			64.4	63.5	65.4
pLDA	0.6	LDA	38,756	3.9 ± 0.4	76.9	76.9	76.9	42,871.9 ± 5295.7	3.8 ± 0.4	66.3	65.4	67.3
		linSVM			77.9	76.9	78.8			64.4	61.5	67.3
pLDA	0.7	LDA	31,534	4.1 ± 0.4	76.9	76.9	76.9	35,341.3 ± 4742.8	4.0 ± 0.4	66.3	65.4	67.3
		linSVM			78.8	76.9	80.8			63.5	59.6	67.3
pLDA	0.8	LDA	25,253	4.3 ± 0.5	76.9	76.9	76.9	28,857.4 ± 4245.3	4.2 ± 0.4	66.3	65.4	67.3
		linSVM			78.8	76.9	80.8			63.5	57.7	69.2
pLDA	0.9	LDA	19,670	4.6 ± 0.6	76.9	76.9	76.9	22,762.5 ± 4087.9	4.4 ± 0.6	68.3	67.3	69.2
		linSVM			80.8	80.8	80.8			64.4	63.5	65.4
pLDA	1.0	LDA	14,923	4.7 ± 0.6	77.9	78.8	76.9	18,001.1 ± 3351.3	4.6 ± 0.6	67.3	65.4	69.2
		linSVM			79.8	80.8	78.8			59.6	55.8	63.5
pLDA	1.1	LDA	10,596	4.9 ± 0.6	80.8	80.8	80.8	13,417.8 ± 2828.5	4.8 ± 0.7	65.4	63.5	67.3
		linSVM			75.0	73.1	76.9			58.7	55.8	61.5
pLDA	1.2	LDA	6543	5.4 ± 1.7	83.7	84.6	82.7	9,624.9 ± 2371.7	5.0 ± 0.9	65.4	61.5	69.2
		linSVM			77.9	75.0	80.8			59.6	53.8	65.4
PCA	-	LDA	-	-	57.7	55.8	59.6	-	-	59.6	57.7	61.5
		linSVM			69.2	69.2	69.2			59.6	57.7	61.5
t-test, $p < 0.01$	-	LDA	44,168	-	76.0	75.0	76.9	43,716.8 ± 3122.2	-	67.3	67.3	67.3
		linSVM			76.9	73.1	80.8			65.4	65.4	65.4
t-test, $p < 0.005$	-	LDA	28,638	-	76.9	76.9	76.9	28,252.1 ± 2331.2	-	67.3	67.3	67.3
		linSVM			79.8	78.8	80.8			69.2	65.4	73.1
t-test, $p < 0.001$	-	LDA	9470	-	79.8	80.8	78.8	9,316.1 ± 1134.6	-	67.3	63.5	71.2
		linSVM			72.1	71.2	73.1			58.7	57.7	59.6
no	-	LDA	748,931	-	63.5	63.5	63.5	-	-	-	-	-
		linSVM			64.4	61.5	67.3			-	-	-

In columns: the number of selected features, the number of iterations of the pLDA algorithm and cross-validated classification performance measures in percentage—accuracy (Acc), sensitivity (Sens), and specificity (Spec).



must be noted that in the case of LOOCV^(rc), slightly different sets of selected features were obtained in each LOOCV iteration. Thus, the color coding in **Figure 4** corresponds to the number of LOOCV iterations in which the feature was among the most discriminative features. In the case of local deformations, the discriminative features form connected regions in the left prefrontal cortex, the right anterior insula, the medial parts of the thalamus, and the cerebellar cortex. The discriminative GM volume features lie also in these brain regions as well

as in the inferior frontal gyrus, the anterior cingulum, the hippocampus, and parahippocampal gyrus, the caudate nucleus, and the superior and middle temporal gyrus. Similar brain areas were identified as the most discriminative ones also using *t*-test. In the case of PCA, the features whose loadings contributed most to the first principal component (based on a threshold at 30% of maximum absolute value of the loadings) covered almost entire brain or gray matter areas, respectively. To further investigate discriminative capability of features selected

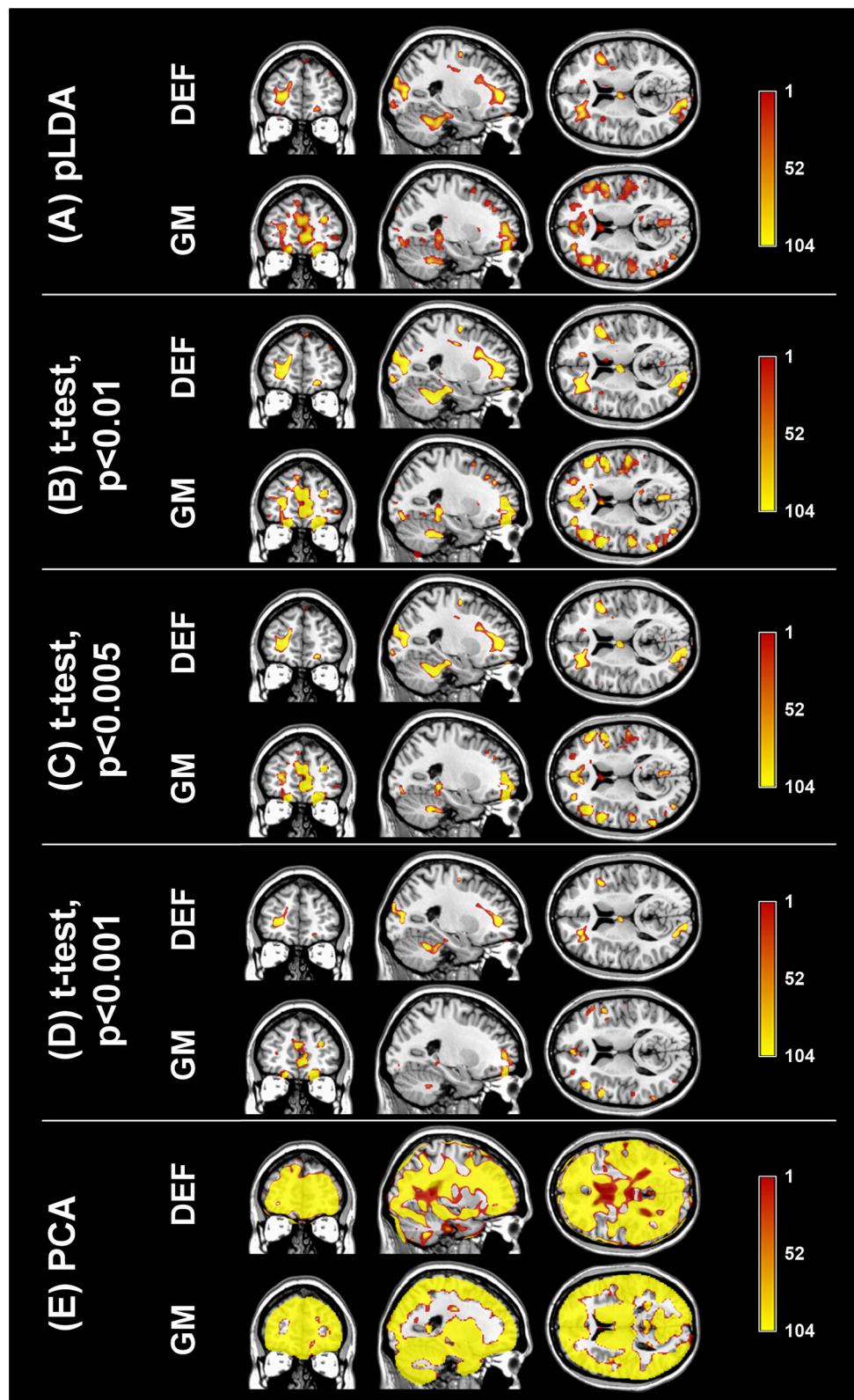


FIGURE 4 | Coronal, sagittal, and transversal slices showing the automatically detected strongly discriminative features in local deformations (DEF) and gray matter volumes (GM) using: (A) pLDA with resampling (for $\lambda = 1.6$ and $\lambda = 0.9$, respectively), (B) *t*-test with $p < 0.01$, (C) *t*-test with $p < 0.005$, (D) *t*-test with $p < 0.001$, (E) PCA. The color coding corresponds to how many times each feature was selected as one of the discriminative features in the case of pLDA with resampling and *t*-test or contributed most to the first principal component in the case of PCA within all 104 LOOCV^(rc) iterations.

by pLDA with resampling in all as well as in more than half of LOOCV^(rc) folds, mean discriminative weights (specifically, mean of non-zero weights over 100 resampling iterations and 104 LOOCV^(rc) folds) were calculated and visualized (Figure 5).

DISCUSSION

In our experiment with T1-weighted MRI data of 52 FES patients and 52 HC, we found that pLDA with resampling, which is a supervised, multivariate, whole-brain reduction technique, did not yield a significantly better classification performance than unsupervised, multivariate reduction using PCA or mass univariate selection using a *t*-test as well as unreduced data in the case of LOOCV^(rc). If LOOCV was performed only during classification, pLDA yielded significantly better classification accuracy than PCA because the pLDA results were considerably optimistically biased even though we performed resampling within the pLDA algorithm. Our results also showed that PCA was less prone to overtraining in the case of LOOCV^(c) because PCA does not use labels during data reduction compared with pLDA and a *t*-test. So, the results indicate that LOOCV^(c) is incorrect and lead to biased classification results (especially in the case of supervised data reduction methods despite of performed resampling within the reduction methods) and LOOCV^(rc) must be used instead.

In our experiment, we also investigated how data preprocessing influenced our classification results. We used two types of features obtained by different preprocessing of the MRI data, namely local deformations and GM volumes. We decided not to use the MRI intensities since they yielded considerably poorer classification results compared with the two other types of features in our previous study (Janousova et al., 2015). Here, we achieved similar performances using LOOCV^(rc) classification based on local deformations and GM volumes reduced using pLDA. The same was true for data reduced using a *t*-test. In the case of reduction using PCA, local deformations enabled a more accurate classification than was possible using GM volumes. When we performed LOOCV^(c), the GM volumes reduced by pLDA as well as a *t*-test led to slightly less overestimated results than in local deformations. The most discriminating local deformation features that were automatically detected by pLDA with resampling were located in the left prefrontal cortex, the cerebellar cortex, the medial parts of the thalamus and the right anterior insula. Furthermore, the most discriminating GM volume features also formed clusters in the inferior frontal gyrus, the anterior cingulum, the hippocampus and parahippocampal gyrus, and the superior and middle temporal gyrus. All of these results are consistent with the findings published in previous studies (Wright et al., 2000; Shenton et al., 2001; Niznikiewicz et al., 2003; Honea et al., 2005; Ellison-Wright et al., 2008). Moreover, it is known that these brain areas are involved in higher cognitive, integrative and regulatory functions that are impaired in schizophrenia (Niznikiewicz et al., 2003; Antonova et al., 2004). For example, prefrontal anomalies have been involved in negative symptoms

and cognitive impairments such as deficits in working memory and executive functions (Goldman-Rakic and Selemon, 1997). The validity of the selected most discriminative brain regions was also confirmed by *t*-test. Unfortunately, a direct comparison of areas of most discriminating local deformations and GM volumes is not possible because of different resolution of the two types of images. Nevertheless, the comparison of brain areas was not a goal of this study. We aimed at using imaging data at their highest possible resolution and we wanted to avoid loss of information in the local deformations by resampling the deformation images to $1.5 \times 1.5 \times 1.5$ mm. Regarding the features whose loadings contributed most to the first principal component, they covered almost entire brain or gray matter areas. Even if the thresholding of the loadings was stricter leading to smaller brain regions, those regions could not be considered as the discriminative ones because PCA does not aim for identification of discriminative features but rather for efficient data representation while preserving maximum amount of variance without regard to data labels (Kasperek et al., 2011, Janousova et al., 2015).

We additionally examined classification performance using two most common classification methods: LDA and linear SVM. We used LDA here for classification since we wished to ensure that our results would be comparable with other studies despite the fact that results obtained using LDA might be unstable in cases with a limited number of subjects for analysis (Thomaz et al., 2007). We also used linear SVM because of comparability with other studies and because we wished to avoid overtraining, which can arise when nonlinear kernels are used (Mourao-Miranda et al., 2012). Performing only the experiments with LOOCV during classification because a reduction using pLDA contained resampling that should prevent overestimating the results, we could state that LDA and linear SVM yield similar classification results (maximum accuracies of 85.6% (82.7% sens., 88.5% spec.) and 86.5% (88.5% sens., 84.6% spec.), respectively, for the classification of local deformations; and maximum accuracies of 83.7% (84.6% sens., 82.7% spec.) and 80.8% (80.8% sens., 80.8% spec.), respectively, for the classification of GM volumes). Such classification performance is superior to previous studies based on LDA that reported classification accuracies ranging from 70.7 to 82.9% based on MRI data (Leonard et al., 1999; Nakamura et al., 2004; Takayanagi et al., 2010; Karageorgiou et al., 2011; Ota et al., 2012) as well as a meta-analytical sensitivity of 76.4% (95% CI: 71.9–80.4%) and specificity of 79.0% (95% CI: 74.6–82.8%) based on 20 structural MRI studies (Kambeitz et al., 2015). Nevertheless, such LOOCV^(c) approach is incorrect. Thus, we also performed experiments with LOOCV used during data reduction and classification and we achieved poorer results; for the classification of local deformations using LDA and linear SVM, the maximum accuracy was 68.3% (63.5% sens., 73.1% spec.) and 70.2% (67.3% sens., 73.1% spec.), respectively. For the classification of GM volumes, the maximum accuracy that we achieved was 68.3% (67.3% sens., 69.2% spec.) and 67.3% (65.4% sens., 69.2% spec.), respectively. Classification based on data reduced using PCA or a *t*-test was even slightly less accurate apart from the classification of GM volumes reduced with a *t*-test with $p < 0.005$ using linear

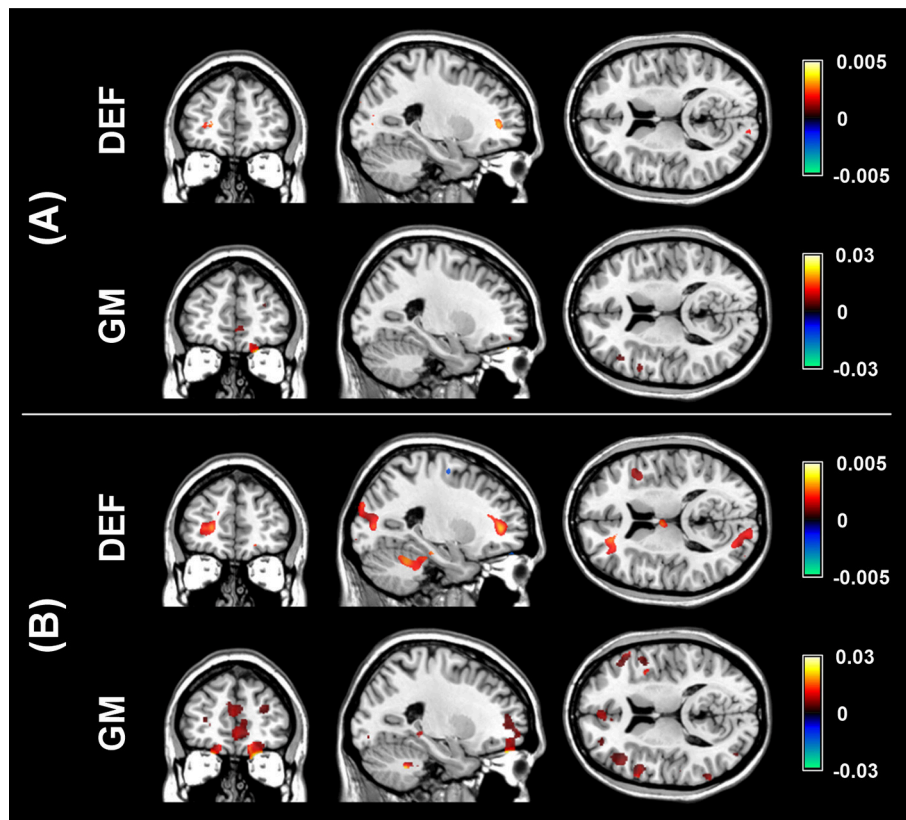


FIGURE 5 | Visualization of mean discriminative weights of local deformation (DEF) and gray matter (GM) volume features selected by pLDA with resampling (for $\lambda = 1.6$ and $\lambda = 0.9$, respectively) in: (A) all LOOCV^(rc) iterations, (B) more than half of LOOCV^(rc) iterations. The mean weights were calculated as an average of non-zero coefficients of \mathbf{v} over 100 resampling iterations and 104 LOOCV^(rc) folds.

SVM with LOOCV^(rc) with an accuracy of 69.2% (65.4% sens., 73.1% spec.).

The reason why our classification performance with LOOCV performed during data reduction and classification is so limited may be due to the fact that schizophrenia is not a single disease entity but instead a group of phenotypically similar disorders (although the current knowledge is not sufficient to isolate these disorders; Silveira et al., 2012). Another possibility is that our cohort of patients contained subjects with a future episodic course who could not be distinguished from healthy individuals (Mourao-Miranda et al., 2012). Such heterogeneity in data samples causes an overlap in discriminative features between patients and controls which leads to a drop of classification accuracy (Schnack and Kahn, 2016). Another reason explaining the limited classification performance of our results might be that the findings of other studies were biased because authors used a mixture of first-episode and chronic schizophrenia patients, a combination of male and female subjects, a small number of subjects, and/or incorrect classification. It is already known that the statistical power of neuroscience studies is often very low and that the effect size is frequently overestimated (Button et al., 2013). To avoid optimistically biased results, we used only patients with a first episode of schizophrenia in our study;

medication affects whole-brain GM volume (Ho et al., 2011), and therefore brain anatomy abnormalities observed at the first episode of schizophrenia are likely to be much more subtle relative to chronic schizophrenia (Ellison-Wright et al., 2008). Therefore, the classification performance can be overestimated if the data set contains chronic schizophrenia patients in addition to FES patients. For example, a previous study recovered an accuracy of 90.8% for male subjects and 91.8% for female subjects while discriminating a mixed group of FES and chronic patients from healthy individuals using the COMPARE algorithm (Fan et al., 2007); whereas the same algorithm yielded an accuracy of 73.4% for the classification of only FES and healthy subjects (Zanetti et al., 2013). Furthermore, to reduce bias in our results, we used only males in our study because it has been shown that there are differences in structural brain abnormalities in males and females (Nasrallah et al., 1990). In terms of the data sets used in classification experiments, an independent validation set should be used ideally (Nieuwenhuis et al., 2012; Schnack and Kahn, 2016). When no external validation set is available (which is unfortunately frequent in neuroimaging studies), Nieuwenhuis et al. (2012) recommend using more than 130 subjects as small number of subjects can lead to unstable classification models. Even though our study does not include such a large number

of subjects, our data set containing 104 subjects is still much larger than the data sets used in many previously published MRI-based classification studies of schizophrenia (Leonard et al., 1999; Kawasaki et al., 2007; Sun et al., 2009; Takayanagi et al., 2010; Karageorgiou et al., 2011; Kasperek et al., 2011; Borgwardt et al., 2012; Pettersson-Yeo et al., 2013). Concerning the influence of type of LOOCV, apart from our current study in which results based on LOOCV performed only during classification are noticeably overestimated, another example is in study of Nieuwenhuis et al. (2012) that achieved an accuracy of 86.8% in their discovery sample while selecting the top 10% of features ranked based on the absolute values of the elements of the weight of the SVM vector compared with an accuracy of 69.1% obtained using an independent validation data set based on the selected features. When no data reduction was performed, these authors obtained similar accuracies in their discovery sample (71.4%) and validation sample (70.4%).

Our current results, as well as a meta-analysis by Kambeitz et al. (2015), suggest that even though much work has focused on using neuroimaging data to distinguish schizophrenia patients from healthy individuals, the classification performance is still too low to be considered for clinical application. Therefore, researchers should focus on continuing schizophrenia research and improving as well as inventing data reduction and classification methods to attain better classification performances; early and accurate diagnoses have the ability to shorten the duration of untreated psychosis and therefore improve therapeutic outcomes and overall patient prognoses (Perkins et al., 2005). One way of improving classification performance may be to use different imaging modalities, such as diffusion tensor imaging (DTI) or functional MRI. These modalities are capable of revealing information about connections inside the brain which could be of help in

classification as schizophrenia is characterized by a deficit of interconnections (Michael et al., 2008). A promising accuracy of 96% for the classification of DTI data has already been reported (Ardekani et al., 2011), but the results may be a bit too optimistic since the authors used a mixture of FES and chronic patients and numbers of patients and controls were low (25 subjects in each group). Resting-state functional MRI also exhibits promising results since its sensitivity was higher than that of structural MRI studies (Kambeitz et al., 2015). However, the resting-state functional MRI studies were based on quite small samples; while using large samples, the classification accuracy will probably decrease (Schnack and Kahn, 2016). To conclude, even though the classification performance of schizophrenia classification studies is still too low for application in schizophrenia diagnostics, schizophrenia research should continue with future studies performed on large samples free of chronic schizophrenia patients using correctly designed and performed classification procedures.

AUTHOR CONTRIBUTIONS

EJ designed the study, performed the analysis, and wrote the manuscript. GM participated in the reduction algorithm design and the manuscript writing. TK contributed to study design, collected the data and participated in interpretation of analysis results. DS contributed to study design, performed image data preprocessing and critically revised the manuscript. All authors read and approved the final manuscript.

ACKNOWLEDGMENTS

The work was supported by grant IGA MH CZ NR No. NT 13359-4/2012.

REFERENCES

- Andreasen, N. C. (1995). Symptoms, signs, and diagnosis of schizophrenia. *Lancet* 346, 477–481. doi: 10.1016/S0140-6736(95)91325-4
- Antonova, E., Sharma, T., Morris, R., and Kumari, V. (2004). The relationship between brain structure and neurocognition in schizophrenia: a selective review. *Schizophr. Res.* 70, 117–145. doi: 10.1016/j.schres.2003.12.002
- Ardekani, B. A., Tabesh, A., Sevy, S., Robinson, D. G., Bilder, R. M., and Szeszko, P. R. (2011). Diffusion tensor imaging reliably differentiates patients with schizophrenia from healthy volunteers. *Hum. Brain Mapp.* 32, 1–9. doi: 10.1002/hbm.20995
- Borgwardt, S., Koutsouleris, N., Aston, J., Studerus, E., Smieskova, R., Riecher-Rössler, A., et al. (2012). Distinguishing prodromal from first-episode psychosis using neuroanatomical single-subject pattern recognition. *Schizophr. Bull.* 39, 1105–1114. doi: 10.1093/schbul/sbs095
- Bunea, F., She, Y., Ombao, H., Gongvatana, A., Devlin, K., and Cohen, R. (2011). Penalized least squares regression methods and applications to neuroimaging. *Neuroimage* 55, 1519–1527. doi: 10.1016/j.neuroimage.2010.12.028
- Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., et al. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nat. Rev. Neurosci.* 14, 365–376. doi: 10.1038/nrn3475
- Castro, E., Gómez-Verdejo, V., Martínez-Ramón, M., Kiehl, K. A., and Calhoun, V. D. (2014). A multiple kernel learning approach to perform classification of groups from complex-valued fMRI data analysis: application to schizophrenia. *Neuroimage* 87, 1–17. doi: 10.1016/j.neuroimage.2013.10.065
- Demirci, O., Clark, V. P., Magnotta, V. A., Andreasen, N. C., Lauriello, J., Kiehl, K. A., et al. (2008). A review of challenges in the use of fMRI for disease classification/characterization and a projection pursuit application from a multi-site fMRI schizophrenia study. *Brain Imaging Behav.* 2, 207–226. doi: 10.1007/s11682-008-9028-1
- Diaz, G., Romero, E., Hernández-Tamames, J. A., Molina, V., and Malpica, N. (2010). Automatic classification of structural MRI for diagnosis of neurodegenerative diseases. *Acta Biol. Colomb.* 15, 165–180.
- Dluhos, P., Schwarz, D., and Kasperek, T. (2014). Wavelet features for recognition of first episode of schizophrenia from MRI brain images. *Radioengineering* 23, 274–281. Available online at: http://www.radioeng.cz/fulltexts/2014/14_01_0274_0281.pdf
- Ellison-Wright, I., Glahn, D. C., Laird, A. R., Thelen, S. M., and Bullmore, E. (2008). The anatomy of first-episode and chronic schizophrenia: an anatomical likelihood estimation meta-analysis. *Am. J. Psychiatry* 165, 1015–1023. doi: 10.1176/appi.ajp.2008.07101562
- Fan, Y., Shen, D. G., and Davatzikos, C. (2005). “Classification of structural images via high-dimensional image warping, robust feature extraction, and SVM,” in *Medical Image Computing and Computer-Assisted Intervention - Miccai 2005*, eds J. S. Duncan and G. Gerig (Berlin: Springer-Verlag Berlin), 1–8.
- Fan, Y., Shen, D., Gur, R. C., Gur, R. E., and Davatzikos, C. (2007). COMPARE: classification of morphological patterns using adaptive regional elements. *IEEE Trans. Med. Imaging* 26, 93–105. doi: 10.1109/TMI.2006.886812

- Goldman-Rakic, P. S., and Selemon, L. D. (1997). Functional and anatomical aspects of prefrontal pathology in schizophrenia. *Schizophr. Bull.* 23, 437–458. doi: 10.1093/schbul/23.3.437
- Good, C. D., Johnsrude, I. S., Ashburner, J., Henson, R. N. A., Friston, K. J., and Frackowiak, R. S. J. (2001). A voxel-based morphometric study of ageing in 465 normal adult human brains. *Neuroimage* 14, 21–36. doi: 10.1006/nimg.2001.0786
- Guyon, I., and Elisseeff, A. (2003). An introduction to variable and feature selection. *J. Mach. Learn. Res.* 3, 1157–1182. Available online at: <http://www.jmlr.org/papers/volume3/guyon03a/guyon03a.pdf>
- Haijma, S. V., Van Haren, N., Cahn, W., Koolschijn, P. C., Hulshoff Pol, H. E., and Kahn, R. S. (2013). Brain volumes in schizophrenia: a meta-analysis in over 18 000 subjects. *Schizophr. Bull.* 39, 1129–1138. doi: 10.1093/schbul/sbs118
- Ho, B.-C., Andreasen, N. C., Ziebell, S., Pierson, R., and Magnotta, V. (2011). Long-term antipsychotic treatment and brain volumes: a longitudinal study of first-episode schizophrenia. *Arch. Gen. Psychiatry* 68, 128–137. doi: 10.1001/archgenpsychiatry.2010.199
- Honea, R., Crow, T. J., Passingham, D., and Mackay, C. E. (2005). Regional deficits in brain volume in schizophrenia: a meta-analysis of voxel-based morphometry studies. *Am. J. Psychiatry* 162, 2233–2245. doi: 10.1176/appi.ajp.162.12.2233
- Janousova, E., Schwarz, D., and Kasperek, T. (2015). Combining various types of classifiers and features extracted from magnetic resonance imaging data in schizophrenia recognition. *Psychiatry Res. Neuroimaging* 232, 237–249. doi: 10.1016/j.pscychres.2015.03.004
- Kambeitz, J., Kambeitz-Illankovic, L., Leucht, S., Wood, S., Davatzikos, C., Malchow, B., et al. (2015). Detecting neuroimaging biomarkers for schizophrenia: a meta-analysis of multivariate pattern recognition studies. *Neuropsychopharmacology* 40, 1742–1751. doi: 10.1038/npp.2015.22
- Karageorgiou, E., Schulz, S. C., Gollub, R. L., Andreasen, N. C., Ho, B.-C., Lauriello, J., et al. (2011). Neuropsychological testing and structural magnetic resonance imaging as diagnostic biomarkers early in the course of schizophrenia and related psychoses. *Neuroinformatics* 9, 321–333. doi: 10.1007/s12021-010-9094-6
- Kasperek, T., Thomaz, C. E., Sato, J. R., Schwarz, D., Janousova, E., Marecek, R., et al. (2011). Maximum-uncertainty linear discrimination analysis of first-episode schizophrenia subjects. *Psychiatry Res. Neuroimaging* 191, 174–181. doi: 10.1016/j.pscychres.2010.09.016
- Kawasaki, Y., Suzuki, M., Kherif, F., Takahashi, T., Zhou, S.-Y., Nakamura, K., et al. (2007). Multivariate voxel-based morphometry successfully differentiates schizophrenia patients from healthy controls. *Neuroimage* 34, 235–242. doi: 10.1016/j.neuroimage.2006.08.018
- Klöppel, S., Abdulkadir, A., Jack, C. R., Koutsouleris, N., Mourao-Miranda, J., and Vemuri, P. (2012). Diagnostic neuroimaging across diseases. *Neuroimage* 61, 457–463. doi: 10.1016/j.neuroimage.2011.11.002
- Kovalev, V. A., Petrou, M., and Suckling, J. (2003). Detection of structural differences between the brains of schizophrenic patients and controls. *Psychiatry Res. Neuroimaging* 124, 177–189. doi: 10.1016/S0925-4927(03)00070-2
- Lemm, S., Blankertz, B., Dickhaus, T., and Mueller, K.-R. (2011). Introduction to machine learning for brain imaging. *Neuroimage* 56, 387–399. doi: 10.1016/j.neuroimage.2010.11.004
- Leonard, C. M., Kuldau, J. M., Breier, J. I., Zuffante, P. A., Gautier, E. R., Heron, D. C., et al. (1999). Cumulative effect of anatomical risk factors for schizophrenia: an MRI study. *Biol. Psychiatry* 46, 374–382. doi: 10.1016/S0006-3223(99)0052-9
- Liddle, P. F. (1987). The symptoms of chronic schizophrenia. a re-examination of the positive-negative dichotomy. *Br. J. Psychiatry* 151, 145–151. doi: 10.1192/bjp.151.2.145
- Marquand, A. F., Rezek, I., Buitelaar, J., and Beckmann, C. F. (2016). Understanding heterogeneity in clinical cohorts using normative models: beyond case-control studies. *Biol. Psychiatry*. doi: 10.1016/j.biopsych.2015.12.023. [Epub ahead of print].
- Meinshausen, N., and Bühlmann, P. (2010). Stability selection. *J. R. Stat. Soc. B-Stat. Methodol.* 72, 417–473. doi: 10.1111/j.1467-9868.2010.00740.x
- Melonakos, E. D., Shenton, M. E., Rathi, Y., Terry, D. P., Bouix, S., and Kubicki, M. (2011). Voxel-based morphometry (VBM) studies in schizophrenia-can white matter changes be reliably detected with VBM? *Psychiatry Res.* 193, 65–70. doi: 10.1016/j.pscychres.2011.01.009
- Michael, A. M., Calhoun, V. D., Andreasen, N. C., and Baum, S. A. (2008). “A method to classify schizophrenia using inter-task spatial correlations of functional brain images,” in *2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Vol. 1-8 (New York, NY: IEEE), 5510–5513.
- Mourao-Miranda, J., Reinders, A. A. T. S., Rocha-Rego, V., Lappin, J., Rondina, J., Morgan, C., et al. (2012). Individualized prediction of illness course at the first psychotic episode: a support vector machine MRI study. *Psychol. Med.* 42, 1037–1047. doi: 10.1017/S0033291711002005
- Nakamura, K., Kawasaki, Y., Suzuki, M., Hagino, H., Kurokawa, K., Takahashi, T., et al. (2004). Multiple structural brain measures obtained by three-dimensional magnetic resonance imaging to distinguish between schizophrenia patients and normal subjects. *Schizophr. Bull.* 30, 393–404. doi: 10.1093/oxfordjournals.schbul.a007087
- Nasrallah, H., Schwarzkopf, S., Olson, S., and Coffman, J. (1990). Gender differences in schizophrenia on MRI brain-scans. *Schizophr. Bull.* 16, 205–209. doi: 10.1093/schbul/16.2.205
- Nieuwenhuis, M., van Haren, N. E. M., Pol, H. E. H., Cahn, W., Kahn, R. S., and Schnack, H. G. (2012). Classification of schizophrenia patients and healthy controls from structural MRI scans in two large independent samples. *Neuroimage* 61, 606–612. doi: 10.1016/j.neuroimage.2012.03.079
- Niznikiewicz, M. A., Kubicki, M., and Shenton, M. E. (2003). Recent structural and functional imaging findings in schizophrenia. *Curr. Opin. Psychiatry* 16, 123–147. doi: 10.1097/00001504-200303000-00002
- Ota, M., Sato, N., Ishikawa, M., Hori, H., Sasayama, D., Hattori, K., et al. (2012). Discrimination of female schizophrenia patients from healthy women using multiple structural brain measures obtained with voxel-based morphometry. *Psychiatry Clin. Neurosci.* 66, 611–617. doi: 10.1111/j.1440-1819.2012.02397.x
- Perkins, D. O., Gu, H. B., Boteva, K., and Lieberman, J. A. (2005). Relationship between duration of untreated psychosis and outcome in first-episode schizophrenia: a critical review and meta-analysis. *Am. J. Psychiatry* 162, 1785–1804. doi: 10.1176/appi.ajp.162.10.1785
- Pettersson-Yeo, W., Benetti, S., Marquand, A. F., Dell’Acqua, F., Williams, S. C. R., Allen, P., et al. (2013). Using genetic, cognitive and multi-modal neuroimaging data to identify ultra-high-risk and first-episode psychosis at the individual level. *Psychol. Med.* 43, 2547–2562. doi: 10.1017/S003329171300024X
- Plis, S. M., Hjelm, D. R., Salakhutdinov, R., Allen, E. A., Bockholt, H. J., Long, J. D., et al. (2014). Deep learning for neuroimaging: a validation study. *Front. Neurosci.* 8:229. doi: 10.3389/fnins.2014.00229
- Pohl, K. M., and Sabuncu, M. R. (2009). “A unified framework for mr based disease classification,” in *Information Processing in Medical Imaging, Proceedings*, eds J. L. Prince, D. L. Pham, and K. J. Myers (Berlin: Springer-Verlag Berlin), 300–313.
- Rosa, M. J., Portugal, L., Hahn, T., Fallgatter, A. J., Garrido, M. I., Shawe-Taylor, J., et al. (2015). Sparse network-based models for patient classification using fMRI. *Neuroimage* 105, 493–506. doi: 10.1016/j.neuroimage.2014.11.021
- Ryali, S., Supekar, K., Abrams, D. A., and Menon, V. (2010). Sparse logistic regression for whole-brain classification of fMRI data. *Neuroimage* 51, 752–764. doi: 10.1016/j.neuroimage.2010.02.040
- Schnack, H. G., and Kahn, R. S. (2016). Detecting neuroimaging biomarkers for psychiatric disorders: sample size matters. *Front. Psychiatry* 7:50. doi: 10.3389/fpsy.2016.00050
- Schwarz, D., Kasperek, T., Provaznik, I., and Jarkovsky, J. (2007). A deformable registration method for automated morphometry of MRI brain images in neuropsychiatric research. *IEEE Trans. Med. Imaging* 26, 452–461. doi: 10.1109/TMI.2007.892512
- Shenton, M. E., Dickey, C. C., Frumin, M., and McCarley, R. W. (2001). A review of MRI findings in schizophrenia. *Schizophr. Res.* 49, 1–52. doi: 10.1016/S0920-9964(01)00163-3
- Silveira, C., Marques-Teixeira, J., and de Bastos-Leite, A. J. (2012). More than one century of schizophrenia: an evolving perspective. *J. Nerv. Ment. Dis.* 200, 1054–1057. doi: 10.1097/NMD.0b013e318275d249
- Simeone, J. C., Ward, A. J., Rotella, P., Collins, J., and Windisch, R. (2015). An evaluation of variation in published estimates of schizophrenia prevalence from 1990-2013: a systematic literature review. *BMC Psychiatry* 15:193. doi: 10.1186/s12888-015-0578-7
- Sun, D., van Erp, T. G. M., Thompson, P. M., Bearden, C. E., Daley, M., Kushan, L., et al. (2009). Elucidating a magnetic resonance imaging-based neuroanatomic

- biomarker for psychosis: classification analysis using probabilistic brain atlas and machine learning algorithms. *Biol. Psychiatry* 66, 1055–1060. doi: 10.1016/j.biopsych.2009.07.019
- Takayanagi, Y., Kawasaki, Y., Nakamura, K., Takahashi, T., Orikabe, L., Toyoda, E., et al. (2010). Differentiation of first-episode schizophrenia patients from healthy controls using ROI-based multiple structural brain variables. *Prog. Neuro-Psychopharmacol. Biol. Psychiatry* 34, 10–17. doi: 10.1016/j.pnpbp.2009.09.004
- Thomaz, C. E., Boardman, J. P., Counsell, S., Hill, D. L. G., Hajnal, J. V., Edwards, A. D., et al. (2007). A multivariate statistical analysis of the developing human brain in preterm infants. *Image Vis. Comput.* 25, 981–994. doi: 10.1016/j.imavis.2006.07.011
- Vounou, M., Janousova, E., Wolz, R., Stein, J. L., Thompson, P. M., Rueckert, D., et al. (2012). Sparse reduced-rank regression detects genetic associations with voxel-wise longitudinal phenotypes in Alzheimer's disease. *Neuroimage* 60, 700–716. doi: 10.1016/j.neuroimage.2011.12.029
- Witten, D. M., and Tibshirani, R. (2011). Penalized classification using Fisher's linear discriminant. *J. R. Stat. Soc. Ser. B-Stat. Methodol.* 73, 753–772. doi: 10.1111/j.1467-9868.2011.00783.x
- Wolfers, T., Buitelaar, J. K., Beckmann, C. F., Franke, B., and Marquand, A. F. (2015). From estimating activation locality to predicting disorder: a review of pattern recognition for neuroimaging-based psychiatric diagnostics. *Neurosci. Biobehav. Rev.* 57, 328–349. doi: 10.1016/j.neubiorev.2015.08.001
- Wright, I. C., Rabe-Hesketh, S., Woodruff, P. W. R., David, A. S., Murray, R. M., and Bullmore, E. T. (2000). Meta-analysis of regional brain volumes in schizophrenia. *Am. J. Psychiatry* 157, 16–25. doi: 10.1176/ajp.157.1.16
- Yoon, U., Lee, J.-M., Im, K., Shin, Y.-W., Cho, B. H., Kim, I. Y., et al. (2007). Pattern classification using principal components of cortical thickness and its discriminative pattern in schizophrenia. *Neuroimage* 34, 1405–1415. doi: 10.1016/j.neuroimage.2006.11.021
- Zanetti, M. V., Schaufelberger, M. S., Doshi, J., Ou, Y., Ferreira, L. K., Menezes, P. R., et al. (2013). Neuroanatomical pattern classification in a population-based sample of first-episode schizophrenia. *Prog. Neuropsychopharmacol. Biol. Psychiatry* 43, 116–125. doi: 10.1016/j.pnpbp.2012.12.005

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2016 Janousova, Montana, Kasperek and Schwarz. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.