Integrating spatial-anatomical regularization and structure sparsity into SVM
Improving interpretation of Alzheimer's disease classification

Sun, Zhuo; Qiao, Yuchuan; Lelieveldt, Boudewijn P.F.; Staring, Marius

DOI
10.1016/j.neuroimage.2018.05.051

Publication date
2018

Document Version
Final published version

Published in
NeuroImage

NeuroImage 178 (2018) 445–460

Contents lists available at ScienceDirect

NeuroImage

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

Integrating spatial-anatomical regularization and structure sparsity into SVM: Improving interpretation of Alzheimer's disease classification

Zhuo Sun [a], Yuchuan Qiao [a], Boudewijn P.F. Lelieveldt [a,b], Marius Staring [a,b,*], for the Alzheimer's Disease NeuroImaging Initiative[1]

[a] Division of Image Processing, Department of Radiology, Leiden University Medical Center, 2300, RC, Leiden, The Netherlands
[b] Intelligent System Group, Faculty of EEMCS, Delft University of Technology, 2600, GA, Delft, The Netherlands

ARTICLE INFO

Keywords:
Alzheimer's disease
Support vector machine (SVM)
Spatial-anatomical regularization
Structure sparsity
Proximal algorithm

ABSTRACT

In recent years, machine learning approaches have been successfully applied to the field of neuroimaging for classification and regression tasks. However, many approaches do not give an intuitive relation between the raw features and the diagnosis. Therefore, they are difficult for clinicians to interpret. Moreover, most approaches treat the features extracted from the brain (for example, voxelwise gray matter concentration maps from brain MRI) as independent variables and ignore their spatial and anatomical relations. In this paper, we present a new Support Vector Machine (SVM)-based learning method for the classification of Alzheimer's disease (AD), which integrates spatial-anatomical information. In this way, spatial-neighbor features in the same anatomical region are encouraged to have similar weights in the SVM model. Secondly, we introduce a group lasso penalty to induce structure sparsity, which may help clinicians to assess the key regions involved in the disease. For solving this learning problem, we use an accelerated proximal gradient descent approach. We tested our method on the subset of ADNI data selected by Cuingnet et al. (2011) for Alzheimer's disease classification, as well as on an independent larger dataset from ADNI. Good classification performance is obtained for distinguishing cognitive normals (CN) vs. AD, as well as on distinguishing between various sub-types (e.g. CN vs. Mild Cognitive Impairment). The model trained on Cuignet's dataset for AD vs. CN classification was directly used without re-training to the independent larger dataset. Good performance was achieved, demonstrating the generalizability of the proposed methods. For all experiments, the classification results are comparable or better than the state-of-the-art, while the weight map more clearly indicates the key regions related to Alzheimer's disease.

Introduction

In recent years, an increasing number of people suffer from neurodegenerative diseases, such as Alzheimer's Disease (AD) or Parkinson's disease. AD is usually diagnosed in people over 65 years old (Alzheimer's Association, 2014), when there are clear symptoms. It is reported that the number of AD patients worldwide will increase from currently 26.6 million to 100 million by the year 2050. Early detection of AD can largely improve the treatment of AD and many groups are focusing on this problem from different angles. Different kinds of biomarkers have been investigated for AD detection, e.g. structural brain MRI (Frisoni et al., 2010), metabolic brain alterations measured by fluorodeoxyglucose positron emission tomography (FDG-PET) (De Santi et al., 2001), or pathological amyloid depositions measured from cerebrospinal fluid (CSF) (Leon et al., 2007; Mattsson et al., 2009). Among all these measurements, Magnetic Resonance Imaging (MRI) plays an increasingly important role, owning to its noninvasiveness, availability, and high sensitivity to brain changes after disease onset (Frisoni et al., 2010). Therefore, it is commonly used as part of the standard clinical assessment for the diagnosis of AD. Due to its ability to visualize the brain morphology at high spatial resolution (Liang and Lauterbur, 2000), it is an ideal tool to study the various brain structures and the morphological changes caused by AD. Automatically distinguishing AD from cognitive normal (CN), or from Mild Cognitive Impairment (MCI) is an important

* Corresponding author. Division of Image Processing, Department of Radiology, Leiden University Medical Center, 2300, RC, Leiden, The Netherlands.
E-mail address: m.staring@lumc.nl (M. Staring).

[1] Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.ucla.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at:_http://adni.loni.ucla.edu/wp-content/uploadsΩ/how_to_apply/ADNI_Acknowledgement_List.pdf.

https://doi.org/10.1016/j.neuroimage.2018.05.051
Received 14 February 2018; Received in revised form 10 April 2018; Accepted 21 May 2018
Available online 23 May 2018

step to understand AD progression and help clinicians to make a decision.

In the last ten years, many structural MRI-based methods have been proposed for automatic AD detection (Cuingnet et al., 2013; Davatzikos et al., 2008; Fan et al., 2005, 2007; 2008b; Klöppel et al., 2008). Based on the features used, these methods can, in general, be divided into three categories: measurements based on brain structures, measurements based on adaptively generated region-of-interest (ROIs), and voxelwise measurements. In the first category, some methods have focused on the structures that are known to be related to AD, such as the hippocampus and the ventricle (Coupé et al., 2011), and perform classification based on features derived from these structures (Hampel et al., 2002; Coupé et al., 2012b; a; Sun et al., 2012). These methods depend strongly on the prior selection of structures. To avoid such bias, other methods consider features from all brain structures (Oliveira et al., 2010; Liu et al., 2013). The description of each structure is typically condensed to a scalar or low dimensional representation, potentially disregarding detailed information inside a structure. Some methods (Fan et al., 2007) divide the brain region into supervoxels, thereby improving on the level of detail. Finally, there are methods that directly work on voxelwise features (Cuingnet et al., 2013; Klöppel et al., 2008), to fully take advantage of the high resolution in structural MRI.

A linear support vector machine (SVM) is frequently used for AD classification based on voxelwise features, because it is easy to use and understand. SVMs learn a weight map, which can be used to indicate the importance of each voxel in distinguishing CN from AD subjects. A linear SVM, however, has several problems. First, spatial information is ignored, as each voxelwise feature is treated completely independent and ignores neighborhood information. However, one would expect spatial smoothness in the weight map, as neighboring tissue tends to be similarly affected by AD. Second, knowledge of the anatomy is neglected, while typically voxels in the same brain structure are similarly involved in AD. Third, in linear SVMs almost all weights will be nonzero, making it difficult to distinguish areas that are highly involved in AD from those that are not. These factors complicate interpretation of the learned model.

To address these problems, several solutions have been proposed. Fan et al. (2008a) proposed a lasso SVM method that replaced the max-margin in the standard linear SVM with ℓ_1 norm regularization to encourage sparsity of the weights. This method solves the third problem by selecting only the most relevant voxels. However, it does not address spatial or anatomical smoothness, so the selected voxels are scattered over the image, making the interpretation difficult. Voxel selection in lasso SVM is also known to be unstable, meaning that small differences in the training images can result in very different weight maps (Dunne et al., 2002; Xin et al., 2016). The elastic-net method (Zou and Hastie, 2005) combines the ℓ_1 sparsity norm with a max-margin term, and is widely used in the field of neuro-imaging (Wu et al., 2017; Wachinger et al., 2016; Kohannim et al., 2012; Shen et al., 2011). It encourages a grouping effect, where strongly correlated voxels tend to be in or out of the model together. The grouping, however, is implicit and not based on spatial or anatomical information. Similar to the elastic-net, the graph-net method (Grosenick et al., 2013) combines the ℓ_1 sparsity norm, but then with a graph-Laplacian regularizer. This regularizer encourages neighboring voxels to have similar weights. Regularization is however also performed across anatomical boundaries, but not all brain structures are equally affected by AD, even if they are close. Even though the weight map is spatially smooth, graph-net may not select anatomically meaningful areas. Cuingnet et al. (2013) proposed a spatial-anatomical regularized SVM model that penalizes both spatial non-smoothness and anatomical non-smoothness. Like graph-net, smoothing is performed across the structure boundary, and also no sparsity is considered. In our previous work (Sun et al., 2015a), we replaced the ℓ_1 sparsity norm in standard lasso SVM by a group lasso sparsity term, which integrates anatomical information. However, spatial smoothness was not included.

Different from the aforementioned papers, we want to simultaneously solve the three problems listed above. In this paper, we propose a new method that integrates regularization and grouping using a group lasso formulation together with a spatial-anatomical regularization term in the SVM cost function. Our goal is for the learned model to more clearly indicate which anatomical regions are important for the classifier to distinguish between clinical groups, which aids the interpretation of the learned model. Simultaneously we aim to improve classification results with respect to the baseline SVM method. In contrast to previous work (Grosenick et al., 2013; Cuingnet et al., 2013), we do not penalize non-smoothness of the weight map across the structure boundary, since tissues belonging to different structures may be affected by AD quite differently (Pegueroles et al., 2017). The introduction of anatomical information into the proposed method, unlike the elastic-net and graph-net methods, also aids in spatial grouping in an anatomically meaningful manner. Compared to our previous work (Sun et al., 2015a), here we add spatial-anatomical regularization to improve the smoothness of the resulting weight map. We propose a mathematical formulation of the combined cost function that does not require the inversion of a potentially large regularization matrix, like Cuingnet et al. (2013). This formulation also allows future extensions with new regularization terms, which seems not easy to do in the dual space used in (Cuingnet et al., 2013). In addition, we introduce a level of detail in between complete anatomical structures and voxels. This is achieved by the use of supervoxels, where we propose a modification of the Simple Linear Iterative Clustering (SLIC) algorithm (Achanta et al., 2012) that respects anatomical boundaries defined by an atlas.

The remainder of the paper is organized as follows. In Section 2, we briefly describe the feature used in this paper and the basic idea of a linear SVM. Then in Section 3, we introduce the new regularization components and describe how to minimize the new cost function using the FISTA algorithm (Beck and Teboulle, 2009). In Section 4, we describe the datasets used in this paper and apply the proposed model to analyze this 3D real brain dataset in Section 5. Discussion and conclusion are given in Section 6 and 7, respectively.

Preliminaries

In this section, we first describe the voxelwise features extracted from brain MRI. Then we offer background on linear SVMs and their regularization.

Voxelwise features

In biomedical imaging, voxelwise features are commonly used to represent local properties. In this paper, we focus on the gray matter density feature (Ashburner and Friston, 2000) derived from T1-weighted MRI, which is commonly used for AD classification (Klöppel et al., 2008; Fan et al., 2007; Cuingnet et al., 2011). However, any type of voxelwise feature or combination thereof can be seamlessly integrated in the proposed framework.

Given T1-weighted images $I_i, \{i = 1, 2, \ldots, n\}$ of n subjects, each image and its corresponding tissue segmentation are deformed to the SPM template T. A modulated gray matter tissue density map G_i is computed by multiplying the spatially normalized gray matter map with the Jacobian determinant of the deformation. For each subject, the feature vector $\mathbf{x}_i$ is extracted from voxels inside the brain mask in the template space as a D-dimensional feature $\mathbf{x}_i \in \mathbb{R}^D$, and its associated clinical diagnosis is $y_i \in \{-1, 1\}$ to indicate different categories in each classification task (e.g. CN and AD).

Linear SVM

In the standard linear SVM, the optimal weight and bias parameters $\{\mathbf{w}^{\text{opt}}, b^{\text{opt}}\}$ are computed by solving the minimization problem:

$$\{\mathbf{w}^{\text{opt}}, b^{\text{opt}}\} = \underset{\mathbf{w}\in\mathbb{R}^D, b\in\mathbb{R}}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n} \mathscr{L}_{\text{hinge}}\big(y_i(<\mathbf{w}, \mathbf{x}_i> + b)\big)^2 + \frac{1}{2}\lambda \parallel \mathbf{w}\parallel_2^2, \qquad (1)$$

where $\lambda \in \mathbb{R}^+$ is a non-negative regularization parameter for the max-margin penalty $||\mathbf{w}||_2^2$ to encourage wider separation of the classes. The max-margin term assumes that each feature is independent, which is suitable for many machine learning tasks that have little prior knowledge about the relationship between the features. The hinge loss function $\mathscr{L}_{\text{hinge}}(u) = \max(0, 1-u)$ introduces a soft margin in the SVM that is zero for samples at the correct side of the hyperplane, and proportional to the distance to hyperplane when at the wrong side.

The dimension of the weights $\mathbf{w}$ is the same as the dimension of the feature vector $\mathbf{x}_i$. For voxelwise features, each $\mathbf{w}_i$ therefore corresponds to a point in the image space. This enables mapping the weight vector $\mathbf{w}$ into the image space, so as to enable natural visualization of the weights, see Fig. 7a for an example.

Methods

In this section, we will introduce a term that enforces weight sparsity in Section 3.2. Section 3.3 proposes a spatial-anatomical regularization (SAR) term that will enforce smoothness between neighbors as well as within predefined structures. These two new terms are integrated into the linear SVM cost function in Section 3.4, where we also describe how to optimize the resulting non-differentiable cost function using the FISTA approach. First, in Section 3.1, we describe a method to obtain the predefined structures by combining atlas-derived brain regions with an adapted supervoxel approach.

Anatomically constrained supervoxels

In neuroimaging, there are several ways to cluster brain voxels. The most common way is to define a brain structure segmentation in a template space; here we use the SPM template space. In this work, we first use the MINC standard pipeline (Coupé et al., 2015) to automatically segment the MNI152 atlas in 35 regions. Then we use elastix (Klein et al., 2010) to register the MNI152 atlas with the SPM atlas, using a B-spline transformation model. The 35 regions are propagated to the SPM space to obtain the segmentation S, see Fig. 2a. At this point, we have obtained grouping of voxels in the template space, based on anatomical structures.

Anatomy-based grouping may, however, be sub-optimal for classification as the two are not related, and secondly since information may be hidden at sub-regional parts of the anatomy. This was indeed demonstrated by Fan et al. (2007), who showed that the use of supervoxels could improve the accuracy of the classifier. However, being based on the watershed method, anatomical boundaries were ignored.

Therefore, we propose a modification of the SLIC supervoxel segmentation method (Achanta et al., 2012), which respects anatomical boundaries. Similar to (Fan et al., 2007), the method takes into account the correlation of the feature $\mathbf{x}(j)$ (gray matter density at point j) with the disease label $\mathbf{y}$, and relates the grouping to the classification problem. This is different from most supervoxel methods, where grouping is based on image intensity. For each voxel j, we compute the Pearson Correlation Coefficient (PCC) as follows:

$$\text{PCC}(j) = \frac{\text{cov}(\mathbf{x}(j), \mathbf{y})}{\text{cov}(\mathbf{x}(j))\text{cov}(\mathbf{y})}. \tag{2}$$

The result is shown in Fig. 1 and we assume voxels within the same supervoxel region tend to have similar discriminative power.

In order to avoid the supervoxels to cross anatomical boundaries, we modify the original cost function D used in the SLIC method. We define the following modified cost function for each cluster k:

$$D(I, S, C_k, j) = \sqrt{d_s(C_k, j)^2 + \eta d_c(I, C_k, j)^2 + d_a(S, C_k, j)^2}, \tag{3}$$

where $d_s(C_k, j)$, $d_c(I, C_k, j)$ are the spatial and content-based distances defined in SLIC (Achanta et al., 2012) and η is a positive coefficient that

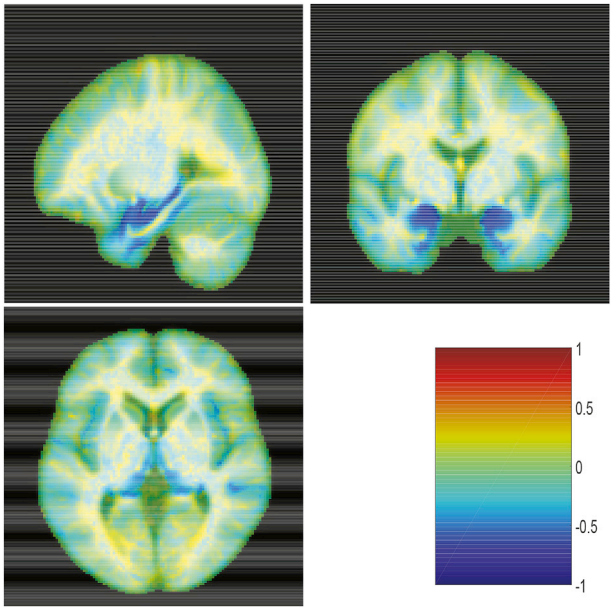


Fig. 1. The Pearson Correlation Coefficient map of AD vs CN on the training set, in three orthogonal views.

controls the balance between these two terms. A smaller η tends to generate more spherical supervoxels with a larger variation in the intensity distribution of the content, and vice versa. We have added a cost $d_a(S, C_k, j)$, which is a distance based on an available anatomical label map S. To forbid a supervoxel to be part of multiple anatomical regions, we define the anatomical distance d_a as an inverse Dirac type function:

$$d_a(S, C_k, j) = \begin{cases} 0 & \text{if } S(j) = S(C_k) \\ \infty & \text{otherwise,} \end{cases} \tag{4}$$

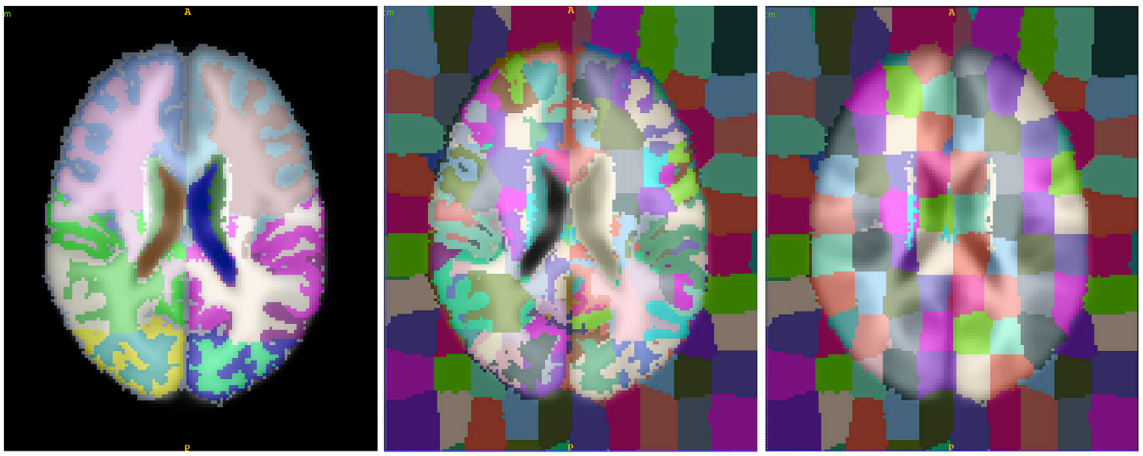
where $S(j)$ and $S(C_k)$ are the anatomical segmentations at voxel j and cluster center C_k. If voxel j and cluster C_k have the same anatomical label ($S(j) = S(C_k)$), this label penalty is 0, while if their anatomical labels are different, this penalty will be positive infinite. Each voxel j is assigned to the cluster k that minimizes the cost function D. Therefore, by adding the penalty term (4) to the overall cost function (3), each voxel j will only be assigned to a cluster C_k that satisfies the condition $S(j) = S(C_k)$. For the segmentation S we use the segmentation in the SPM space, and for the input image I we use the PCC result instead of image intensity.

The difference between the proposed ac-SLIC method with the standard SLIC method can be seen in Fig. 2. We can see that the proposed method follows anatomical boundaries as defined by the anatomical labels S.

In (Achanta et al., 2012), the authors showed that the complexity of the standard SLIC method is linear with respect to the number of voxels in the image and independent of the number of supervoxels. Therefore, SLIC is faster and more memory efficient than state-of-the-art methods (Veksler et al., 2010; Vedaldi and Soatto, 2008; Moore et al., 2008). The proposed ac-SLIC method has the same level of complexity as SLIC.

Sparsity

In standard linear SVM, the weights are a vector of coefficients, with mostly nonzero entries. Such a dense result indicates that almost all features were found to aid in the predictive power of the classifier. However, as mentioned in (Tohka et al., 2016; Bron et al., 2015), the use of all voxelwise brain features may be suboptimal, and a feature selection step or the use of a sparsity-inducing norm can both improve the classification accuracy as well as generate clinically more meaningful results. In this paper, we integrate a sparsity-inducing norm into the linear SVM cost function, using either lasso or group lasso penalties.



(a) (b) (c)

Fig. 2. Comparison of the SLIC and ac-SLIC segmentation results. a) the brain segmentation S; b) the proposed ac-SLIC supervoxel segmentation; c) the SLIC supervoxel segmentation.

Lasso

Since its introduction in (Tibshirani, 1996), the lasso penalty term is widely used to select the important factors in a high dimensional space. Besides its original use in linear regression, it is increasingly used for classification. For example, the lasso penalty is used in linear SVM and implemented in the liblinear toolbox (Fan et al., 2008a). Intuitively, we want to minimize the number of selected variables, which can be represented by the sum of non-zero terms, i.e. $w_0 = \sum_j 1_{(x(j)\neq 0)}$, using the ℓ_0 norm. However, this will lead to an NP-hard problem, which is not solvable for high dimensions. To approximate it, the lasso method uses the ℓ_1 norm to replace the ℓ_0 norm. The lasso SVM used in liblinear is modeled as:

$$\{\mathbf{w}^{\text{opt}}, b^{\text{opt}}\} = \underset{\mathbf{w}\in\mathbb{R}^D, b\in\mathbb{R}}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n} \mathscr{L}_{\text{hinge}}(y_i(<\boldsymbol{w}, \boldsymbol{x}_i> + b))^2 + \lambda \parallel \boldsymbol{w}\parallel_1, \tag{5}$$

where $\boldsymbol{w}_1$ replaces the max-margin term $\boldsymbol{w}_2^2$. To solve this lasso term minimization, a soft thresholding is used for each element of $\boldsymbol{w}$.

Group lasso

A weakness of the lasso SVM is that each variable is treated independently. Therefore, it fails to select groups of strongly correlated variables, and the selection of features is known to be unstable (Tološi and Lengauer, 2011). To overcome these limitations, we exploit the spatial-anatomical information available from the ac-SLIC segmentation. Neighboring voxels from the same subregion are then considered a group, which are used in a group lasso penalty term:

$$\text{GL}(\boldsymbol{w}) = \sum_{g=1}^{G} \beta_g \left\|\boldsymbol{w}_g\right\|_2, \tag{6}$$

where β_g is a scaling factor that compensates for size differences among groups and $\boldsymbol{w}_g$ is the coefficients subvector for group g. Note that we have used the ℓ_2 norm in Eq. (6) to avoid sparsity within a group, and moreover to avoid reducing GL to a standard lasso approach when the ℓ_1 norm would be used. The group lasso term then replaces the sparsity term in Eq. (5). Soft thresholding is now performed on the group level instead of on the feature level. This will result in the selection of a number of predictive groups g, instead of the spurious selection of isolated voxels when using the standard lasso approach.

Spatial-anatomical regularization (SAR)

Our aim is to regularize the weight map, such that two points that are spatial neighbors and additionally belong to the same anatomical region should have similar weights. Mathematically, we define a Spatial-Anatomical Regularization term as $\text{SAR} = \sum_{j,j'}(f(j,j')(w(j) - w(j')))^2$, with $f()$ an indicator function encoding which point pairs require regularization:

$$f(j,j') = \begin{cases} 1 & \text{if } ||j - j'||_2 \leq d \quad \text{and} \quad S(j) = S(j'), \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

where d controls neighborhood extent. Note that the term $S(j) = S(j')$ introduces the anatomical information. In the experiments we compare with a variant that only considers spatial regularization (SR), where the condition $S(j) = S(j')$ is removed from the function $f()$. We encode the relations defined by $f()$ in a matrix $\boldsymbol{L}$. As most point pairs are no neighbors, $\boldsymbol{L}$ is a sparse matrix. We rewrite the SAR as:

$$\text{SAR} = \sum_{c=1}^{C} \langle \boldsymbol{L}(c), \boldsymbol{w}\rangle^2 = (\boldsymbol{L}\boldsymbol{w})'(\boldsymbol{L}\boldsymbol{w}) = \boldsymbol{w}'(\boldsymbol{L}'\boldsymbol{L})\boldsymbol{w}, \tag{8}$$

where C is the total number of point pairs.

It can be seen that the SAR has a quadratic form ($\boldsymbol{w}'\boldsymbol{L}'\boldsymbol{L}\boldsymbol{w}$), similar to the max-margin penalty term ($\boldsymbol{w}'\boldsymbol{I}\boldsymbol{w}$, with $\boldsymbol{I}$ the identity matrix). In addition, both $\boldsymbol{L}'\boldsymbol{L}$ and $\boldsymbol{I}$ are semi-positive definite. When both SAR and max-margin are used, they can be easily merged in a single quadratic regularizer $\mathscr{R}$:

$$\mathscr{R}(\boldsymbol{w}, \boldsymbol{L}'\boldsymbol{L}, \lambda_1, \lambda_2) = \lambda_1 \boldsymbol{w}_2^2 + \lambda_2 \boldsymbol{w}'(\boldsymbol{L}'\boldsymbol{L})\boldsymbol{w} = \boldsymbol{w}'(\lambda_1 + \lambda_2 \boldsymbol{L}'\boldsymbol{L})\boldsymbol{w}. \tag{9}$$

The derivative of this new regularization term is $\frac{d\mathscr{R}}{d\boldsymbol{w}} = 2(\lambda_1 + \lambda_2 \boldsymbol{L}'\boldsymbol{L})\boldsymbol{w}$. This approach can easily be extended with new quadratic regularization terms.

Cost function and optimization

Now that we introduced sparsity-inducing norms and spatial-anatomical regularization, we need to integrate these terms into a unified cost function. When using lasso as a sparsity term to encourage feature level sparsity, the overall cost function becomes:

$$\mathscr{C}_{\text{lasso}} = \frac{1}{n}\sum_{i=1}^{n} \mathscr{L}_{\text{hinge}}(y_i(<\boldsymbol{w}, \boldsymbol{x}_i> + b))^2 + \frac{1}{2}\lambda_1 \parallel \boldsymbol{w}\parallel_2^2 + \frac{1}{2}\lambda_2 \boldsymbol{w}'(\boldsymbol{L}'\boldsymbol{L})\boldsymbol{w} + \lambda_3 ||\boldsymbol{w}||_1, \tag{10}$$

where λ_1, λ_2 and λ_3 are non-negative parameters to control the contribution of each term to the cost function.

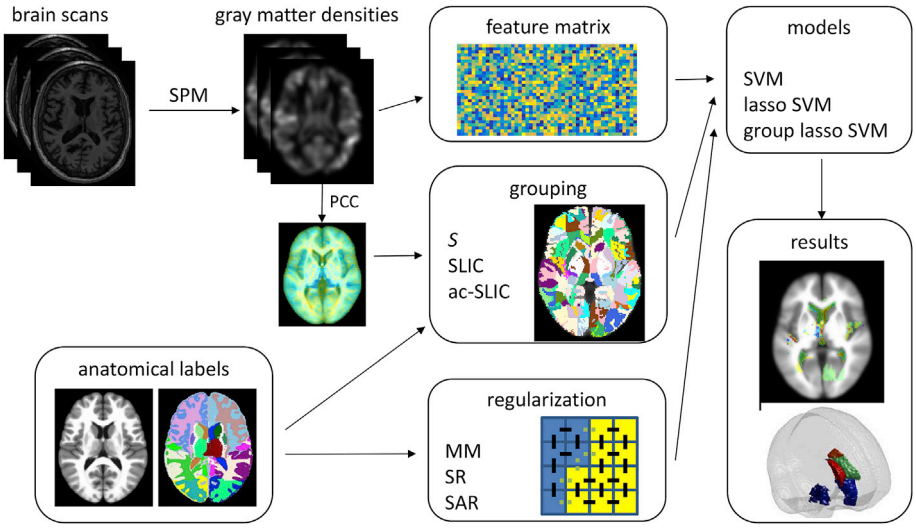


Fig. 3. Workflow of the proposed methods.

When using the group lasso term to encourage group level sparsity, the overall cost function becomes:

$$\mathscr{C}_{\text{grouplasso}} = \frac{1}{n}\sum_{i=1}^{n}\mathscr{L}_{\text{hinge}}(y_i(<\boldsymbol{w},\boldsymbol{x}_i>+b))^2 + \frac{1}{2}\lambda_1 \| \boldsymbol{w}\|_2^2 + \frac{1}{2}\lambda_2 \boldsymbol{w}'(\boldsymbol{L}'\boldsymbol{L})\boldsymbol{w} + \lambda_3\sum_{g=1}^{G}\beta_g\|\boldsymbol{w}_g\|_2. \tag{11}$$

Analyzing Eqs. (10) and (11) we observe that the hinge loss, the max-margin penalty and the SAR are all convex and differentiable. The lasso and group lasso sparsity terms are however convex and not differentiable. We therefore cannot use standard gradient descent to solve the minimization problem.

Instead, we use the proximal gradient descent method, which alternatively optimizes the convex part by gradient descent and the non-convex part by the proximal method (Combettes and Pesquet, 2011), to solve the cost minimization problem. In this work, we choose to use the Fast Iterative Soft Threshold Algorithm (FISTA) (Beck and Teboulle, 2009), due to its efficiency. In (Beck and Teboulle, 2009), the authors proved that by adding a momentum term, FISTA has a better convergence rate $\mathscr{O}(1/k^2)$ compared to the convergence rate $\mathscr{O}(1/k)$ of its predecessor Iterative Soft Threshold Algorithm (ISTA) (Bredies and Lorenz, 2008). In each iteration, we first compute the gradient of the differentiable terms, after which an update is computed using the proximal operator. The derivatives to $\boldsymbol{w}$ and b for the differentiable terms are:

$$\nabla_{\boldsymbol{w}} = \sum_{i=1}^{n}\frac{\partial\mathscr{L}_{\text{hinge}}(\boldsymbol{x}_i,y_i,\boldsymbol{w},b)}{\partial\boldsymbol{w}} + \lambda_1\boldsymbol{w} + \lambda_2(\boldsymbol{L}'\boldsymbol{L})\boldsymbol{w}, \tag{12}$$

$$\nabla_b = \sum_{i=1}^{n}\frac{\partial\mathscr{L}_{\text{hinge}}(\boldsymbol{x}_i,y_i,\boldsymbol{w},b)}{\partial b}. \tag{13}$$

The gradient of the hinge loss is computed as:

$$\frac{\partial\mathscr{L}_{\text{hinge}}}{\partial\boldsymbol{w}} = \begin{cases} -y_i\boldsymbol{x}_i & \text{if } y_i(\langle\boldsymbol{x}_i,\boldsymbol{w}\rangle+b)<1 \\ 0 & \text{otherwise}\end{cases} \tag{14}$$

$$\frac{\partial\mathscr{L}_{\text{hinge}}}{\partial b} = \begin{cases} -y_i & \text{if } y_i(\langle\boldsymbol{x}_i,\boldsymbol{w}\rangle+b)<1 \\ 0 & \text{otherwise}\end{cases} \tag{15}$$

After computing the gradient of the differentiable terms, we update $\{\boldsymbol{w},b\}$ using gradient descent $\boldsymbol{w}=\boldsymbol{w}-\alpha\nabla_{\boldsymbol{w}}$ and $b=b-\alpha\nabla_b$. Then the optimal solution is computed using the proximal operator to minimize a proximal cost: $\min_{\boldsymbol{u}}\lambda_3||\boldsymbol{w}||_1+\frac{1}{2\alpha}\|\boldsymbol{u}-\boldsymbol{w}\|_2^2$ for lasso, and $\min_{\boldsymbol{u}}\lambda_3\sum_{g=1}^{G}\beta_g\|\boldsymbol{w}_g\|_2+\frac{1}{2\alpha}||\boldsymbol{u}-\boldsymbol{w}||_2^2$ for group lasso. The proximal operators are defined as follows:

$$\mathscr{P}_{\text{lasso}} = \begin{cases} (||w_m||_1-\lambda_3\alpha)\text{sign}(w_m) & \text{if } ||w_m||_1\geq\lambda_3\alpha \\ 0 & \text{otherwise}\end{cases} \tag{16}$$

for each element of $\boldsymbol{w}$ in lasso, and

$$\mathscr{P}_{\text{group lasso}} = \begin{cases} \left(1-\frac{\lambda_3\beta_g\alpha}{||\boldsymbol{w}_g||_2}\right)\boldsymbol{w}_g, & \text{if } \|\boldsymbol{w}_g\|_2\geq\lambda_3\beta_g\alpha \\ 0 & \text{otherwise}\end{cases} \tag{17}$$

for each group in group lasso. The overall FISTA optimization is summarized in Algorithm 1.

Table 1
Model variation under the proposed framework. SR and SAR stand for spatial regularization and spatial-anatomical regularization, respectively. Note that for the group lasso models we need to choose an anatomical grouping, which is one of *S*, SLIC or ac-SLIC.

Model name	Max-margin	Regularizer	Sparsity	Similar or equivalent to
linear SVM	ON	OFF	OFF	(Klöppel et al., 2008)
+ SR	OFF	SR	OFF	(Cuingnet et al., 2013)
+ SAR	OFF	SAR	OFF	
lasso SVM	OFF	OFF	lasso	
+ MM	ON	OFF	lasso	(Wu et al., 2017), (Wachinger et al., 2016), (Kohannim et al., 2012), (Shen et al., 2011)
+ SR	OFF	SR	lasso	(Grosenick et al., 2013)
+ SAR	OFF	SAR	lasso	
group lasso SVM	OFF	OFF	group lasso	
+ MM	ON	OFF	group lasso	
+ SR	OFF	SR	group lasso	
+ SAR	OFF	SAR	group lasso	

Summary

An overview of the proposed method is given in Fig. 3. To summarize, brain structural information S is derived from segmentation of the MNI atlas and defined in the SPM template space. A gray matter density feature GM is derived from the brain scans by SPM 8. Based on the Pearson correlation of GM, the segmentation S is refined into a supervoxel segmentation by the proposed ac-SLIC method. Spatial-Anatomical Regularization (SAR) is derived directly from the segmentation S. By optionally combining the feature GM with SAR and the supervoxel segmentation, different SVM-based models are constructed. The list of possible variations is given in Table 1. In this table, we also indicate if some of the variants are similar or equivalent to previous methods. A more detailed comparison can be found in the discussion, Section 6.1. In the remainder of this paper, we adhere to the naming conventions given in this table.

Algorithm 1. FISTA optimization for the proposed methods.

Require: feature $\boldsymbol{x}_i$, subject label y_i, spatial-anatomical matrix $L'L$, initial weight vector $\boldsymbol{w}_0$, non-negative parameters $\lambda_1, \lambda_2, \lambda_3, t_1 = 1, k = 1$

while not converged **do**
  // Gradient descent
  $\boldsymbol{w}_{k+1} \leftarrow \boldsymbol{w}_k - \alpha\nabla_{\boldsymbol{w}}$ ▷ update $\boldsymbol{w}$, use Eq. (12)
  $b_{k+1} \leftarrow b_k - \alpha\nabla_b$ ▷ update b, use Eq. (13)
  if $\lambda_3 > 0$ **then**
    // Proximal operator
    if use lasso **then**
      $\boldsymbol{w}_{k+1} \leftarrow \mathcal{P}_{\text{lasso}}(\boldsymbol{w}_{k+1})$ ▷ use Eq. (16)
    end if
    if use group lasso **then**
      $\boldsymbol{w}_{k+1} \leftarrow \mathcal{P}_{\text{group lasso}}(\boldsymbol{w}_{k+1})$ ▷ use Eq. (17)
    end if
  end if
  // Update
  $t_{k+1} \leftarrow \frac{1}{2}\left(1 + \sqrt{1 + 4t_k^2}\right)$
  $\boldsymbol{w}_{k+1} \leftarrow \boldsymbol{w}_{k+1} + \frac{t_k - 1}{t_{k+1}}(\boldsymbol{w}_{k+1} - \boldsymbol{w}_k)$
end while
Return $\boldsymbol{w}$ and b

Data and implementation details

ADNI brain data

All data used in this paper was obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset (http://adni.loni.usc.edu). The ADNI was launched in 2003 by the National Institute on Aging (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the Food and Drug Administration (FDA), private pharmaceutical companies and non-profit organizations, as a $60 million, 5-year public-private partnership. The primary goal of ADNI has been to test whether serial MRI, PET, other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of MCI and early AD. Determination of sensitive and specific markers of very early AD progression is intended to aid researchers and clinicians to develop new treatments and monitor their effectiveness, as well as lessen the time and cost of clinical trials.

The principal investigator of this initiative is Michael W. Weiner, MD, VA Medical Center and University of California San Francisco. ADNI is the result of efforts of many co-investigators from a broad range of academic institutions and private corporations, and subjects have been recruited from over 50 sites across the U.S. and Canada. The initial goal of ADNI was to recruit 800 subjects but ADNI has been followed by ADNI-GO and ADNI-2.

Participants

The MR images used in this study are from the same population as used in (Cuingnet et al., 2011) and (Cuingnet et al., 2013). In these studies, 509 subjects were selected from ADNI, including 162 cognitively normal (CN) subjects, 137 subjects with AD, 76 subjects with MCI who had converted to AD within 18 months (MCIc) and 134 subjects with MCI who remained stable (MCIs). To obtain an unbiased estimation, we use the same splitting between the training and testing set as used in (Cuingnet et al., 2011). This splitting preserves the age and sex distribution. Detailed demographic characteristics of the selected subjects and the training-testing division are presented in Table 2.

To test the proposed method in a large scale dataset, we additionally included AD and CN subjects from the publicly available "ADNI1: Complete 2Yr 1.5 T" dataset, which contains 346 AD and 575 CN subjects. After removing the subjects that are also in Cuingnet's set, an independent test set is obtained consisting of 508 CN subjects and 311 AD subjects (67 CN and 35 AD excluded). This set is used strictly as a test set, meaning that inference is performed using the model trained on Cuingnet's training set. Similar to Cuingnet's dataset, detailed demographic characteristics of this dataset are presented in Table 3.

Table 3
Demographic characteristics of the ADNI1:Complete 2Yr 1.5 T dataset, after removing the overlap with Cuingnet's dataset summarized in Table 2. MMSE stands for Mini-Mental State Examination.

Diagnosis	Number	Age	Gender	MMSE
CN	508	77.1 ± 5.1 [61–92]	263M/245F	29.1 ± 1.1 [24–30]
AD	311	75.9 ± 7.6 [57–91]	163M/148F	20.9 ± 4.7 [2–29]

Table 2
Demographic characteristics of the ADNI subset defined by (Cuingnet et al., 2011). MMSE stands for Mini-Mental State Examination.

Group	Diagnosis	Number	Age	Gender	MMSE
Complete set	CN	162	76.3 ± 5.4 [60–90]	76M/86F	29.2 ± 1.0 [25–30]
	AD	137	76.0 ± 7.3 [55–91]	67M/70F	23.2 ± 2.0 [18–27]
	MCIc	76	74.8 ± 7.4 [55–88]	43M/33F	26.5 ± 1.9 [23–30]
	MCIs	134	74.5 ± 7.2 [58–91]	84M/50F	27.2 ± 1.7 [24–30]
Training set	CN	81	76.1 ± 5.6 [60–89]	38M/43F	29.2 ± 1.0 [25–30]
	AD	69	75.8 ± 7.5 [55–89]	34M/35F	23.3 ± 1.9 [18–26]
	MCIc	39	74.7 ± 7.8 [55–88]	22M/17F	26.0 ± 1.8 [23–30]
	MCIs	67	74.3 ± 7.3 [58–87]	42M/25F	27.1 ± 1.8 [24–30]
Test set	CN	81	76.5 ± 5.2 [63–90]	38M/43F	29.2 ± 0.9 [26–30]
	AD	68	76.2 ± 7.2 [57–91]	33M/35F	23.2 ± 2.1 [20–27]
	MCIc	37	74.9 ± 7.0 [57–87]	21M/16F	26.9 ± 1.8 [24–30]
	MCIs	67	74.7 ± 7.3 [58–88]	42M/25F	27.3 ± 1.7 [24–30]

MRI acquisition

The MRI images are T1-weighted 1.5*T* MR images. The MRI acquisition had been done according to the ADNI acquisition protocol (Jack et al., 2008). To further enhance the standardization across different clinical sites, all images have undergone same post acquisition correction (3D gradwarp correction, B1 non-uniformity correction, and N3 bias field correction) to remove imaging artifacts.

Implementation details and settings

For computation of the gray matter feature we use the SPM 8 toolbox (http://www.fil.ion.ucl.ac.uk/spm/), with the default parameters. Only voxels inside the brain are used for feature construction. We employ elastix (Klein et al., 2010) to register the brain segmentation to the template space. The Matlab code of all models listed in Table 1 is made publicly available via GitHub (https://github.com/ZhuoSun1987/GroupLassoSVM_SAR.git).

For the spatial (-anatomical) regularization (SR and SAR), we restrict the neighborhood to 26 neighbors ($d = \sqrt{3}$), based on which $\boldsymbol{L}'\boldsymbol{L}$ is computed. The segmentation S contains 35 brain regions, according to the MINC pipeline. For the group lasso model we need to select a grouping strategy. In the experiments we compare the SPM template segmentation S, with the supervoxel-based groupings SLIC and ac-SLIC. The grid size for SLIC and ac-SLIC is set to $10 \times 10 \times 10$, based on visual inspection of the generated supervoxels. The hyper-parameter η balancing the spatial and content-based distances is set to $\eta = 1$, according to Achanta et al. (2012). When not mentioned, the ac-SLIC method is used to provide grouping information.

In our experiments, SPM takes around 10 min per scan to compute the gray matter features. It takes around 5 min to compute the supervoxel map using ac-SLIC for the 3D template brain, and around 2 min to generate the SAR matrix (both only required once). For a given set of parameters $\lambda_1, \lambda_2, \lambda_3$, it takes around 20 min to train an SVM model when using the group lasso sparsity term and less than 10 min when using the lasso sparsity term. After offline model training, the online inference phase takes 10 min per scan to compute the gray matter features and less than a second to apply the trained model.

Experiments and results

We first compare the influence of the separate model terms on a synthetic 2D dataset, see Section 5.1. Then, in Section 5.2, we show the influence of model choices on the resulting weight maps based on Cuingnet's AD vs. CN dataset. In Section 5.3, we explore the influence of the ac-SLIC grid size parameter and the hyper-parameters λ_2 and λ_3, on the proposed group lasso SVM + SAR model using Cuingnet's AD vs. CN dataset. In Section 5.4 we investigate the stability of the generated feature maps on the same dataset, and find the main regions involved in the AD vs. CN task. In Section 5.5, we report the classification performance of the several models, on distinguishing AD from CN using Cuingnet's dataset. To test the performance and generalizability of the proposed method, we apply the model learned from Cuingnet's AD vs. CN training set directly to the ADNI1 AD vs. CN data summarized in Table 3 and report the performance in Section 5.6. Finally, we test the proposed method on three harder tasks (CN vs. MCI, MCIc vs. MCIs, and MCI vs. AD) as described in Table 2 to distinguish the different clinical groups in Section 5.7.

Model comparison on 2D synthetic data

In order to highlight the differences between the several models derived from the proposed framework, we create a synthetic experiment. We generated two baseline images mimicking two population means, as shown in Fig. 4a and b. The two images are structurally the same, except for the small green square. For each population 20 realizations are generated by adding Gaussian noise to the population, mean, as shown in Fig. 4c for population B. The resulting images act as the input features for the classifier. Fig. 4a and b also act as the anatomical segmentation and thus provide grouping information (grouping A and B).

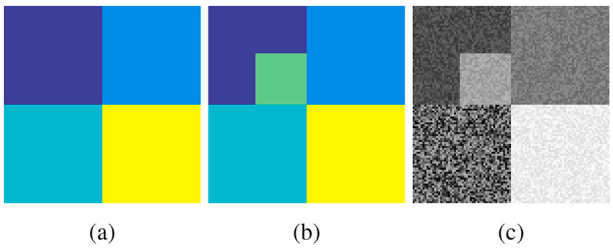


Fig. 4. Synthetic 2D data. (a) Mean of population A; (b) mean of population B; (c) simulated subject from population B, after adding Gaussian noise to the population mean.

Then we train the following models and display the resulting normalized weight map in Fig. 5. For Fig. 5a we use the linear SVM model with $\lambda_1 = 1.0$; For Fig. 5b and c we use the SVM + SR or SAR with $\lambda_2 = 1.0$; For Fig. 5d–f we use the group lasso SVM + SAR with $\lambda_2 = 1.0$ and $\lambda_3 = 0.01, 0.1$ and 0.1 respectively.

From all figures, we can see that the small green structure is highlighted, and was therefore found to be important in distinguishing population A from B, by all methods. The linear SVM (Fig. 5a) returns a noisy weight map, making the results harder to interpret. When comparing Fig. 5b and c, we can see that SAR indeed avoids smoothing across anatomical boundaries (edges are visible between the big squares), while SR tends to over smooth it. For the group lasso models (Fig. 5d–f), it is possible to select the true distinctive regions. For Fig. 5d and e, we used segmentation A as grouping information, and indeed the top left square was found to be distinctive. For Fig. 5f we used segmentation B, and the small square was correctly found to be distinctive. Therefore, a proper choice of grouping information is beneficial for selectively locating distinctive regions.

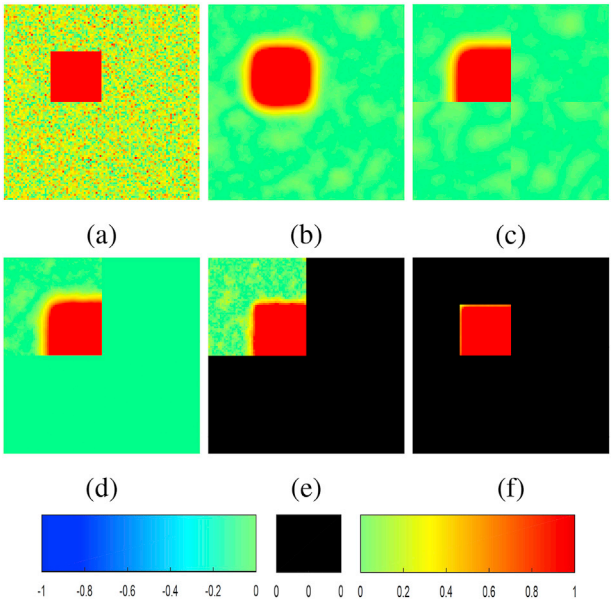


Fig. 5. Normalized weight maps of the various models for the 2D synthetic experiment. The black color indicates weights that are exactly zero. (a) linear SVM; (b) SVM + SR; (c) SVM + SAR; (d) group lasso SVM + SAR using grouping A and low sparsity weight; (e) group lasso SVM + SAR using grouping A and high sparsity weight; (f) group lasso SVM + SAR using grouping B and high sparsity weight.

Qualitative model comparison on ADNI data

Inside the linear SVM framework, the optimal weight map $\mathbf{w}^{opt}$ enables visualization of the amount of involvement of the several brain regions. In this section we qualitatively compare the weight maps of the several models. All visualized weight maps are z-score normalized.

The effect of sparsity

As mentioned in the introduction, the sparsity term may lead to a better spatial localization of the learned model. In this section we compare linear SVM ($\lambda_1 = 1$) with lasso SVM ($\lambda_3 = 0.2$) and group lasso SVM (using ac-SLIC and $\lambda_3 = 5$). Normalized weight maps for these models are given in Fig. 6. From this figure, we can see that different sparsity patterns are obtained. In linear SVM, which has no sparsity penalty, all voxels have non-zero weight. In lasso SVM, only a few voxels are selected and they are scattered over the brain. Although this weight map shows the brain locations that impact classification the most, these locations are quite isolated and not well structured, making it hard to identify which brain regions are actually involved in AD. In group lasso SVM larger connected neighborhoods at supervoxel size, are selected, that show more structure. In addition, anatomical boundaries are preserved. In Fig. 6c, it can be seen that the hippocampus and the frontal part of the ventricle are very important for classification of AD.

The effect of regularization on linear SVM

In this section we show the influence of the several quadratic regularization terms (MM, SR and SAR) on the weight map. We set the relevant parameters to the same number, i.e. $\lambda_1 = 10$ for linear SVM and $\lambda_2 = 10$ for SVM + SR and also for SVM + SAR. The learned optimal weight maps $\mathbf{w}^{opt}$ are shown in Fig. 7.

From these weight maps, we can see that both SR and SAR regularization will lead to smoother weight maps than linear SVM (with MM). The blue and red colors indicate regions that are closely related to the classification of AD. Comparing SR (Fig. 7b) with SAR (Fig. 7c), it is clear that with SAR, the weight map can avoid over-smoothing across anatomical boundaries (Fig. 7d). For example, compare the hippocampus region and the corpus callosum region. To better illustrate this effect, we increased λ_2 to 100. Fig. 8 shows a clear difference on the boundaries of several anatomical regions.

The effect of regularization on lasso SVM

To better understand the relation between spatial regularization and sparsity, we plot the weight maps of lasso SVM, lasso SVM + MM, lasso SVM + SR, and lasso SVM + SAR, for different settings of $\lambda_3 \in \{0.2, 0.1, 0.05\}$. The results are shown in Fig. 9. It is clear that introducing spatial regularization, either SR or SAR, leads to a more clustered feature selection. There is little visual difference between lasso SVM and lasso SVM + MM, and also between the use of SR or SAR.

The effect of regularization on group lasso SVM

From the previous results, it is clear that spatial regularization can lead to a more clustered feature selection, which simplifies localization. Here we illustrate the effect of regularization on the SVM models with group level sparsity. We set $\lambda_2 = 10$ and $\lambda_3 = 5$ for group lasso SVM + SR and for group lasso SVM + SAR. We omit group lasso SVM + MM here, as little visual difference with group lasso SVM without spatial regularization was observed. The resulting weight maps are shown in Fig. 10. We can see that both group lasso SVM + SR and group lasso SVM + SAR are smoother and more clustered than group lasso SVM, for example the ventricle region. Comparing SR and SAR, SAR is somewhat smoother within regions and obeys anatomical boundaries; see the red box for example. Compared to the lasso SVM models, see Fig. 9, much more (anatomical) structure can be observed.

Optimization and influence of hyper-parameters

The proposed model (group lasso SVM + SAR) is directly influenced by the hyper-parameters λ_2 and λ_3. As there is no analytical gradient

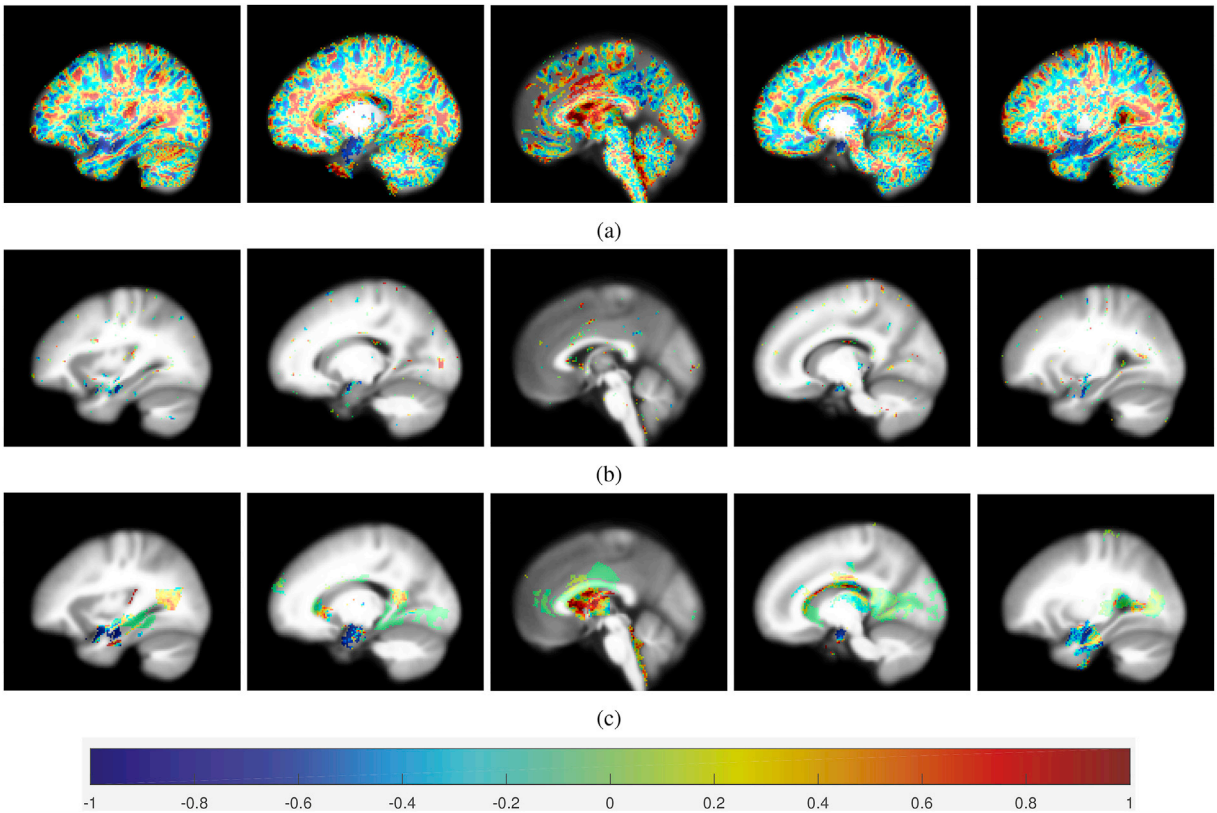


Fig. 6. Normalized weight map $\mathbf{w}^{opt}$ of (a) linear SVM; (b) lasso SVM; and (c) group lasso SVM using ac-SLIC. From left to right different slices are shown.

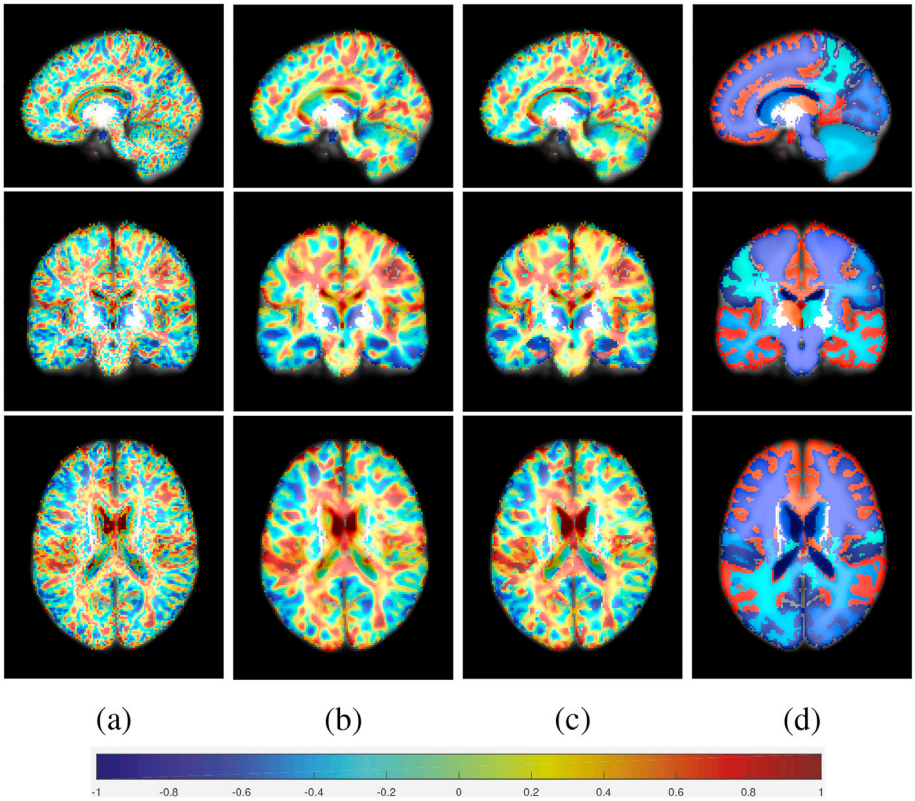


Fig. 7. Normalized weight map $\mathbf{w}^{\text{opt}}$ of (a) linear SVM; (b) SVM + SR; (c) SVM + SAR; and (d) the brain structure segmentation S.

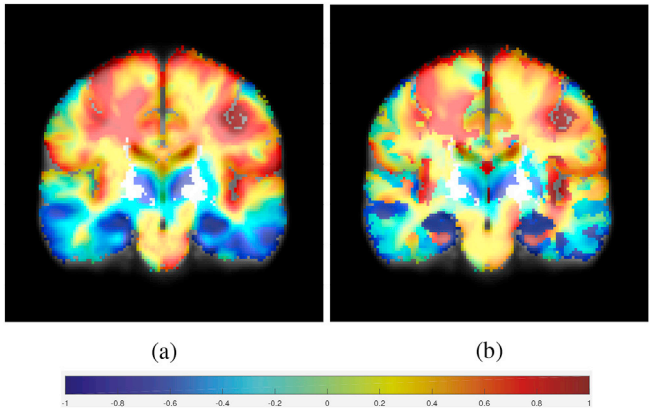


Fig. 8. Normalized weight map $\mathbf{w}^{\text{opt}}$ of (a) SVM + SR; and (b) SVM + SAR, now using $\lambda_2 = 100$.

available for these hyper-parameters, we use a grid search approach to explore the influence of these parameters. Both λ_2 and λ_3 are optimized from $\{0.00001, 0.0001, 0.001, 0.1, 0.2, 0.5, 1, 2, 5, 10, 100\}$. They are optimized for Cuingnet's AD vs. CN dataset, using 5-fold cross-validation on only the training data. The optimal results on the test data, for several grid sizes, are reported in Table 4. It can be seen that the a priori chosen grid size of $10 \times 10 \times 10$, indeed yielded best performance. For this grid size, the influence of λ_2 and λ_3 on the classification accuracy is shown in Fig. 11a. Fig. 11b shows the influence of λ_2 and λ_3 on the number of selected supervoxel regions. For the other models from Table 1 we performed similar grid searches, resulting in optimized hyper-parameters. In the remainder of the paper, these optimized parameters are used unless mentioned otherwise.

Feature map stability

When retraining the models over a (slightly) different training set, the resulting weight map changes accordingly. For the models that encourage sparsity, different brain locations may be identified as contributing to explaining AD. This is a known phenomenon (Dunne et al., 2002; Xin et al., 2016). These unstable effects may lead to differences in interpretation of the results, depending on the training set.

To measure the variation in the selection of important brain locations, we propose the following procedure. We repeat model training K times, each with a slightly different training set. Then for each voxel $\boldsymbol{x}$ we count the fraction of times $n(\boldsymbol{x})$ it was selected. From this fraction, we count the total number of voxels in the brain that were selected at least p percent of the time. The latter is normalized by the total number of voxels that were selected at least once (activated voxels). So, we define $n(x) = \frac{1}{K}\sum_k 1_{\mathbf{w}_k^{\text{opt}}(\boldsymbol{x})\neq 0}$. Then, we come to the following definition of stability at a level p:

$$\mathbb{S}(p) = \frac{\sum_{\boldsymbol{x}} 1_{n(\boldsymbol{x})\geq p}}{\sum_{\boldsymbol{x}} 1_{n(\boldsymbol{x})>0}} \times 100\%. \tag{18}$$

For example, the stability of $\mathbb{S}(50) = 60$ means that 60% of the activated voxels have been selected 50% of the time. By sampling over different stability levels $p \in [0, 100]$, we can make a complete stability profile. In our experiment we selected $K = 100$. For each new training, we randomly selected 75% of the available data to train the model.

We compute the $\mathbb{S}$-curves for lasso SVM + SAR and group lasso SVM + SAR. For both models we set $\lambda_2 = 10$. We use $\lambda_3 = 10$ for group lasso SVM + SAR, and for lasso SVM + SAR we set λ_3 such that the two models have the same amount of activated voxels (same sparsity level). With $\lambda_3 = 0.38$, we have approximately 2.4×10^4 activated voxels for both models. The resulting $\mathbb{S}$-curves are given in Fig. 12, showing that the proposed group lasso model is more stable than the lasso model.

In Fig. 13 we show the supervoxels that were selected at least 95/100

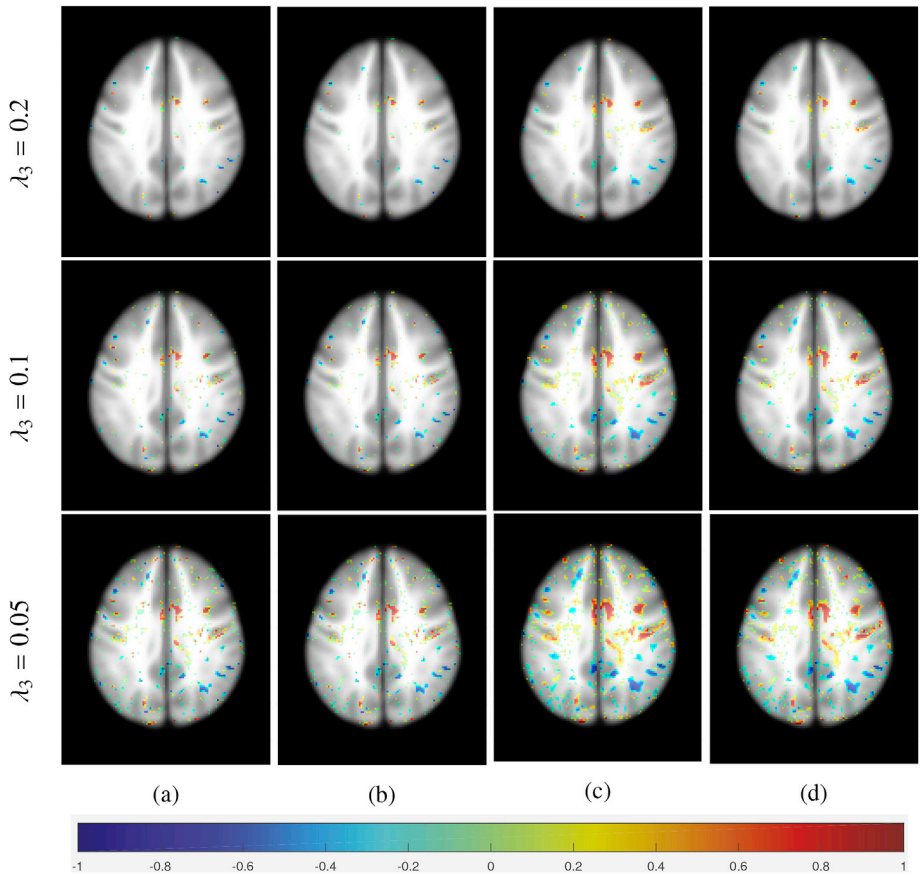


Fig. 9. Normalized weight map $\mathbf{w}^{opt}$ of (a) lasso SVM; (b) lasso SVM + MM; (c) lasso SVM + SR; and (d) lasso SVM + SAR. Top row: $\lambda_3 = 0.2$, middle row: $\lambda_3 = 0.1$, bottom row: $\lambda_3 = 0.05$.

times by the proposed model. The hippocampus and the frontal parts of the ventricles were the most stable regions, important for AD classification. This is in line with current knowledge (Apostolova et al., 2012; Pini et al., 2016). In Fig. 14 we can see that by ac-SLIC only the most relevant part of each anatomical structure is selected.

Classification performance for AD vs CN

A summary is given in Table 5 to compare the classification results from different models. First, for the linear SVM without sparsity, both SR and SAR improve the classification accuracy. For all model categories SAR improves upon SR. Second, comparing the model categories, the group lasso SVM model performs among the best, indicating the usefulness of structure sparsity in neuroimaging applications. Third, the proposed model (group lasso SVM + SAR using ac-SLIC) performs slightly better than the other models. Overall, the results in classification performance are however very similar. The *p*-value of the McNemar test comparing each method to the linear SVM shows that the SAR term and all group lasso models can statistically improve the classification results.

Generalization performance

In this section we test how well the proposed models generalize to an independent dataset, which is much larger than Cuingnet's training set. This resembles a real-life application, where a previously trained model is used online to classify new inputs. We use the pre-trained model from the previous section, trained on Cuingnet's AD vs. CN training set, without modifications, and apply it to the large ADNI1 dataset described in Section 4.1.1. The classification results for this new dataset are summarized in Table 6. From this table, it is clear that the proposed models generalize very well to new data. An accuracy of 92.6% was obtained, and the proposed model still outperformed the baseline model (linear SVM) with statistical significance.

Classification performance for harder tasks

Until now we have trained and tested the proposed models to distinguish AD from CN. In this section, we report the classification performance of the proposed model on three more difficult tasks, i.e., CN vs. MCI, MCI vs. AD and MCIc vs. MCIs. The model is trained and tested on Cuingnet's dataset given in Table 2. In the current experiments, we combine the MCIc and MCIs group into a single MCI group. The hyper-parameters (λ_1, λ_2 and λ_3) are re-optimized on the training data, similar to the procedure described in Section 5.3: 5-fold cross-validation is used again on the training set to find the optimal hyper-parameters, which are subsequently used for re-training on the complete training set.

The performance of the proposed group lasso SVM + SAR model and the baseline linear SVM model is reported in Table 7 for each task. It can be seen that the proposed method leads to an improvement in the classification compared to the linear SVM model.

Discussion

Related work

In recent years, several methods have been proposed that integrate spatial or anatomical information, in order to generate more interpretable models. Klöppel et al. (2008) used an SVM for AD classification, and introduced the idea of visualizing the local contribution to the classification by overlaying the weight map $\boldsymbol{w}$ on the template image. However,

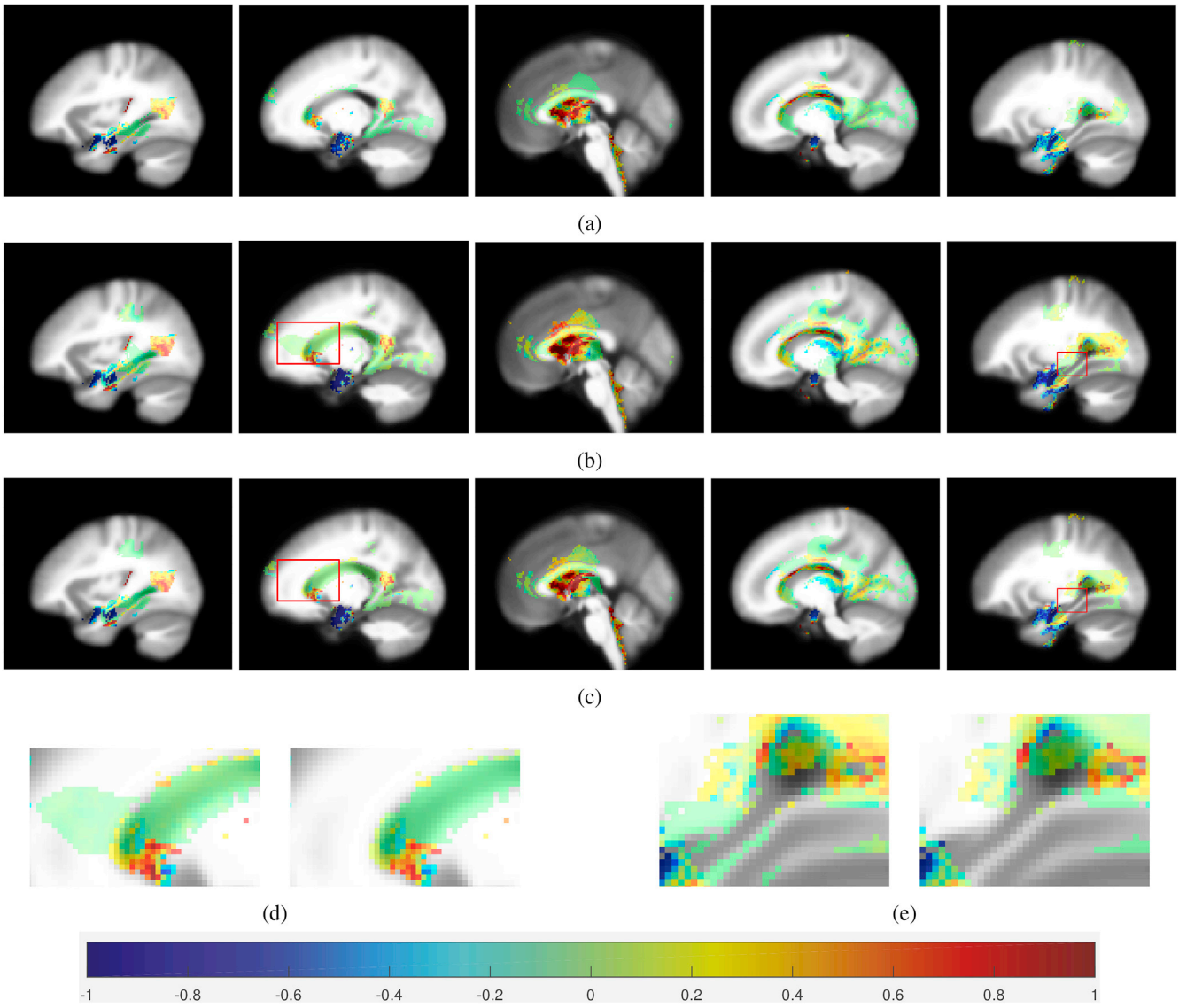


Fig. 10. Normalized weight map $\mathbf{w}^{opt}$ of (a) group lasso SVM; (b) group lasso SVM + SR; and (c) group lasso SVM + SAR, using ac-SLIC to group features. From left to right different slices are shown. The red box in middle and bottom rows shows the difference in the selected regions between group lasso SVM + SR and group lasso SVM + SAR. The zoomed-in version of the second slice (d) and the fifth slice (e) show the differences of the weight maps resulting from the group lasso SVM + SR (left) and group lasso SVM + SAR (right) models.

Table 4
Classification performance of AD vs CN, using Cuingnet's dataset, for different grid sizes for ac-SLIC. The generated supervoxels are used as groups in the proposed group lasso SVM + SAR method.

Grid size	Acc	AUC	SPE	SEN
$1 \times 1 \times 1$	0.879	0.938	0.938	0.809
$5 \times 5 \times 5$	0.872	0.936	0.951	0.779
$10 \times 10 \times 10$	0.893	0.951	0.938	0.838
$15 \times 15 \times 15$	0.852	0.932	0.914	0.779
Image resolution	0.859	0.932	0.926	0.779

they did not consider neighborhood relations and did not identify key involved regions (by sparsity). Comparing to Klöppel et al. (2008), the proposed model uses completely different regularization terms, which can take anatomical information into consideration to encourage both smoothness as well as localization of the weight map.

Fan et al. (2007) introduced the use of supervoxels for schizophrenia classification on brain MRI. These supervoxels span multiple anatomical regions, as they ignored anatomical information similar to SLIC. In addition, features were computed per supervoxel, thus ignoring detailed information available at the voxel level, unlike the proposed method.

Cuingnet et al. (2013) proposed to use two independent terms for spatial and anatomical regularization in a linear SVM. In the computation, these two terms were added together as a large regularization matrix, and the inverse of that regularization matrix was used to transfer the original feature into a new domain. However, the adding operation of the two regularization terms will lead to unavoidable over-smoothing across the anatomical boundary. Also, the inversion of this large regularization matrix may cause computation and memory problems. Although this problem can be solved by a diffusion-like approximation of matrix inversion, it will limit the range of regularization matrix designs. In the presented work, we propose a mathematical reformation of the problem that does not require the inversion of a potentially large regularization matrix and directly optimizes the cost function in the original feature domain. This makes the computation in our paper completely different from Cuingnet et al. (2013). It is exactly this reformulation that enables the proposed method to integrate voxel-level or group-level sparsity in the cost function and offers the possibility to include a wide range of

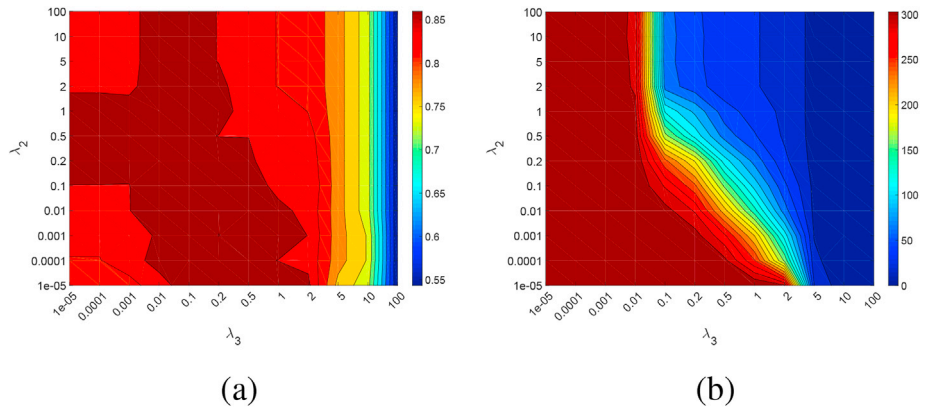


Fig. 11. (a) Classification accuracy and (b) number of selected regions, with respect to the hyper-parameters λ_2 and λ_3 for the group lasso SVM + SAR model.

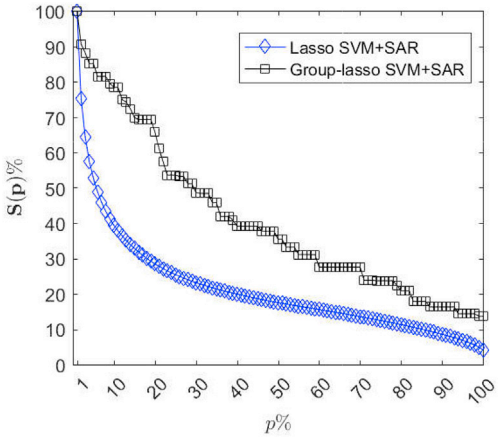


Fig. 12. The $\mathbb{S}$ tability-curves, see Eq. (18), for lasso SVM + SAR and group lasso SVM + SAR.

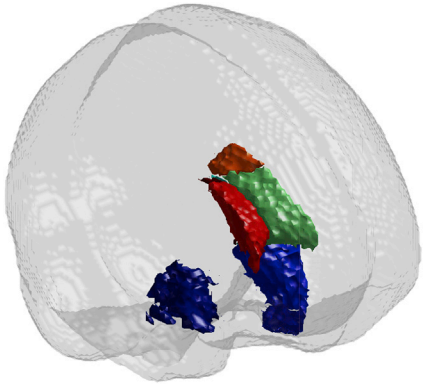


Fig. 13. Supervoxels (ac-SLIC) that were selected 95/100 times by the group lasso SVM + SAR model.

regularization matrix designs, which are difficult to achieve within Cuingnet's computational framework. Also, the proposed method uses a different regularization matrix, which avoids smoothing across anatomical boundaries.

In the field of neuroimaging, the group lasso penalty is typically used for multi-task learning (Zhang et al., 2012a; Liu et al., 2014; Jie et al., 2015), to allow (predefined) groups of covariate features to be turned on or off simultaneously for all tasks. Grouping is thus done on the tasks, whereas in our paper we perform grouping spatially. Compared to our

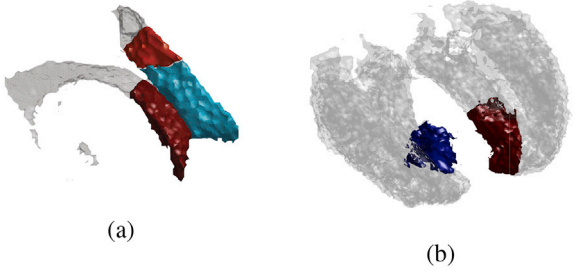


Fig. 14. Stable selection of anatomical sub-regions. (a) The transparent gray surface are the complete ventricles. Only the frontal parts are selected 95% of the time; (b) The transparent gray surface is the complete temporal lobe, as there is no separate label for the hippocampus. The hippocampus is nevertheless selected 95% of the time by the proposed model.

previous method (Sun et al., 2015a), this work performs grouping on supervoxels instead of anatomical labels (S), and we added a new spatial-anatomical regularization term (SAR). This leads to a more smooth and localized model of the brain regions involved in AD.

Although sharing ideas with elastic-net (Wu et al., 2017; Wachinger et al., 2016; Kohannim et al., 2012; Shen et al., 2011) and graph-net (Grosenick et al., 2013), the proposed method has several differences. Compared to both elastic-net and graph-net, the proposed method adds anatomically meaningful grouping. Compared to elastic-net the proposed spatial-anatomical regularization generates a smoother weight map; compared to graph-net our method avoids across-boundary smoothing.

Despite the differences, many of the aforementioned methods fit the proposed framework (Eq. (10) and (11)). The method used in (Klöppel et al., 2008) can be obtained by setting $\lambda_2 = \lambda_3 = 0$. The method used in (Cuingnet et al., 2013) can be obtained by setting $\lambda_1 = \lambda_3 = 0$, and by replacing SAR with two terms representing spatial regularization and anatomical regularization separately. The only remaining difference is the optimization domain, potentially leading to numerical differences. Our previous method (Sun et al., 2015b) can be obtained by setting $\lambda_1 = \lambda_2 = 0$ and using the anatomical segmentation S for grouping. Elastic-net and graph net can be obtained by using lasso as the sparsity term, and setting $\lambda_2 = 0$ to compute the elastic-net, and $\lambda_1 = 0$ with SR for graph-net.

Experimental results

As shown in Section 5.2.1, the different sparsity terms yield different weight maps. The model without sparsity employs only the max margin term, which is a quadratic term on the weights $\boldsymbol{w}$. As its gradient will become small when $\boldsymbol{w}$ is near zero, its strength is relatively small

Table 5
Classification performance for AD vs CN, for the several models. For the grouping we have *S* for anatomical regions without supervoxels, and the two supervoxel methods SLIC and ac-SLIC. † indicates a statistically significant difference with the baseline model (linear SVM), using the McNemar test ($p < 0.05$).

Model	group	Acc	AUC	Spe	Sen
linear SVM		0.846	0.923	0.864	0.765
+ SR		0.852	0.931	0.914	0.779
+ SAR		0.872†	0.949	0.901	**0.838**
lasso SVM		0.859	0.930	0.926	0.779
+ MM		0.866	0.912	0.951	0.765
+ SR		0.832	0.921	0.901	0.750
+ SAR		0.859†	0.932	0.926	0.779
group lasso SVM	ac-SLIC	0.879†	0.943	0.938	0.809
+ MM	ac-SLIC	0.879†	0.943	0.938	0.809
+ SR	ac-SLIC	0.886†	0.944	**0.951**	0.809
+ SAR	ac-SLIC	0.893†	**0.951**	0.938	**0.838**
+ SAR	SLIC	0.879†	0.941	0.926	0.824
+ SAR	*S*	0.879†	0.938	0.938	0.809

compared with the gradient of the hinge loss. Therefore, the weights will likely not become zero exactly, thus not encouraging sparsity. The lasso term is not quadratic, but uses an ℓ_1 norm on the weights. In order to minimize this sum of absolute values, some weights will be forced to be zero, thus inducing sparsity. In group lasso, we use the ℓ_2 norm on each group to avoid encouraging sparsity inside a group, while the ℓ_1 norm is used to encourage sparsity among groups. Comparing the max margin, lasso and group lasso models (see Fig. 6), max margin yields no sparsity, lasso yields very scattered results that are hard to interpret, while the group lasso model incorporates anatomical prior knowledge yielding much improved localization. From Table 5 we can see that the group lasso models perform among the best in terms of classification accuracy, while simultaneously and in particular featuring much improved smoothness and localization.

Comparing the three regularization terms, we note that the max margin term can be considered as assumption-free in the sense that the relation between voxels is ignored. For SR the weak assumption is that neighbors yield similar effect on classification, very comparable to an isotropic smoothing approach. SAR has a stronger assumptions that neighbors from the same anatomical structure yield similar effect on classification, very comparable to an anisotropic smoothing approach. The latter makes sense, however, since indeed different brain structures are differently affected by AD. For example, Pegueroles et al. (2017) reported the different brain structural changes in a longitudinal AD study. As can be seen from Figs. 7, 9 and 10 the proposed group lasso SVM + SAR model indeed results in smooth and sparsely structured weight maps, obeying anatomical boundaries.

When using the group lasso models, we need to specify the grouping information from the three available choices: grouping by anatomical structure (*S*), grouping by supervoxels (SLIC), or grouping by a combination (ac-SLIC). Both SLIC and ac-SLIC divide the image into smaller subregions, and thus have a finer granularity than the atlas *S*. Therefore, these two grouping methods enable more localized selection of involved brain areas. This for example allows pinpointing of the hippocampus from the temporal lobe (*S* only includes the temporal lobe and not the hippocampus), or identifying that the frontal part of some brain structure is more affected by AD than the occipital part. While both SLIC and ac-SLIC have good granularity, ac-SLIC obeys anatomical boundaries (cf Fig. 2). As shown in Fig. 14, ac-SLIC indeed has these desirable properties, leading to anatomically more meaningful localization in the

Table 6
Classification performance for AD vs CN on an independent large scale dataset (ADNI1) using the pre-computed models trained over the training set from Cuingnet et al. (2013). † indicates a statistically significant difference with the baseline model (linear SVM), using the McNemar test ($p < 0.05$).

Model	group	Acc	AUC	Spe	Sen
linear SVM		0.900	0.958	0.898	0.904
+ SR		0.910	0.961	0.909	0.910
+ SAR		0.922†	0.964	0.925	0.916
lasso SVM		0.915†	0.953	0.933	0.884
+ MM		0.902	0.939	0.923	0.868
+ SR		0.907	0.948	0.923	0.881
+ SAR		0.915†	0.953	0.931	0.887
group lasso SVM	ac-SLIC	0.924†	0.959	0.937	0.904
+ MM	ac-SLIC	0.924†	0.959	0.937	0.904
+ SR	ac-SLIC	0.916†	0.959	0.933	0.887
+ SAR	ac-SLIC	0.926†	0.960	0.933	0.913
+ SAR	SLIC	0.928†	0.958	0.939	0.910
+ SAR	*S*	0.904	0.950	0.927	0.865

Table 7
Classification performance for three harder tasks (CN vs MCI, MCI vs AD and MCIs vs MCIc) using the linear SVM model and the proposed group lasso SVM + SAR method. † indicates a statistically significant difference with the baseline model (linear SVM), using the McNemar test ($p < 0.05$).

Application	Model	group	Acc	AUC	Spe	Sen
CN vs MCI	linear SVM	ac-SLIC	0.691	0.793	0.531	0.817
	group lasso SVM + SAR		0.708	0.779	0.691	0.721
MCIs vs MCIc	linear SVM	ac-SLIC	0.615	0.675	0.597	0.649
	group lasso SVM + SAR		0.654	0.683	0.642	0.676
MCI vs AD	linear SVM	ac-SLIC	0.639	0.705	0.673	0.588
	group lasso SVM + SAR		0.657	0.705	0.673	0.632

resulting weight map. From Table 5 ac-SLIC even performs slightly better in terms of classification accuracy than when using S or SLIC for grouping.

Hyper-parameters

In the unified cost function, Eq. (10) for lasso or Eq. (11) for group lasso, there are three parameters λ_1, λ_2, and λ_3 controlling the behavior of the model. The first weight λ_1 relates to the max-margin term and thus to the amplitude of $\boldsymbol{w}$. Therefore, a larger λ_1 encourages the weights $\boldsymbol{w}$ to be closer to $\mathbf{0}$. The second weight λ_2 relates to the Spatial-Anatomical Regularization. A larger λ_2 will lead to a weight map that is smoother inside the pre-defined structures. The third weight λ_3 controls the contribution of the sparsity terms in the cost function. Larger λ_3 will decrease the number of selected regions (in group lasso) or voxels (in lasso), selecting the features that are more important for the classifier. An example of the latter behavior is shown in Fig. 15. Here, we use the proposed group lasso SVM + SAR model, and vary λ_3 between 8 and 10 while $\lambda_2 = 10$. We can indeed see that the number of stable selected supervoxel regions (see Section 5.4) is smaller for larger λ_3. Comparing this figure with Fig. 13 (AD vs. CN), these may indicate that the frontal ventricle parts are discriminative for both MCI vs. AD and AD vs. CN. Furthermore, it seems that the most discriminative regions for MCI vs. AD are mainly in the sub-cortical areas, while for MCIc vs. MCIs these are mainly on the cortical surface.

Both the max-margin term and the sparsity terms will penalize a large amplitude of the weight vector $\boldsymbol{w}$. Since for weights $\boldsymbol{w}$ close to zero the sparsity term (an ℓ_1 norm) has a larger gradient than the max-margin term (an ℓ_2 norm), it has a stronger effect in reducing the amplitude of $\boldsymbol{w}$. Therefore, in the proposed model we chose to omit the max-margin term.

Feature choice

In this paper we have used gray matter density as a feature for Alzheimer's disease classification, since it is arguably the most common feature used for this task. As reported in a recent survey paper (Arbabshirani et al., 2017), other features however may perform better. Focusing on MCIc vs. MCIs classification, Wee et al. (2013) and Li et al. (2012) for example use cortical thickness as a feature and obtained 75.0% respectively 80.3% accuracy, compared to 65.4% for the gray matter density reported in this paper. Fig. 15 confirms that indeed the cortical region was important for MCIc vs. MCIs classification, but in our work found by a completely data driven approach without prior assumptions. Some state-of-the-art methods use other features containing more information, such as multi-modal features (Zhang et al., 2012b; Yu et al., 2014; Moradi et al., 2015; Ota et al., 2015) (accuracy ranges from 67.2% to 78.4%), or multiple type of features (gray matter, white matter, cortical thickness) (Plant et al., 2010; Cui et al., 2011; Klöppel et al., 2015) (accuracies of 75.0%, 67.0% and 73.0%) to improve classification. In our previous paper (Sun et al., 2017), we showed that a feature encoding anatomical change over time is even more powerful than cross-sectional features, with accuracies ranging from 89.0% to 92.0%. All these features can be plugged into the proposed framework.

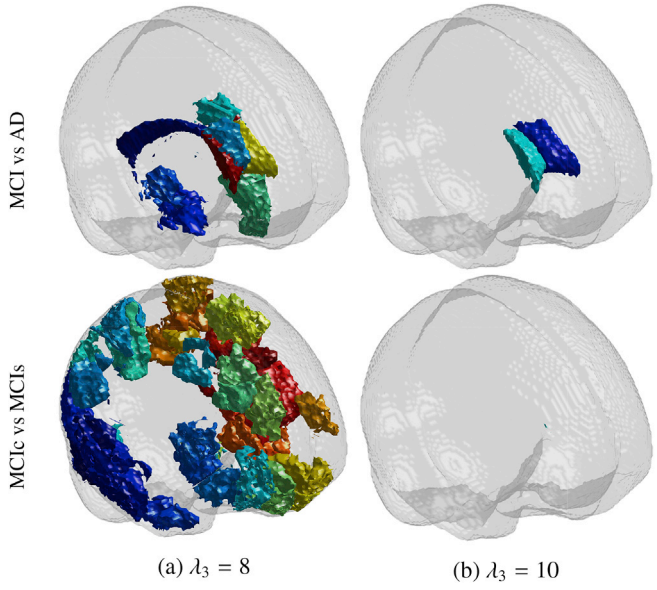


Fig. 15. The stable selected supervoxels for the MCI vs AD task (top row) and the MCIc vs MCIs task (bottom row) using the proposed group lasso SVM + SAR model, using $\lambda_2 = 10$.

Kernelization

The standard linear SVM classifier can be easily represented by a kernel approach, which has potential benefits for the computational complexity. For linear SVM this is done by replacing the classification function with a kernel representation $f(\mathbf{x}') = \sum_{j=1}^{N} a_j y_j \langle \mathbf{x}_j, \mathbf{x}' \rangle + b$ and optimizing the dual variables a_j. In this kernel representation, the kernel function $\mathbf{K}(\mathbf{x}_j, \mathbf{x}') = \langle \mathbf{x}_j, \mathbf{x}' \rangle$. Although most of a_j will be zero, the recovered weights vector $\mathbf{w} = \sum_{j=1}^{N} a_j y_j \mathbf{x}_j$ is still a dense vector and moreover not smoothed.

Xu et al. (2010) and Cuingnet et al. (2013) proposed methods to incorporate sparsity respectively smoothness into the kernel domain representation of SVM. In (Xu et al., 2010), the kernel was rewritten as a weighted sum of inner products of sub-vectors $\mathbf{x}_j^{(g)}$ and $\mathbf{x}'^{(g)}$: $K(\mathbf{x}_j, \mathbf{x}') = \sum_{g=1}^{G} \beta_g \langle \mathbf{x}_j^{(g)}, \mathbf{x}'^{(g)} \rangle$, where β_g is encouraged to be zero through a cost function. The weight vector is then recovered as $\mathbf{w} = \sum_{j=1}^{N} a_j y_j (\mathbf{x}_j \circ \sqrt{\boldsymbol{\beta}})$, where $\circ$ denotes point-wise multiplication. This results in a sparse weight vector due to the sparsity of $\boldsymbol{\beta}$. In (Cuingnet et al., 2013), the authors reformulated the kernel function as $K(\mathbf{x}_j, \mathbf{x}') = \mathbf{x}_j^T \boldsymbol{Q}^{-1} \mathbf{x}'$, where $\boldsymbol{Q}$ is used to enforce smoothing of $\mathbf{w}$ by a term $\mathbf{w}^T Q \mathbf{w}$ in the cost function. However, the recovered weights $\mathbf{w} = \sum_{j=1}^{N} a_j y_j \boldsymbol{Q}^{-1} \mathbf{x}_j$ are not sparse, and cannot easily be made sparse.

Now we show that the proposed group lasso SVM + SAR model may in principle be kernelized as well, where we build on the ideas from Xu et al. (2010) and Cuingnet et al. (2013). We first reorder the features so that features from each anatomical structure are grouped. As our spatial-anatomical regularization (SAR) term does not allow links between groups, the SAR matrix $\boldsymbol{Q}$ is a block diagonal sparse matrix, with sub-blocks $\boldsymbol{Q}^{(g)}$. The kernel for the proposed method can then be written as $K(\mathbf{x}_j, \mathbf{x}') = \sum_{g=1}^{G} \beta_g \mathbf{x}_j^{(g)^T} \boldsymbol{Q}^{(g)^{-1}} \mathbf{x}'^{(g)}$. The optimal parameters α_j and β_g can be computed using (Xu et al., 2010), where both α_j and β_g are sparse. When $\beta_g = 0$, the corresponding $\boldsymbol{w}^{(g)} = \sum_{j=1}^{N} a_j \sqrt{\beta_g} y_j \boldsymbol{Q}^{(g)^{-1}} \mathbf{x}_j^{(g)} = 0$. Otherwise, $\boldsymbol{w}^{(g)}$ is a spatially smoothed weight map inside the anatomical structure g by the regularization via $\boldsymbol{Q}^{(g)}$. This way the proposed method can be represented in the kernel domain as well.

Comparison to forward activation patterns

Current multi-variable pattern analysis methods in neuroimaging give a new possibility in understanding brain development (or neuroactivation patterns) by visualizing the weight map. However, as pointed out by (Haufe et al., 2014), the direct interpretation of weight maps derived from non-regularized discriminative models, should be performed with caution. To improve the interpretability, Haufe et al. (2014) proposed to use the covariance matrix of the inputs to modulate the weight map resulting from the discriminative model. In their experiments on EEG and fMRI data, the relationship between variables,

which is encoded in the covariance matrix, helps to generate a smoother and more interpretable weight map.

In our approach, we aim to solve a similar problem in the context of high-resolution structural MR images. In the proposed model, the (spatial) relation between the variables are encoded by the SAR term in the cost function. This also modulates the learned model, smoothing the weight map, similar in effect to the modulation technique of (Haufe et al., 2014). Different from them, SAR modulation is performed iteratively, where Haufe et al. only perform the modulation afterward. C.f. Figs. 5 and 7 from our paper and Fig. 6 from (Haufe et al., 2014). Additionally, we only consider the relation between weights from the same anatomical region, which helps to avoid over-smoothing across the boundary, as shown in Fig. 8.

Future work

Although the proposed model (group lasso SVM + SAR) obtains a good classification accuracy and generates a smoother, more localized and stable classifier, there are still some points that can be improved. In this work, we only considered single level grouping (using S, SLIC or ac-SLIC), which may introduce a granularity bias. Such bias can be avoided by using multiple levels of grouping, which may then be used in a hierarchical group lasso method (Lim and Hastie, 2015; Jenatton et al., 2011). Secondly, the proposed method can be extended to include other features than the gray matter feature used in this paper. Candidate features are cortical thickness (Moradi et al., 2017), the curvature of the cortical surface (Kim et al., 2016; Lyu et al., 2017), or combinations of features (de Vos et al., 2016; Adeli et al., 2017). While these features can be readily incorporated into the proposed framework, visualizing the weight maps requires more work. For example, instead of visualizing a scalar map for one feature, one should visualize the local magnitude of multiple features. Thirdly, we may extend the evaluation to a larger dataset, such as the complete ADNI database (http://adni.loni.usc.edu/), BRAINnet (http://www.brainnet.net/)or Enigma (http://enigma.ini.usc.edu/).

Conclusions

In conclusion, this paper introduces a general linear SVM framework that integrates spatial-anatomical regularization and sparsity in a single cost function. The framework can be solved efficiently using FISTA, and allows extension with new (quadratic) regularization terms to encode additional types of prior information. By introducing a novel anatomically constrained supervoxel method called ac-SLIC, the proposed method (group lasso SVM + SAR) can select subregions of brain structures in a data-driven way.

In an experiment distinguishing Cognitive Normals from Alzheimer's Disease subjects, classification results improved from 84.6% for linear SVM to 89.3% for the proposed method, while yielding a visually more smooth and localized model, which aids interpretation of the resulting weight map. The selected brain regions can be used to localize the important patterns of morphological brain changes associated with AD. This selection was also more stable for the proposed model than alternative strategies. The most stable selected areas were the ventricles and the hippocampus, in line with current knowledge.

Acknowledgments

This project is funded by the Joint Scientific Thematic Research Programme (JSTP) between The Netherlands and China (Project 116350001). This research has received partial funding from the European Union Seventh Framework Programme (FP7/2007–2013) under Grant Agreement 604102 (Human Brain Project).

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir,Inc.; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

References

Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Süsstrunk, S., 2012. SLIC superpixels compared to state-of-the-art superpixel methods. IEEE Trans. Pattern Anal. Mach. Intell. 34 (11), 2274–2282.

Adeli, E., Meng, Y., Li, G., Lin, W., Shen, D., 2017. Joint sparse and low-rank regularized multi-task multi-linear regression for prediction of infant brain development with incomplete data. In: International Conference on Medical Image Computing and Computer-assisted Intervention. Vol. 10433 of Lecture Notes in Computer Science. Springer, pp. 40–48.

Alzheimer's Association, 2014. 2014 Alzheimer's disease facts and figures. Alzheimer's Dementia 10 (2), 47–92.

Apostolova, L.G., Green, A.E., Babakchanian, S., Hwang, K.S., Chou, Y.-Y., Toga, A.W., Thompson, P.M., 2012. Hippocampal atrophy and ventricular enlargement in normal aging, mild cognitive impairment and alzheimer's disease. Alzheimer Dis. Assoc. Disord. 26 (1), 17.

Arbabshirani, M.R., Plis, S., Sui, J., Calhoun, V.D., 2017. Single subject prediction of brain disorders in neuroimaging: promises and pitfalls. Neuroimage 145, 137–165.

Ashburner, J., Friston, K.J., 2000. Voxel-based morphometrythe methods. Neuroimage 11 (6), 805–821.

Beck, A., Teboulle, M., 2009. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM J. Imag. Sci. 2 (1), 183–202.

Bredies, K., Lorenz, D.A., 2008. Linear convergence of iterative soft-thresholding. J. Fourier Anal. Appl. 14 (5), 813–837.

Bron, E.E., Smits, M., Niessen, W.J., Klein, S., 2015. Feature selection based on the SVM weight vector for classification of dementia. IEEE J. Biomed. Health Inf. 19 (5), 1617–1626.

Combettes, P.L., Pesquet, J.-C., 2011. Proximal splitting methods in signal processing. In: Fixed-point Algorithms for Inverse Problems in Science and Engineering. Springer, pp. 185–212.

Coupé, P., Eskildsen, S.F., Manjón, J.V., Fonov, V.S., Collins, D.L., Alzheimer's disease Neuroimaging Initiative, et al., 2012a. Simultaneous segmentation and grading of anatomical structures for patient's classification: application to Alzheimer's disease. Neuroimage 59 (4), 3736–3747.

Coupé, P., Eskildsen, S.F., Manjón, J.V., Fonov, V.S., Pruessner, J.C., Allard, M., Collins, D.L., Alzheimer's Disease Neuroimaging Initiative, et al., 2012b. Scoring by nonlocal image patch estimator for early detection of Alzheimer's disease. Neuroimage: Clin. 1 (1), 141–152.

Coupé, P., Fonov, V.S., Bernard, C., Zandifar, A., Eskildsen, S.F., Helmer, C., Manjón, J.V., Amieva, H., Dartigues, J.-F., Allard, M., et al., 2015. Detection of Alzheimer's disease signature in mr images seven years before conversion to dementia: toward an early individual prognosis. Hum. Brain Mapp. 36 (12), 4758–4770.

Coupé, P., Manjón, J.V., Fonov, V., Pruessner, J., Robles, M., Collins, D.L., 2011. Patch-based segmentation using expert priors: application to hippocampus and ventricle segmentation. Neuroimage 54 (2), 940–954.

Cui, Y., Liu, B., Luo, S., Zhen, X., Fan, M., Liu, T., Zhu, W., Park, M., Jiang, T., Jin, J.S., et al., 2011. Identification of conversion from mild cognitive impairment to alzheimer's disease using multivariate predictors. PLoS One 6 (7), e21896.

Cuingnet, R., Gerardin, E., Tessieras, J., Auzias, G., Lehéricy, S., Habert, M.-O., Chupin, M., Benali, H., Colliot, O., 2011. Automatic classification of patients with Alzheimer's disease from structural MRI: a comparison of ten methods using the ADNI database. Neuroimage 56 (2), 766–781.

Cuingnet, R., Glaunes, J.A., Chupin, M., Benali, H., Colliot, O., 2013. Spatial and anatomical regularization of SVM: a general framework for neuroimaging data. IEEE Trans. Pattern Anal. Mach. Intell. 35 (3), 682–696.

Davatzikos, C., Resnick, S.M., Wu, X., Parmpi, P., Clark, C.M., Jul. 2008. Individual patient diagnosis of AD and FTD via high-dimensional pattern classification of MRI. Neuroimage 41 (4), 1220–1227.
De Santi, S., de Leon, M.J., Rusinek, H., Convit, A., Tarshish, C.Y., Roche, A., Tsui, W.H., Kandil, E., Boppana, M., Daisley, K., 2001. Hippocampal formation glucose metabolism and volume losses in MCI and AD. Neurobiol. Aging 22 (4), 529–539.
de Vos, F., Schouten, T.M., Hafkemeijer, A., Dopper, E.G., van Swieten, J.C., de Rooij, M., van der Grond, J., Rombouts, S.A., 2016. Combining multiple anatomical mri measures improves Alzheimer's disease classification. Hum. Brain Mapp. 37 (5), 1920–1929.
Dunne, K., Cunningham, P., Azuaje, F., 2002. Solutions to instability problems with sequential wrapper-based approaches to feature selection. J. Mach. Learn. Res. 1–22.
Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., Lin, C.-J., 2008a. Liblinear: a library for large linear classification. J. Mach. Learn. Res. 9 (Aug), 1871–1874.
Fan, Y., Batmanghelich, N., Clark, C.M., Davatzikos, C., 2008b. Spatial patterns of brain atrophy in MCI patients, identified via high-dimensional pattern classification, predict subsequent cognitive decline. Neuroimage 39 (4), 1731–1743.
Fan, Y., Shen, D., Davatzikos, C., 2005. Classification of structural images via high-dimensional image warping, robust feature extraction, and SVM. In: Duncan, J., Gerig, G. (Eds.), Medical Image Computing and Computer-assisted Intervention MICCAI 2005. Vol. 3749 of Lecture Notes in Computer Science, pp. 1–8.
Fan, Y., Shen, D., Gur, R.C., Gur, R.E., Davatzikos, C., 2007. COMPARE: classification of morphological patterns using adaptive regional elements. IEEE Trans. Med. Imag. 26 (1), 93–105.
Frisoni, G.B., Fox, N.C., Jack, C.R., Scheltens, P., Thompson, P.M., Feb. 2010. The clinical use of structural MRI in Alzheimer disease. Nat. Rev. Neurol. 6 (2), 67–77.
Grosenick, L., Klingenberg, B., Katovich, K., Knutson, B., Taylor, J.E., 2013. Interpretable whole-brain prediction analysis with graphnet. Neuroimage 72, 304–321.
Hampel, H., Teipel, S.J., Bayer, W., Alexander, G.E., Schwarz, R., Schapiro, M.B., Rapoport, S.I., Mller, H.-J., 2002. Age transformation of combined hippocampus and amygdala volume improves diagnostic accuracy in Alzheimer's disease. J. Neurol. Sci. 194 (1), 15–19.
Haufe, S., Meinecke, F., Görgen, K., Dähne, S., Haynes, J.-D., Blankertz, B., Bießmann, F., 2014. On the interpretation of weight vectors of linear models in multivariate neuroimaging. Neuroimage 87, 96–110.
Jack, C.R., Bernstein, M.A., Fox, N.C., Thompson, P., Alexander, G., Harvey, D., Borowski, B., Britson, P.J., L Whitwell, J., Ward, C., et al., 2008. The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods. J. Magn. Reson. Imag. 27 (4), 685–691.
Jenatton, R., Mairal, J., Obozinski, G., Bach, F., 2011. Proximal methods for hierarchical sparse coding. J. Mach. Learn. Res. 12 (Jul), 2297–2334.
Jie, B., Zhang, D., Cheng, B., Shen, D., 2015. Manifold regularized multitask feature learning for multimodality disease classification. Hum. Brain Mapp. 36 (2), 489–507.
Kim, S.H., Lyu, I., Fonov, V.S., Vachet, C., Hazlett, H.C., Smith, R.G., Piven, J., Dager, S.R., Mckinstry, R.C., Pruett, J.R., et al., 2016. Development of cortical shape in the human brain from 6 to 24months of age via a novel measure of shape complexity. Neuroimage 135, 163–176.
Klein, S., Staring, M., Murphy, K., Viergever, M.A., Pluim, J.P., 2010. Elastix: a toolbox for intensity-based medical image registration. IEEE Trans. Med. Imag. 29 (1), 196–205.
Klöppel, S., Peter, J., Ludl, A., Pilatus, A., Maier, S., Mader, I., Heimbach, B., Frings, L., Egger, K., Dukart, J., et al., 2015. Applying automated mr-based diagnostic methods to the memory clinic: a prospective study. J. Alzheim. Dis. 47 (4), 939–954.
Klöppel, S., Stonnington, C.M., Chu, C., Draganski, B., Scahill, R.I., Rohrer, J.D., Fox, N.C., Jack, C.R., Ashburner, J., Frackowiak, R.S.J., 2008. Automatic classification of MR scans in Alzheimer's disease. Brain 131 (3), 681–689.
Kohannim, O., Hibar, D.P., Jahanshad, N., Stein, J.L., Hua, X., Toga, A.W., Jack, C.R., Weinen, M.W., Thompson, P.M., 2012. Predicting temporal lobe volume on mri from genotypes using l^1–l^2 regularized regression. In: IEEE International Symposium on Biomedical Imaging (ISBI). IEEE, pp. 1160–1163.
Leon, M.J., Mosconi, L., Li, J., Santi, S., Yao, Y., Tsui, W.H., Pirraglia, E., Rich, K., Javier, E., Brys, M., Glodzik, L., Switalski, R., Saint Louis, L.A., Pratico, D., 2007. Longitudinal CSF isoprostane and MRI atrophy in the progression to AD. J. Neurol. (12), 1666–1675.
Li, Y., Wang, Y., Wu, G., Shi, F., Zhou, L., Lin, W., Shen, D., 2012. Discriminant analysis of longitudinal cortical thickness changes in alzheimer's disease using dynamic and network features. Neurobiol. Aging 33 (2), 427–e15.
Liang, Z.-P., Lauterbur, P.C., 2000. Principles of Magnetic Resonance Imaging: a Signal Processing Perspective. The Institute of Electrical and Electronics Engineers Press.
Lim, M., Hastie, T., 2015. Learning interactions via hierarchical group-lasso regularization. J. Comput. Graph Stat. 24 (3), 627–654.
Liu, F., Wee, C.-Y., Chen, H., Shen, D., 2014. Inter-modality relationship constrained multi-modality multi-task feature selection for alzheimer's disease and mild cognitive impairment identification. Neuroimage 84, 466–475.
Liu, M., Suk, H.-I., Shen, D., 2013. Multi-task sparse classifier for diagnosis of MCI conversion to AD with longitudinal MR images. In: International Workshop on Machine Learning in Medical Imaging. Vol. 8184 of Lecture Notes in Computer Science. Springer, pp. 243–250.
Lyu, I., Kim, S.H., Bullins, J., Gilmore, J.H., Styner, M.A., 2017. Novel local shape-adaptive gyrification index with application to brain development. In: International Conference on Medical Image Computing and Computer-assisted Intervention. Vol. 10433 of Lecture Notes in Computer Science. Springer, pp. 31–39.
Mattsson, N., Zetterberg, H., Hansson, O., Andreasen, N., Parnetti, L., Jonsson, M., Herukka, S.-K., van der Flier, W.M., Blankenstein, M.A., Ewers, M., 2009. CSF biomarkers and incipient Alzheimer disease in patients with mild cognitive impairment. J. Am. Med. Assoc. 302 (4), 385–393.
Moore, A.P., Prince, S.J., Warrell, J., Mohammed, U., Jones, G., 2008. Superpixel lattices. In: Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on. IEEE, pp. 1–8.
Moradi, E., Khundrakpam, B., Lewis, J.D., Evans, A.C., Tohka, J., 2017. Predicting symptom severity in autism spectrum disorder based on cortical thickness measures in agglomerative data. Neuroimage 144, 128–141.
Moradi, E., Pepe, A., Gaser, C., Huttunen, H., Tohka, J., Initiative, A.D.N., et al., 2015. Machine learning framework for early mri-based alzheimer's conversion prediction in mci subjects. Neuroimage 104, 398–412.
Oliveira Jr., P.P. d. M., Nitrini, R., Busatto, G., Buchpiguel, C., Sato, J.R., Amaro Jr., E., 2010. Use of SVM methods with surface-based cortical and volumetric subcortical measurements to detect Alzheimer's disease. J. Alzheim. Dis. 19 (4), 1263–1272.
Ota, K., Oishi, N., Ito, K., Fukuyama, H., Group, S.-J.S., Initiative, A.D.N., et al., 2015. Effects of imaging modalities, brain atlases and feature selection on prediction of alzheimer's disease. J. Neurosci. Meth. 256, 168–183.
Pegueroles, J., Vilaplana, E., Montal, V., Sampedro, F., Alcolea, D., Carmona-Iragui, M., Clarimon, J., Blesa, R., Lleó, A., Fortea, J., et al., 2017. Longitudinal brain structural changes in preclinical Alzheimer's disease. Alzheimer's Dementia 13 (5), 499–509.
Pini, L., Pievani, M., Bocchetta, M., Altomare, D., Bosco, P., Cavedo, E., Galluzzi, S., Marizzoni, M., Frisoni, G.B., 2016. Brain atrophy in Alzheimers disease and aging. Ageing Res. Rev. 30, 25–48.
Plant, C., Teipel, S.J., Oswald, A., Böhm, C., Meindl, T., Mourao-Miranda, J., Bokde, A.W., Hampel, H., Ewers, M., 2010. Automated detection of brain atrophy patterns based on mri for the prediction of alzheimer's disease. Neuroimage 50 (1), 162–174.
Shen, L., Kim, S., Qi, Y., Inlow, M., Swaminathan, S., Nho, K., Wan, J., Risacher, S.L., Shaw, L.M., Trojanowski, J.Q., et al., 2011. Identifying neuroimaging and proteomic biomarkers for mci and ad via the elastic net. In: International Workshop on Multimodal Brain Image Analysis. Vol. 7012 of Lecture Notes in Computer Science. Springer, pp. 27–34.
Sun, Z., Fan, Y., Lelieveldt, B.P., van de Giessen, M., 2015a. Detection of alzheimer's disease using group lasso svm-based region selection. In: SPIE Medical Imaging. International Society for Optics and Photonics, 941414–941414.
Sun, Z., Lelieveldt, B.P.F., Staring, M., 2015b. Fast linear geodesic shape regression using coupled logdemons registration. In: IEEE International Symposium on Biomedical Imaging (ISBI), pp. 1276–1279.
Sun, Z., van de Giessen, M., Lelieveldt, B.P., Staring, M., 2017. Detection of conversion from mild cognitive impairment to alzheimer's disease using longitudinal brain mri. Front. Neuroinf. 11.
Sun, Z., Veerman, J.A., Jasinschi, R.S., 2012. A method for detecting interstructural atrophy correlation in mri brain images. In: IEEE International Conference on Image Processing (ICIP). IEEE, pp. 1253–1256.
Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J. Roy. Stat. Soc. B 267–288.
Tohka, J., Moradi, E., Huttunen, H., Initiative, A.D.N., et al., 2016. Comparison of feature selection techniques in machine learning for anatomical brain MRI in dementia. Neuroinformatics 1–18.
Toloşi, L., Lengauer, T., 2011. Classification with correlated features: unreliability of feature ranking and solutions. Bioinformatics 27 (14), 1986–1994.
Vedaldi, A., Soatto, S., 2008. Quick shift and kernel methods for mode seeking. Comput. Vis.-ECCV 705–718, 2008.
Veksler, O., Boykov, Y., Mehrani, P., 2010. Superpixels and supervoxels in an energy optimization framework. Comput. Vis.-ECCV 211–224, 2010.
Wachinger, C., Reuter, M., Initiative, A.D.N., et al., 2016. Domain adaptation for Alzheimer's disease diagnostics. Neuroimage 139, 470–479.
Wee, C.-Y., Yap, P.-T., Shen, D., 2013. Prediction of alzheimer's disease and mild cognitive impairment using cortical morphological patterns. Hum. Brain Mapp. 34 (12), 3411–3425.
Wu, M.-J., Mwangi, B., Bauer, I.E., Passos, I.C., Sanches, M., Zunta-Soares, G.B., Meyer, T.D., Hasan, K.M., Soares, J.C., 2017. Identification and individualized prediction of clinical phenotypes in bipolar disorders using neurocognitive data, neuroimaging scans and machine learning. Neuroimage 145, 254–264.
Xin, B., Kawahara, Y., Wang, Y., Hu, L., Gao, W., 2016. Efficient generalized fused lasso and its applications. ACM Trans. Intell. Syst. Technol. 7 (4), 60.
Xu, Z., Jin, R., Yang, H., King, I., Lyu, M.R., 2010. Simple and efficient multiple kernel learning by group lasso. In: Proceedings of the 27th International Conference on Machine Learning (ICML-10), pp. 1175–1182.
Yu, G., Liu, Y., Thung, K.-H., Shen, D., 2014. Multi-task linear programming discriminant analysis for the identification of progressive mci individuals. PLoS One 9 (5), e96458.
Zhang, D., Shen, D., Initiative, A.D.N., et al., 2012a. Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in Alzheimer's disease. Neuroimage 59 (2), 895–907.
Zhang, D., Shen, D., Initiative, A.D.N., et al., 2012b. Predicting future clinical changes of mci patients using longitudinal and multimodal biomarkers. PLoS One 7 (3), e33182.
Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. J. Roy. Stat. Soc. B 67 (2), 301–320.