CHAPTER

Machine learning as a means toward precision diagnostics and prognostics 10

A. Sotiras, B. Gaonkar, H. Eavani, N. Honnorat, E. Varol, A. Dong, C. Davatzikos
University of Pennsylvania, Philadelphia, PA, United States

CHAPTER OUTLINE

10.1 INTRODUCTION

The advent of new imaging modalities providing high-resolution depictions of the anatomy (Hsieh, 2009; Atlas, 2009; Liang and Lauterbur, 2000) and function (Detre et al., 1992; Baxton, 2009; Phelps, 2000) of the brain in disease and health has resulted in medical imaging becoming increasingly indispensable for patients' healthcare. The way medical images are analyzed has been greatly shaped by machine learning, which has found application in numerous fields, including image segmentation (Pham et al., 2000; Heimann and Meinzer, 2009), image registration

Machine Learning and Medical Imaging. http://dx.doi.org/10.1016/B978-0-12-804076-8.00010-4

(Maintz and Viergever, 1998; Sotiras et al., 2013), image fusion, and computer-aided diagnosis (Sajda, 2006).

Among the reasons behind the success of machine learning in medical imaging are increased automation, high sensitivity and specificity. Machine learning has fueled automated approaches that provide measurements by circumventing the error-prone and labor-intensive manual procedures that are typically involved in traditional interest-based analyses. Moreover, contrary to conventional automated approaches, such as mass univariate analyses, high-dimensional multivariate pattern analysis (MPVA) (Norman et al., 2006; McIntosh and Mišić, 2013) driven methods fully harness the potential of high-dimensional data by examining statistical relationships between elements that span the whole image domain.

Integrating information from the whole domain while taking advantage of prior knowledge allows MVPA techniques to identify and measure subtle and spatially complex structural and functional changes in the brain that are induced by disease or pharmacological interventions, despite important normal variability. As a consequence, sophisticated pattern analysis techniques have been employed to identify disease-specific signatures and elucidate the selective vulnerability of different brain networks to different pathologies (Mour ao-Miranda et al., 2005; Zhu et al., 2008; Sun et al., 2009; Zeng et al., 2012; Davatzikos et al., 2008a; Klöppel et al., 2008; Vemuri et al., 2008; Duchesne et al., 2008). This has led to the construction of sensitive biomarkers that are able to quantify the risk of developing a disease, track the disease progression or the effect of pharmacological interventions in clinical trials, and deliver patient-specific diagnosis before measurable clinical effects occur.

Neurodegenerative diseases such as Alzheimer's disease (AD) have been in the epicenter of the development of computerized biomarkers. Machine learning diagnostic and prognostic tools have been developed to identify patients with neurodegenerative diseases such as dementia (Davatzikos et al., 2008a; Klöppel et al., 2008; Vemuri et al., 2008; Duchesne et al., 2008; Fan et al., 2008b; McEvoy et al., 2009; Hinrichs et al., 2009; Gerardin et al., 2009; Hinrichs et al., 2011; Zhang et al., 2011; Cuingnet et al., 2011), to differentially distinguish between AD and frontotemporal dementia (FTD) (Davatzikos et al., 2008b), or to predict clinical progression of patients (Fan et al., 2008a; Davatzikos et al., 2011). Studies into mental disorders have also benefited from the application of computer-assisted imaging techniques. Fully automated classification algorithms have been successfully applied to diagnose a wide range of neurological and psychiatric diseases, including schizophrenia (Davatzikos et al., 2005; Koutsouleris et al., 2015), psychosis (Koutsouleris et al., 2012), and depression (Mour ao-Miranda et al., 2011).

However, despite important advances and successes, there remain significant challenges to be addressed. Three of the most important challenges comprise (i) dimensionality reduction; (ii) interpreting the learned model; and (iii) elucidating disease heterogeneity.

The first challenge tackles a fundamental problem one encounters when training machine learning models to identify imaging signatures towards automated

diagnosis and prognosis, namely the sheer dimensionality of imaging data along with the relatively small sample size that is typically available. This problem is further exacerbated by the increasing resolution of the imaging data, as well as the increasing availability of multiparametric imaging, which further increase the dimensionality and complexity of the available data. The main challenge here is summarizing the imaging information through a reduced number of features that is compatible with the sample size of a typical imaging study, while retaining the necessary information that will allow the learning system to recognize relevant imaging patterns.

The second challenge relates to the interpretability of the learned model. Machine learning models are generally treated as "black-boxes" that provide us with an index of the presence of a disease. While this index may be used to perform diagnosis, it does not inform us about how each brain region contributes to the construction of the discriminative multivariate pattern. This information is of significant importance since it provides key insight regarding the selective vulnerability of different brain systems to different pathologies, thus elucidating disease mechanisms, paving the road for more effective treatments.

The third challenge addresses the problem of elucidating disease heterogeneity. Most existing methodologies assume a single, unifying pathophysiological process and aim to reveal it by identifying a unique imaging pattern that can distinguish between healthy and diseased populations, or between two subgroups of patients. However, this assumption effectively disregards ample evidence for the heterogeneous nature of brain diseases. Neurodegenerative, neuropsychiatric and neurodevelopmental disorders are characterized by high clinical heterogeneity, which is likely due to the underlying neuroanatomical heterogeneity of various pathologies. Elucidating disease heterogeneity is crucial for deepening our understanding of the involved pathological mechanisms, and may lead to more precise diagnosis, prognosis, and specialized treatment.

In this chapter, we are going to present solutions for tackling the aforementioned challenges. In Section 10.2, we present a clustering and statistical-based approached for dimensionality reduction of both structural and functional data. In Section 10.3, we detail an efficient technique for deriving statistical significance maps in classification tasks using support vector machines (SVMs), while in Section 10.4 we present a palette of techniques to tackle disease heterogeneity under different methodological assumptions. In Section 10.5, we provide evidence of the usefulness of machine learning techniques at the clinical and research level, while Section 10.6 concludes the chapter.

10.2 DIMENSIONALITY REDUCTION

During the past decades, the advent of high-resolution imaging techniques has given rise to high-dimensional, complex clinical datasets consisting of hundreds of patient scans that comprise millions of voxels (Van Essen et al., 2013; Satterthwaite

et al., 2014). The high dimensionality of the data, along with the relatively small sample size that is typically available, poses an important challenge when aiming to holistically analyze imaging patterns in association with brain diseases. This challenge is further exacerbated by the increasing availability of multiparametric imaging data, which results in an additional increase in both the dimensionality and complexity of the data. Moreover, the emergence of sophisticated imaging techniques, such as diffusion tensor imaging and functional magnetic resonance imaging that derive complex representations of the axonal anatomy and brain activity, not only emphasizes the aforementioned challenge but also calls for tailored analysis tools.

To address this challenge, dimensionality reduction is typically performed. The aim is to extract, in an optimal way, a few imaging features, thus reducing the dimensionality of the data to a level that is compatible with the sample size of a typical imaging study. Additionally, these features should retain the important image information that will allow for the identification of imaging patterns that offer good predictive value.

Numerous approaches have been proposed to reduce the dimensionality of imaging data. Dimensionality reduction methods can be typically categorized into two groups: (i) spatial grouping and (ii) statistically driven reduction, depending on the driving assumption behind its method. In the first case, one aims to group together elements that are spatially close and similar in terms of imaging measurements. In the second case, emphasis is put on considering together image elements that vary in consistent ways across the population. This taxonomy may be further refined by taking into account the nature of the imaging data the method handles.

10.2.1 DIMENSIONALITY REDUCTION THROUGH SPATIAL GROUPING

Methods of this class typically formulate the problem as clustering, and dimensionality reduction is achieved by summarizing the data through a restricted set of features that correspond to the estimated clusters. Features are typically extracted by computing a single average measure per estimated cluster, while clusters are obtained by segmenting the brain into contiguous regions that encompass elements with imaging measurements that are similar to each other. Defining an appropriate similarity measure is of significant importance for the success of these methods and should take into account the nature of the imaging signal, leading to data-specific algorithms. In the following, we summarize two such algorithms for structural magnetic resonance imaging (MRI) scans and resting-state functional MRI (rs-fMRI), respectively.

10.2.2 SPATIAL GROUPING OF STRUCTURAL MRI

Structural imaging based on magnetic resonance provides information regarding the integrity of gray and white matter structures in the brain, making it an integral part of the clinical assessment of patients with dementia, such as AD and FTD. Automated

classification approaches applied on structural MRI data have shown promise for the diagnosis of AD and the identification of whole-brain patterns of disease-specific atrophy. In this scenario, when dimensionality reduction is performed prior to a supervised machine learning task, such as patient classification, it is appealing to adopt a *supervised* clustering approach. The goal is to exploit prior information (ie, disease diagnosis) in order to generate regions of interest that are adapted not only to the data, but also to the machine learning task, with the aim to improve its performance.

This supervised approach was adopted by the COMPARE method (Fan et al., 2007) that aims to perform classification of morphological patterns using adaptive regional elements. COMPARE extracts spatially smooth clusters that can be used to train a classifier to predict patient diagnosis by combining information stemming from both the imaging signal and subjects' diagnosis. The two types of information are integrated at each image location, p, in a multiplicative fashion though the score:

$$s(p) = P(p)C(p), \tag{10.1}$$

where $C(p)$ measures the spatial consistency of the imaging signal, while $P(p)$ measures discriminative power. More precisely, P is calculated as the following leave-one-out absolute Pearson correlation:

$$P(p) = \texttt{argmin}_{i=1..n} |\rho(p, i)|, \tag{10.2}$$

where $\rho(p, i)$ denotes the Pearson correlation measured between the imaging signal at p and the classification labels when excluding the ith subject/sample. The consistency $C(p)$ is the intra-class coefficient measuring the proportion of neighboring feature variance that is explained by the inter-subject variability (McGraw and Wong, 1996; Fan et al., 2007). It takes values between 0 and 1, with higher values indicating that the variance of the measurements across neighboring brain location is small with respect to the inter-subject variability of the imaging signal. As a result, the score $s(p)$ is bounded between 0 and 1, with values close to 1 indicating that the imaging signal around p is simultaneously highly reliable and discriminative (ie, highly correlated or anticorrelated with patient diagnosis).

This score map is subsequently smoothed, and its gradient is used in conjunction with a watershed segmentation algorithm (Vincent and Soille, 1991) to partition the brain into different regions (Fig. 10.1 presents brain regions generated by watershed from white matter tissue density maps of demented and normally aging subjects (Shen and Davatzikos, 2002; Fan et al., 2007)). These regions are then refined by considering only locations that optimize the classification power of the extracted features. This is performed in a region growing fashion where, initially, only the node of the region with the highest discriminative score is selected, and adjacent locations are incrementally aggregated as long as the discriminative power does not decrease. The previous steps are summarized in Algorithm 10.1. This approach extracts a single connected component per watershed region. Each component comprises highly discriminative elements whose average imaging signal may be used as a feature for training a classifier, such as an SVM (Vapnik, 2000).

ALGORITHM 10.1 PSEUDO-CODE FOR COMPARE.

```
Data: Structural MRI images and group labels
Result: Regional elements optimized for classification
/* Initialization                                          */
for each image location p do
    compute discriminative power P(p) using Eq. 10.2;
    compute spatial consistency C(p);
    compute score s(p) using Eq. 10.1;
end
/* Region segmentation                                     */
Gaussian smoothing of s(p);
gradient computation of score map;
parcellate score map using watershed;
/* Region refinement/feature extraction                    */
for each parcel do
    select voxel with highest score s(p);
    /* Region growing                                      */
    while regional discriminative power increases do
        find voxel with highest discriminative
        power at the border of growing region;
        add voxel to growing region;
    end
end
```

FIG. 10.1

Coronal and sagittal cross-sectional views of a watershed segmentation generated by COMPARE.

The efficiency of this supervised dimensionality reduction scheme was demonstrated in classifying patients with clinical dementia versus normal individuals, as well as distinguishing between schizophrenic patients and normal controls (Fan et al., 2007). COMPARE is generic and can be readily extended to incorporate different forms of prior information, such as the ones provided in regression and multiclass classification settings.

10.2.2.1 Spatial grouping of rs-fMRI

Functional MRI is an imaging technique that reflects neural activity in the whole brain by detecting changes in oxygen consumption. Resting-state fMRI reveals brain networks (Biswal, 2012) by evaluating regional interactions that occur when the subjects are relaxed and do not perform a particular mental task during the brain scan. The dynamic nature of this imaging modality results in extremely voluminous and complex datasets, underlining the need for efficient dimensionality reduction.

Clustering approaches have received considerable attention towards reducing the dimensionality of functional data. This is due to the fact that clustering is not only an efficient way to reduce the spatial dimension of rs-fMRI data, but also a biologically meaningful one. Clustering sheds light on the mid-scale functional structure of the brain that is considered to follow a *segregation and integration principle*. In other words, information is thought to be processed by compact groups of neurons in the brain, or *functional units*, that collaborate together towards addressing complex tasks (Tononi et al., 1994).

Clustering approaches typically aim to divide the brain into spatially smooth areas that are likely to correspond to the functional units that constitute the brain. This is usually performed by first representing the brain in the form of a graph, where nodes represent brain locations and edges connect nodes that correspond to spatially adjacent locations. The weight of the edges represents the strength of the connectivity between nodes and is estimated by computing the similarity between the rs-fMRI signals that are measured at each node. The similarity is commonly measured by the Pearson correlation or partial correlation (Smith et al., 2011). Once the node is constructed, adjacent brain locations that are strongly connected are grouped together in the same parcel.

Numerous methods have been proposed for this task. Among the most popular methods, one may cite hierarchical clustering (Cordes et al., 2002), normalized clustering (Shen et al., 2010; Craddock et al., 2012), k-means (Bellec et al., 2010), region growing (Blumensath et al., 2013; Heller et al., 2006), and Markov random fields (MRFs) (Ryali et al., 2013; Golland et al., 2007; Honnorat et al., 2015). Different methods exhibit distinct advantages and disadvantages. Generally, many of the above methods are either initialization dependent (eg, region growing (Blumensath et al., 2013; Heller et al., 2006) and k-means (Bellec et al., 2010)), or rely on complex models that involve a large number of parameters (Golland et al., 2007). As a result, they are sensitive to initialization and suffer from limitations related to the employed heuristics (eg, hierarchical clustering may lead to the creation of poorly fit parcels at coarser scales (Cordes et al., 2002; Honnorat et al., 2015)), and the large number of inferred parameters that may negatively impact the quality of the locally optimal solution that is obtained (Golland et al., 2007). Moreover, not all methods produce contiguous parcels.

In order to address the aforementioned concerns, a discrete MRF approach, termed GraSP (Graph based segmentation with Shape Priors), was recently introduced in (Honnorat et al., 2015). This approach adopts an *exemplar-based clustering*

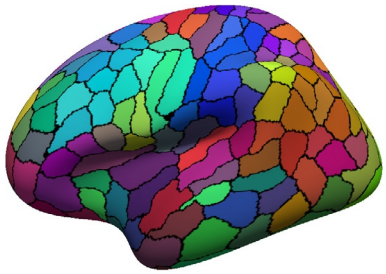

FIG. 10.2

Functional parcellation of the left hemisphere of the brain, projected on an inflated brain surface.

approach that allows for the reduction of the number of parameters by representing the rs-fMRI time series of each parcel by the signal of one of the nodes that are assigned to it. Thus, the clustering framework is simplified through the encoding of the parcels with their functional center. Only one parameter needs to be chosen by the user, the label cost K. This corresponds to the cost of introducing a new parcel into the clustering result, which indirectly determines the size of the produced parcels (Delong et al., 2012). Contrary to other MRF clustering methods (Ryali et al., 2013), these parcels are connected (Fig. 10.2 presents a functional parcellation that was produced for reducing the dimension of rs-fMRI scans from a neurodevelopmental study (Satterthwaite et al., 2014)). Parcel connectedness is promoted without any spatial smoothing by the inclusion of a shape prior term into the MRF energy formulation (Veksler, 2008; Gulshan et al., 2010). Lastly, the energy is optimized in a single step, thus removing the need for initialization and specification of a stopping criterion.

The MRF energy is summarized in the following form:

$$\min \sum_{p} V_p(l_p) + L_p(\{l_p\}) + S_p(\{l_p\}),$$

where p denotes a node of the brain graph, l_p the parcel that should contain this node, $V_p(l_p)$ is a cost that decreases when the node p is assigned to a parcel l_p with highly correlated rs-fMRI signal, $L_p(\{l_p\})$ penalizes by a positive cost K the introduction of a parcel of functional center p, and the $S_p(\{l_p\})$ are the shape priors that enforce the connectedness of each parcel p. This energy is optimized by exploiting advanced solvers (Delong et al., 2012) that could provide a substantial advance over existing methods. Experimental results on large datasets demonstrated that this approach is capable of generating parcels that are all highly coherent, while the overall parcellation is slightly more reproducible than the results produced by hierarchical clustering and normalized cuts (Honnorat et al., 2015).

10.2.3 STATISTICALLY DRIVEN DIMENSIONALITY REDUCTION

The second family of dimensionality reduction methods is based on exploiting statistical procedures to project the data in a space of lower dimension. This is typically performed within a regularized matrix factorization framework where a tall matrix $\mathbf{X}$ comprising N samples/images of dimension D, each one arrayed per column $\left(\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N], \mathbf{x}_i \in \mathbb{R}^D\right)$, is approximated by a product of matrices ($\mathbf{X} \approx \mathbf{BC}$). $\mathbf{B}$ is a matrix of the basis vectors that span the estimated subspace, and $\mathbf{C}$ contains the loading coefficients that provide the low-dimensional description of the data. Depending on the implemented modeling assumptions, $\mathbf{B}$ and $\mathbf{C}$ exhibit different properties.

Among the most widely used methods of this class, one may cite principal component analysis (PCA) (Friston et al., 1993; Strother et al., 1995; Hansen et al., 1999) and independent component analysis (ICA). PCA maps the data to a lower dimensional space through an orthogonal linear transformation, while preserving the variance of the data. The transformation is performed in such a way that basis vectors (or principal components) are ordered in descending order according to the amount of the variance they explain. ICA (McKeown et al., 1998; Calhoun et al., 2001; Beckmann and Smith, 2004), on the other hand, maps the data into a set of components that are as statistically independent from each other as possible.

Despite their widespread use in neuroimaging, conventional factorization methods that are used for dimensionality reduction suffer from limitations related to the interpretability and the reproducibility of the derived representation. For example, both PCA and ICA estimate components and coefficients of mixed sign, thus approximating the data through complex mutual cancelation between component regions of opposite sign. This complex modeling of the data, along with the fact that the estimated components highly overlap due to their often global spatial support, results in representations that lack specificity. In other words, while it is possible to interpret individual components, it is difficult to associate a specific brain region with a specific effect. Lastly, conventional factorization methods, and especially PCA, aim to approximate the data as faithfully as possible, thus capturing both relevant and irrelevant sources of variation, resulting in poor generalization in unseen datasets.

Next, we summarize our group's work to derive efficient, interpretable, and reproducible statistically driven dimensionality reduction techniques for structural and functional MRI data. The key idea behind the developed frameworks is to derive highly parsimonious representations. The reason behind this choice is twofold: (i) sparse methods achieve a higher degree of specificity than conventional multivariate analysis methods (Lee and Seung, 1999); and (ii) they show improved generalizability (Avants et al., 2010). The above underline the importance of sparsity in brain modeling and analysis (Daubechies et al., 2009). Sparsity is introduced in a tailored way, taking into account the specific properties of different imaging modalities.

10.2.3.1 Statistically driven dimensionality reduction of structural MRI

Structural MRI scans typically encode the physical properties of the image tissue through the use of non-negative values. This fact allows us to derive parsimonious representations through the use of non-negative matrix factorization (NNMF) (Lee and Seung, 1999; Sotiras et al., 2015). NNMF was proposed as an analytical and interpretive tool in structural neuroimaging in Sotiras et al. (2015).

NNMF produces a factorization that constrains the elements of both the components and the loading coefficients matrix to be non-negative. This is achieved by minimizing the following energy:

$$\underset{\mathbf{B},\mathbf{C}}{\text{minimize}} \, ||\boldsymbol{X} - \mathbf{BC}||_F^2 \text{ subject to } \mathbf{B} \geq 0, \; \mathbf{C} \geq 0,$$

where $\mathbf{B} = [\mathbf{b}_1, \ldots, \mathbf{b}_K]$, $\mathbf{b}_i \in \mathbb{R}^D$, $\mathbf{C} = [\mathbf{c}_1, \ldots, \mathbf{c}_N], \mathbf{c}_i \in \mathbb{R}^K$, N is the number of images/samples, and K is the dimension of the estimated subspace. The non-negativity constraints lead to a sparse, parts-based representation (Lee and Seung, 1999). NNMF minimizes the reconstruction error by aggregating variance through positively weighting variables of the data matrix that tend to co-vary across the population. This provides a useful way of reducing the dimensionality of structural data. The structural data of each individual is approximated through an additive combination of the estimated components. In general, the estimated components identify regions that co-vary across individuals in a consistent way, thus forming patterns of structural co-variance that may potentially be parts of underlying networks or influenced by common mechanisms. The loading coefficients matrix $\mathbf{C}$ summarizes the integrity of each pattern of structural co-variance in each individual with a scalar value. These values provide an efficient and interpretable representation, and can be used for comparing the integrity of structural networks across individuals.

This method was applied in a cohort of normal aging adults and was compared against PCA and ICA in (Sotiras et al., 2015). It was shown to derive representations that are more parsimonious and coherent than the ones estimated by PCA and ICA. Moreover, the derived representation was quantitatively shown to be more relevant to age-related phenomena, while allowing for accurate age prediction as demonstrated through cross-validated age regression experiments. NNMF captured less of the variance in the data than PCA and ICA, resulting in higher reconstruction error. However, the high prediction accuracy suggests that the discarded information is not pertinent, leading to the conclusion that NNMF is able to retain important information while discarding irrelevant variations, which may potentially lead to increased generalizability. Indeed, split-sample experiments demonstrated that the non-negative components are more reproducible than the principal components.

Typical components estimated by NNMF are shown in Fig. 10.3. Note that the representation amounts to a soft clustering that segments the brain into structurally coherent units in a data-driven way by exploiting group statistics. The derived components are characterized by high spatial connectedness even though spatial smoothness was not explicitly enforced in the design of the method. Another important characteristic of the obtained representation is the symmetry of the

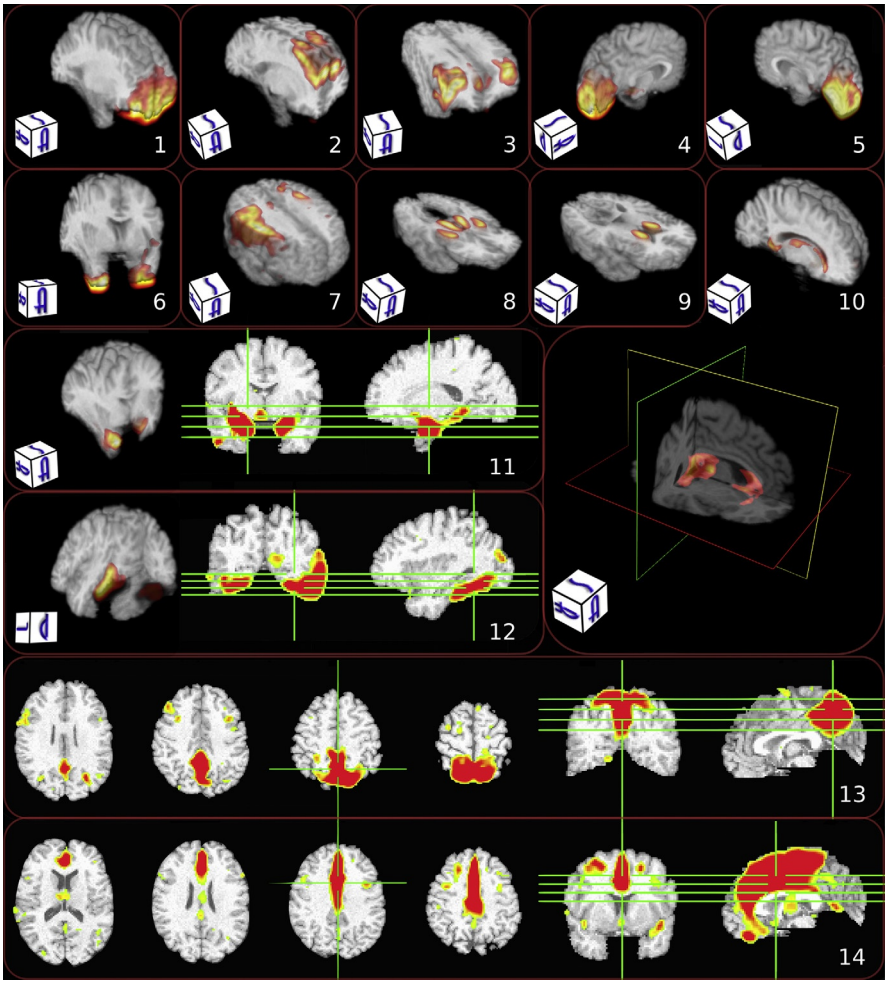

FIG. 10.3

Characteristic components estimated by NNMF. Different visualization strategies were used in order to enhance the visual perception of the components (note that the 2D images use radiographic convention). Warmer colors correspond to higher values. Note the alignment with anatomical regions: (1) prefrontal cortex; (2) superior frontal cortex; (3) superior lateral cortex; (4) left occipital lobe; (5) right occipital lobe; (6) inferior anterior temporal; (7) motor cortex; (8) thalamus and putamen; (9) head of caudate; (10) periventricular structures; (11) amygdala and hippocampus; (12) fusiform; (13) medial parietal including precuneus; (14) anterior and middle cingulate.

Source: Reprinted from Sotiras, A., Resnick, S.M., Davatzikos, C., 2015. Finding imaging patterns of structural covariance via Non-Negative Matrix Factorization. NeuroImage 108, 1–16.

estimated components. This symmetry is completely data-driven and it breaks when not supported by the group statistics. Lastly, and most importantly, the estimated components are not a solely statistical construct, but highly correspond to known structural and functional networks of the brain, or in some cases reflect underlying pathological processes.

10.2.3.2 Statistically driven dimensionality reduction of functional MRI

Resting-state functional MRI is typically used to analyze interactions between regions, aiming to reveal the brain's functional organization. Resting-state functional connectivity is used to reveal functional networks that can be found consistently in healthy populations by examining the connectivity between all pairs of regions in the brain. Pearson correlation is typically used to measure connectivity between different brain regions due to its simplicity and robustness (Smith et al., 2011; Lashkari et al., 2010). The resulting functional connectivity data is high dimensional and of mixed sign. The high dimensionality of the data makes subsequent group-wise analysis and interpretation of results difficult, underlining the need for an efficient and interpretable dimensionality reduction framework. However, the mixed sign nature of the data does not allow the application of the previously described non-negative framework. Instead, sparsity needs to be explicitly modeled through the inclusion of sparsity-inducing priors in the objective function of the matrix factorization framework.

A sparsity-based matrix factorization approach was proposed for functional connectivity data in Eavani et al. (2015). In this approach, each subject-specific correlation matrix $\mathbf{\Sigma}_n$ is approximated by a non-negative sum of sparse rank one matrices $\mathbf{b}_k\mathbf{b}_k^{\mathrm{T}}$. These sparse rank one matrices can be interpreted as functionally coherent subsets of brain regions, or sparse patterns of connectivity (SCPs), which occur in many of the subjects. A non-negative, subject-specific combination of SCPs, denoted by the set of coefficients $\mathbf{c}_n$, approximates the input correlation matrix $\mathbf{\Sigma}_n$:

$$\underset{\mathbf{B},\mathbf{C}}{\text{minimize}} \sum_{n=1}^{N} \left|\left|\mathbf{\Sigma}_n - \mathbf{B}\,\mathrm{diag}(\mathbf{c}_n)\,\mathbf{B}^{\mathrm{T}}\right|\right|_F^2$$

subject to

$$\begin{aligned} ||\mathbf{b}_k||_1 \leq \lambda, \quad & k = 1,\ldots,K, \\ -1 \leq \mathbf{b}_k(i) \leq 1, \quad \max_i |\mathbf{b}_k(i)| = 1, \quad & i = 1,\ldots,P, \\ \mathbf{c}_n \geq 0, \quad & n = 1,\ldots,N, \end{aligned}$$

where $\mathbf{B} = [\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_K]$. Sparse connectivity patterns (SCPs) provide a useful manner of reducing the dimensionality of the connectivity data, while summarizing the connectivity within each SCP in each individual with a scalar SCP coefficient value. These values can be used for comparing functional connectivity across individuals.

Applied to a normative sample of young adults, the resulting SCPs were shown to be reproducible across datasets, while explaining more of the variance in the second-order connectivity data when compared to spatial and temporal ICA (Calhoun et al.,

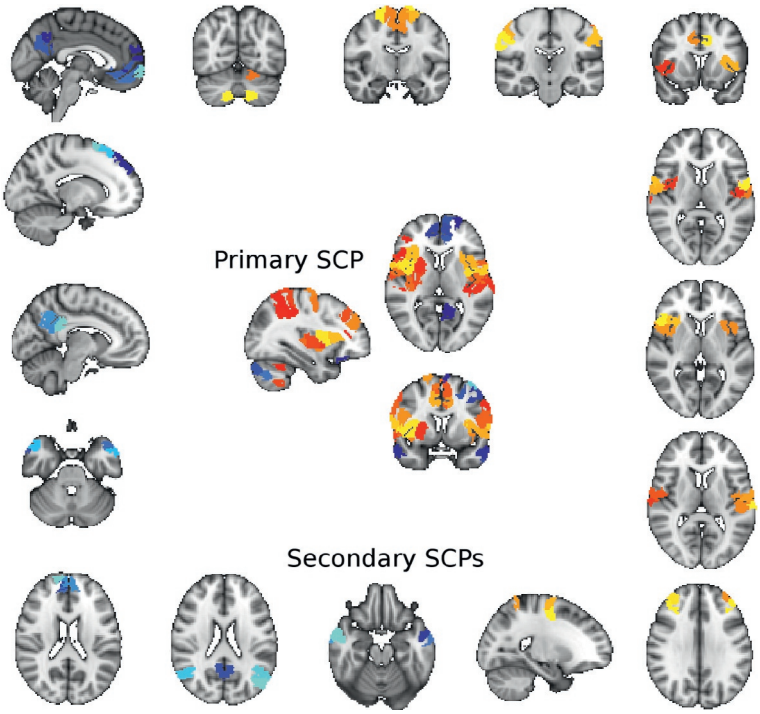

FIG. 10.4

Primary SCP (middle) showing the cingulum, operculum (red-yellow) and anticorrelated with the default mode (blue-light blue). Sixteen of its associated secondary SCPs are shown around it.

2003; Smith et al., 2012). This method can also be applied within a hierarchical framework, where each "primary" SCP with a large spatial extent can be split up into multiple smaller "secondary" SCPs, providing greater spatial specificity. Fig. 10.4 shows a large primary SCP with contributions from the operculum and anticorrelated with parts of the default mode. Its associated secondary SCPs, which represent a much smaller set of regions, are shown around it. Note the high specificity of the representation that is due to the sparsity of the derived networks.

10.3 MODEL INTERPRETATION: FROM CLASSIFICATION TO STATISTICAL SIGNIFICANCE MAPS

Once an appropriate set of features has been extracted, machine learning algorithms are employed to analyze neuroimaging data. This is typically performed by treating machine learning algorithms as "black-boxes" that are to able to integrate patterns

of disease-induced morphological signals into subject-specific indices. Even though these indices carry significant prognostic and diagnostic value, this usage paradigm does not fully exploit the potential of machine learning methods. In order to fully harness this potential, it is important to be able interpret the learned model in terms of identifying brain regions that significantly contribute to the construction of the discriminative pattern. This could significantly improve our understanding of the disease mechanisms that selectively influence-specific brain systems, while at the same time making the automated system transparent to human expert-driven verification.

In this section, we present such a framework for SVMs (Burges, 1998; Vapnik, 2000). SVMs are significantly popular in neuroimaging (Fan et al., 2007; Cuingnet et al., 2011; Klöppel et al., 2008; Gaonkar and Davatzikos, 2013; Batmanghelich et al., 2012; Varol et al., 2013), mainly due to their simplicity and the fact that the resulting problem is convex, allowing for efficient and globally optimal solutions. The SVM operates by constructing a hyperplane in a high-dimensional space that separates samples from two classes (eg, disease group vs. healthy controls) by the largest possible margin (see Fig. 10.5 for an illustration of the principle). The hyperplane coefficients denoted by $\mathbf{w}^*$ and b^* are estimated by solving the following optimization problem:

$$\{\mathbf{w}^*, b^*\} = \min_{\mathbf{w},b} \frac{1}{2}||\mathbf{w}||^2 + C\sum_{i=1}^{m}\xi_i$$
$$\text{such that } \; y_i(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b) \geq 1 - \xi_i \;\; \forall i = 1, \ldots, N$$
$$\xi_i \geq 0 \;\; \forall i = 1, \ldots, N,$$

where $\mathbf{x}_i \in \mathbb{R}^D$ denotes the vectorized image of the ith subject of the study, $y_i \in \{+1, -1\}$ denotes its respective binary label, ξ_i denotes the slack variable that accounts for the case where the classes are not separable, and C is a penalty parameter on the training error. The weight vector $\mathbf{w}^* \in \mathbb{R}^D$ describes the combination of all imaging elements that, along with the intercept b^*, best discriminates between the two classes.

It is tempting to use the weight image $\mathbf{w}^*$ to interpret the model by assigning more importance to elements that have higher weights. However, this is problematic (Haufe et al., 2014) and does not readily yield to a well-understood p-value-based statistical paradigm. One way to derive such a paradigm on the basis of SVM theory is to use permutation testing (see Fig. 10.6 for an illustration of the process). This is typically performed by generating a large number of shuffled instances of data labels by random permutations. Each shuffled instance is subsequently used for training one SVM, generating a new hyperplane parameterized by a vector $\mathbf{w}$. Thus for every element of $\mathbf{w}$, there is a set of possible values, each one corresponding to a specific shuffling of the labels. Collecting these values allows for the construction of the corresponding empirically obtained null distribution. Finally, comparing each component of $\mathbf{w}^*$ with the corresponding null distribution allows for the estimation of statistical significance. The number of permutations determines the minimal

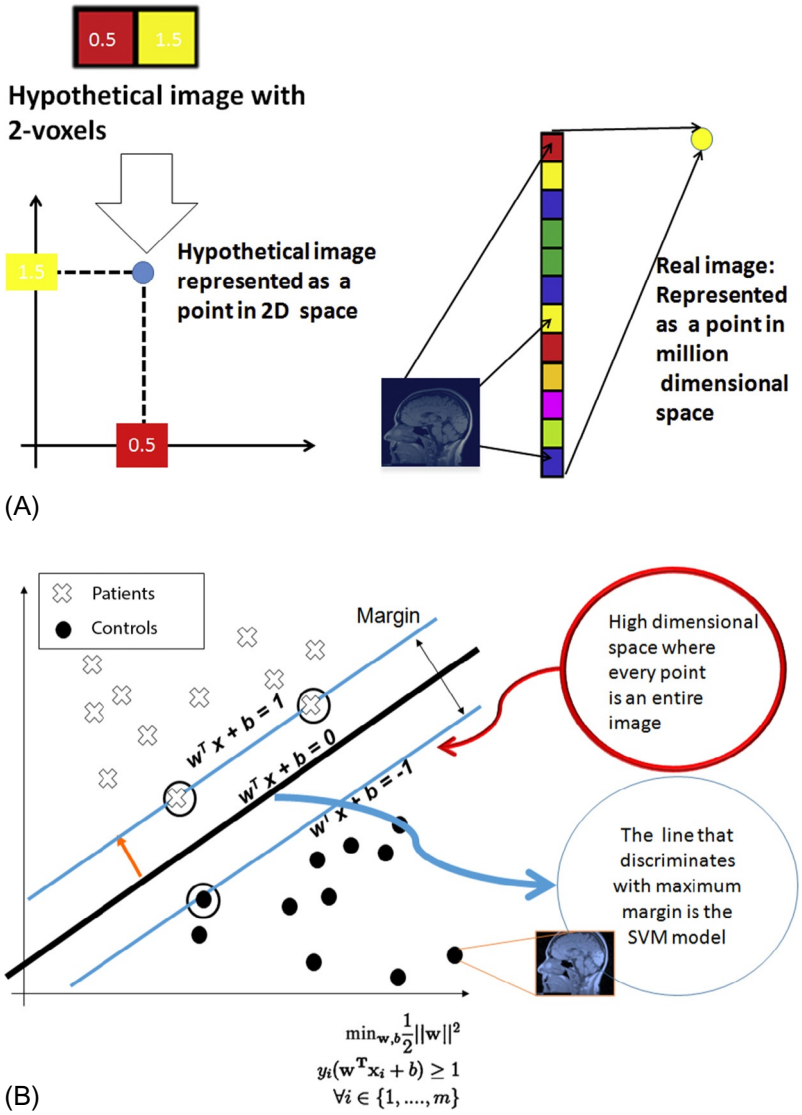

FIG. 10.5

The concept of imaging-based diagnosis using SVMs. (A) Images are treated as points located in a high-dimensional space. (B) The maximum margin principle of classification used in SVMs. Dots and crosses represent imaging scans taken from two groups. Even though the two groups cannot be separated on the basis of values along any single dimension, the combination of two dimensions gives perfect separation. This corresponds to the situation where a single anatomical region may not provide the necessary discriminative power between groups, whereas the multivariate SVM can still find the relevant hyperplane.

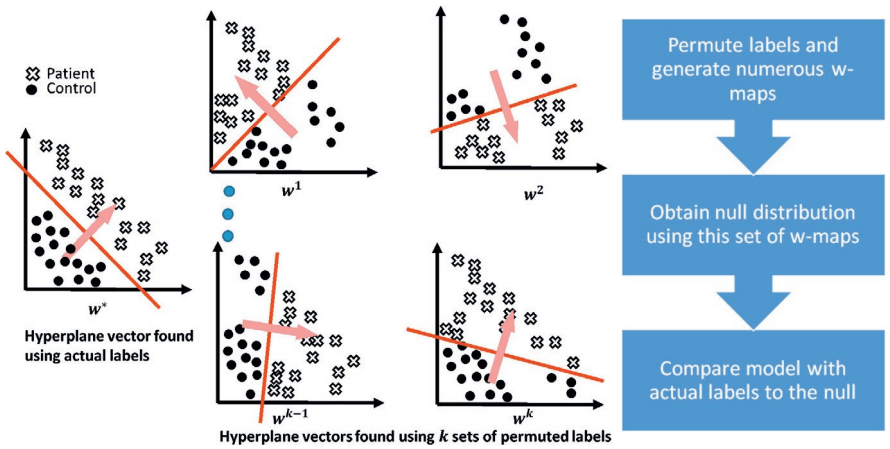

FIG. 10.6

Illustration of the permutation testing procedure.

Source: Reprinted from Gaonkar, B., Davatzikos, C., 2013. Analytic estimation of statistical significance maps for support vector machine based multi-variate image analysis and classification. NeuroImage 78, 270–283.

obtainable p-value as well as the resolution of the p-value. Increasing the number of permutations to a high number that will allow for the estimation of low p-values requires training a high number of support vector classifiers, which in turn requires a considerable amount of computational time and resources. Thus, a framework that would allow the analytic computation of the p-values in a computationally economic fashion would be of significant value.

Such a theoretical framework, that describes an analytic alternative to permutation testing, was introduced in Gaonkar and Davatzikos (2013) and Gaonkar et al. (2015). This analytical framework makes use of a certain set of simplifying assumptions that can be applied to the SVM formulations in high-dimensional spaces to derive an approximate null distribution, obviating the need for performing actual permutation testing. The first assumption regards the high-dimension, low-sample size setting that is typically encountered in medical imaging. In such a setting, it is always possible to find hyperplanes that can separate any possible labeling of points/samples. Thus when using linear SVMs, for any permutation of the labeling, one can always find a separating hyperplane that perfectly separates the training data. This allows us to use the hard margin SVM formulation. The second assumption regards the observation that, for most permutations, most data are support vectors. Taken together, these assumptions indicate that, for most permutations, it is possible to solve the following optimization problem:

$$\min_{\mathbf{w},b} \frac{1}{2}||\mathbf{w}||^2 \text{ such that } \mathbf{Xw} + \mathbf{J}b = \mathbf{y},$$

where $\mathbf{J}$ is a column matrix of ones, and $\mathbf{X}$ is a tall matrix with each row representing one image. Solving for $\mathbf{w}$ yields

$$\mathbf{w} = \underbrace{\mathbf{X}^{\mathrm{T}}\left[\left(\mathbf{X}\mathbf{X}^{\mathrm{T}}\right)^{-1} + \left(\mathbf{X}\mathbf{X}^{\mathrm{T}}\right)^{-1}\mathbf{J}\left(-\mathbf{J}^{\mathrm{T}}\left(\mathbf{X}\mathbf{X}^{\mathrm{T}}\right)^{-1}\mathbf{J}\right)^{-1}\mathbf{J}^{\mathrm{T}}\left(\mathbf{X}\mathbf{X}^{\mathrm{T}}\right)^{-1}\right]}_{=\mathbf{C}}\mathbf{y}.$$

Note that each element w_j of $\mathbf{w}$ is expressed as a linear combination of elements of $\mathbf{y}$. Thus, it is possible to hypothesize about the probability distribution of the elements of $\mathbf{w}$ given the distributions of y_i. If y_i attains any of the labels with equal probability, then $E(y_i) = 0$ and $Var(y_i) = 1$, which in turns lead to $E(w_j) = 0$ and $Var(w_j) = \sum_{i=1}^{N} C_{ij}^2$. At this point, there is an analytical method to approximate the mean and the variance of the null distributions of components w_j of $\mathbf{w}$. By taking advantage of the Lyapunov central limit theorem, it was demonstrated in Gaonkar and Davatzikos (2013) and Gaonkar et al. (2015) that the distribution of the individual components of $\mathbf{w}$ can be approximated using the normal distribution for a sufficiently large number of subjects. Thus, w_j^* computed by an SVM model using true labels can now simply be compared to the previous distribution and statistical inference can be made. The accuracy of this approximation is shown in Fig. 10.7. Note that the analytic and experimental p-maps are visually indistinguishable, while the scatter plot shows a good correspondence between the experimental and analytical p-values. Fig. 10.8 shows the regions that were identified by the method in Gaonkar et al. (2015) to be most statistically significant for classifying AD patients from controls. Note that the hippocampal complex, along with parahippocampal regions and amygdala, are clearly highlighted.

10.4 HETEROGENEITY

A common assumption behind automated group analysis methods applied in neuroimaging is that there is a single pattern that distinguishes the two contrasted groups. In other words, most approaches assume a single pathophysiological process that converts healthy controls to patients, and aim to reveal it through monistic analysis. However, this approach ignores ample evidence regarding the heterogeneous nature of diseases. For example, autism (Geschwind and Levitt, 2007; Jeste and Geschwind, 2014), schizophrenia (Buchanan and Carpenter, 1994; Koutsouleris et al., 2008; Zhang et al., 2015), Parkinson's disease (Graham and Sagar, 1999; Lewis et al., 2005), AD (Murray et al., 2011; Noh et al., 2014) or mild cognitive impairment (MCI) (Huang et al., 2003; Whitwell et al., 2007) are all characterized by clinical heterogeneity (see Fig. 10.9A for a graphical illustration of the problem).

Disentangling disease heterogeneity may greatly contribute to our understanding and lead to more accurate diagnosis, prognosis, and targeted treatment. We present here three recently proposed methods to tackle disease heterogeneity under different

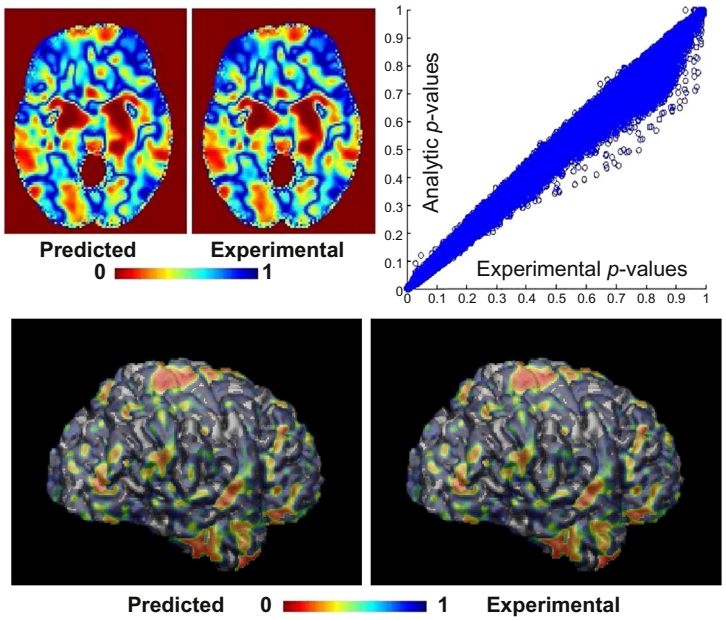

FIG. 10.7

(Top left) Analytic and experimental *p*-value maps thresholded at 0.01 are overlaid on the template brain. (Top right) A scatter plot of *p*-values comparing experimental and analytical *p*-values. (Bottom) A 3D rendering representing the predicted and experimental *p*-value maps.

Source: Reprinted from Gaonkar, B., Davatzikos, C., 2013. Analytic estimation of statistical significance maps for support vector machine based multi-variate image analysis and classification. NeuroImage 78, 270–283.

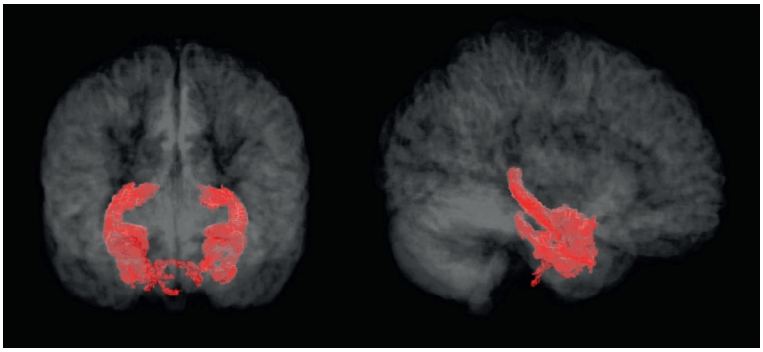

FIG. 10.8

3D views of the hippocampal and parahippocampal regions used by the SVM ($\alpha \leq 0.01$ FDR corrected).

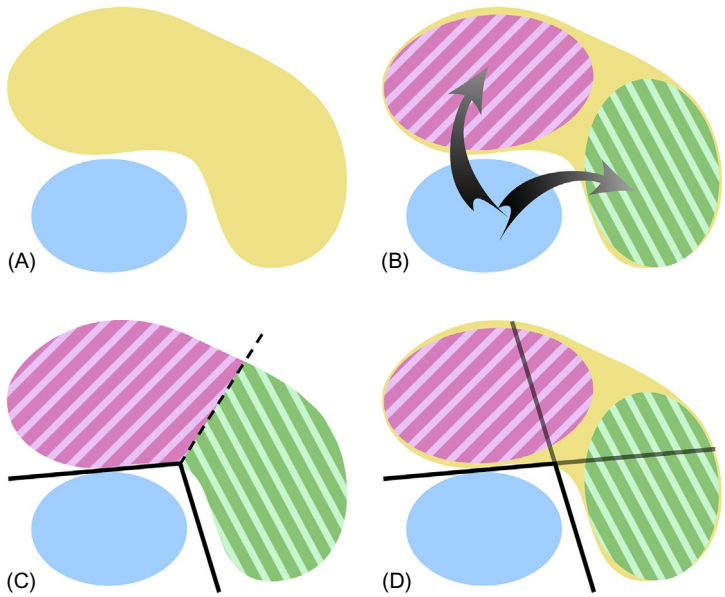

FIG. 10.9

Heterogeneity problem setting and different methods. (A) Problem setting. (B) CHIMERA. (C) HYDRA. (D) Mixture of Experts.

methodological assumptions. The first method is based on a generative clustering framework; the second adopts a purely discriminative approach, while the third combines discrimination and clustering.

10.4.1 GENERATIVE FRAMEWORK

The first method treats subjects as points in a high-dimensional feature space, where both the patient and the normal control group may be viewed as point distributions. In such a setting, the disease heterogeneity can be addressed by partitioning the patient distribution with a clustering method. However, directly clustering the patients would be driven by the distances between individuals, which would result in clustering the largest factor of data variability instead of the disease effect. In order to address this challenge, the generative approach proposed in Dong et al. (2016) considers the disease effect to be a transformation from the normal control distribution to the patient distribution (see Fig. 10.9b for a graphical illustration).

As a consequence, the patient distribution can be generated by transforming the normal control distribution with the assumption that if points of the patients had been spared from the disease, they would be covered by the normal control distribution. Heterogeneous disease effects are modeled by considering multiple distinct transformations. These transformations can be found by solving

for a distribution matching of the true patient and generated patient distributions. The distribution matching takes into account both imaging and covariate features (known variables, such as age, sex, and height). In this way, the clustering of patient distribution is regularized by the structure of the normal control distribution.

More formally, let us assume that there are M normal control subjects, $\mathbf{X} = \{x_1, ..., x_M\}$, and N patient subjects, $\mathbf{Y} = \{y_1, ..., y_N\}$. They are described by two sets of features: a set of D_1-dimensional imaging features, $x_m^v, y_n^v \in \mathbb{R}^{D_1}$; and a set of D_2-dimensional covariate features, $x_m^c, y_n^c \in \mathbb{R}^{D_2}$. For simplicity, subjects are denoted in compact vector forms: $x_m = (x_m^v, x_m^c)$, $y_n = (y_n^v, y_n^c)$. The clustering model minimizes the following energy $\mathcal{E}$:

$$\mathcal{E}(\mathbf{X}, \mathbf{Y}, \Theta) = -\mathcal{L}(\mathbf{X}, \mathbf{Y}, \Theta) + \mathcal{R}(\Theta),$$

where Θ denotes the parameters of the model, such as transformations that are applied to $\mathbf{X}$ in order to generate $\mathbf{Y}$; $\mathcal{L}$ is the log-likelihood of the distributions $\mathbf{X}$ and $\mathbf{Y}$ given the parameters; and $\mathcal{R}$ is a regularization term aiming to improve the stability of the clustering results.

The distribution transformation is denoted as $\mathbf{T}$, which is a convex combination of K linear transformations, each one corresponding to a different disease effect. $\mathbf{T}$ maps the imaging feature x_m of a normal control sample to the patient distribution, while keeping its covariate feature unchanged: $\mathbf{T}(x_m) = (\sum_{k=1}^{K} \zeta_{km}(A_k x_m^v + b_k), x_m^c)$. The distribution matching is conducted as a variant of the coherent point drift algorithm Myronenko and Song (2010). Each transformed normal control point is considered as a centroid of a spherical Gaussian cluster, and patient points are treated as independent and identically distributed data generated by a Gaussian Mixture Model (GMM) with equal weights for each cluster. The data likelihood of this mixture model is optimized during the distribution matching, where covariate features are embedded in the distance between points with a multikernel setting. These model assumptions lead to the log-likelihood term $\mathcal{L}$ being:

$$\mathcal{L}(\mathbf{X}, \mathbf{Y}, \Theta) = \sum_{n=1}^{N} \log \sum_{m=1}^{M} \frac{1}{M} \frac{r^{D_2/2}}{(\sqrt{2\pi}\sigma)^{D_1+D_2}} \exp\left\{ \frac{\|y_n^v - \sum_{k=1}^{K} \zeta_{km}(A_k x_m^v + b_k)\|^2 + r\|y_n^c - x_m^c\|^2}{-2\sigma^2} \right\}.$$

The Frobenius norm of $A_k - \mathbf{I}$ and the ℓ_2 norm of b_k are to be regularized, where $\mathbf{I}$ is an identity matrix. This regularization is equivalent to posing Gaussian priors for the parameters:

$$\mathcal{R}(\Theta) = \frac{\lambda_1}{2\sigma^2} \sum_k \|b_k\|_2^2 + \frac{\lambda_2}{2\sigma^2} \sum_k \|A_k - \mathbf{I}\|_F^2$$

The energy objective $\mathcal{E}$ is optimized with an expectation-maximization Moon (1996) approach. The heterogeneous disease subgroups of patients are further clustered by the estimated transformations.

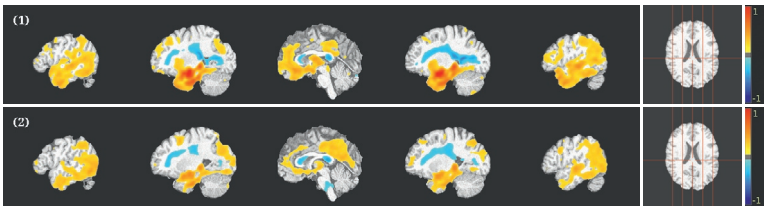

FIG. 10.10

VBM performed on gray matter RAVENS (Davatzikos, 1998) maps between (1) Subgroup 1 and Control group; (2) Subgroup 2 and Control group. Group comparison results are overlaid on the registration template image. Regions that are significant under a corrected threshold of FDR = 0.01 are shown. Color maps indicate the scale of the *t*-statistic. Warmer colors indicate volume loss, while colder colors indicate volume increase.

This method was applied to an AD dataset[1] comprising 390 T1 structural MRI scans with 177 AD patients and 213 normal controls. Multi-Atlas ROI volumes were generated and used as imaging features, while age and sex information were used as covariate features. With the cross-validated parameters, two subgroups were discovered. Voxel-based morphometry (VBM) Ashburner and Friston (2000) was employed to examine the differences between the estimated subgroups and the control population. The VBM results obtained from gray matter group comparisons are shown in Fig. 10.10. Subgroup 1 has more gray matter atrophy in limbic lobe and frontal insular regions, while it exhibits unique deep gray matter atrophy in basal ganglia. Subgroup 2 exhibits unique parietal and occipital gray matter atrophy on both lateral and medial structures.

10.4.2 DISCRIMINATIVE FRAMEWORK

The second method takes a purely discriminative approach. It is based upon the observation that, in high-dimensional spaces, the modeling capacity of linear SVMs is theoretically rich enough to discriminate between two homogeneous classes. However, while two classes may be linearly separable with high probability, the resulting margin could be small. This case arises, for example, when one class is generated by a multimodal distribution that models a heterogeneous process. This may be remedied by the use of nonlinear classifiers, allowing for larger margins and thus, better generalization. However, while kernel methods, such as Gaussian kernel SVM, provide nonlinearity, they lack interpretability when aiming to characterize heterogeneity.

[1] http://adni.loni.usc.edu/

In order to tackle the aforementioned limitations, a novel maximum margin nonlinear learning algorithm for simultaneous binary classification and subtype identification, termed HYDRA (HeterogeneitY through DiscRiminative Analysis) was introduced in Varol et al. (2015) and Varol et al. (2016). HYDRA aims to tackle disease subtype discovery in a principled machine learning framework. Neuroanatomical or genetic subtypes are effectively captured by multiple linear hyperplanes, which form a convex polytope that separates two groups (eg, healthy controls from pathologic samples); each face of this polytope effectively defines a disease subtype (see Fig. 10.9c for a graphical illustration).

More formally, let us assume an imaging (or genetic) dataset consisting of n binary labeled d-dimensional data points ($\mathcal{D} = (\mathbf{x}_i, y_i)_{i=1}^{n}, \mathbf{x}_i \in \mathbb{R}^d$ and $y_i \in \{-1, 1\}$). The maximum margin polytope that separates the assumed heterogenous patients from the controls can be solved by optimizing the following objective:

$$\min_{\substack{\{\mathbf{w}_j, b_j\}_{j=1}^{K} \\ \{s_{i,j}\}_{i,j}^{n-,K}}} \sum_{j=1}^{K} \frac{\|\mathbf{w}_j\|_2^2}{2} + C \sum_{\substack{i|y_i=+1 \\ j}} \frac{1}{K} \max\{0, 1 - \mathbf{w}_j^{\mathrm{T}} \mathbf{x}_i - b_j\} + C \sum_{\substack{i|y_i=-1 \\ j}} s_{i,j} \max\{0, 1 + \mathbf{w}_j^{\mathrm{T}} \mathbf{x}_i + b_j\}.$$

The first term encourages maximum average margin across all K faces of the convex polytope classifier. The second term forces the control samples to be confined **inside** the polytope with slack. Lastly, the third term enforces the patient samples to lie **outside** the assigned face of the polytope with slack. The assignment of patient samples to the faces of the polytope is handled by the indicator variable $s_{i,j}$, which can be estimated by solving a linear program. The objective is optimized by following a two-step procedure that iterates between assigning samples to faces of the polytope, and solving for hyperplanes that maximize the overall margin. This is similar in spirit to unsupervised clustering methods, such as K-means, where centroids and assignments are iteratively solved.

This approach was applied to a genetic dataset comprising 53 AD patients and 68 cognitively normal (CN) older adults (see demographic information in Table 10.1), obtained from the ADNI study[2]. ADNI genotyping is performed using the Human610-Quad Bead-Chip (Illumina, Inc., San Diego, CA), which results in a set of 620,901 single nucleotide polymorphisms (SNPs) and copy number variation markers. Due to the weak, or spurious, signal in most of the genome, the features were pruned and only SNP loci that were found to be associated with AD in a recent large-scale genome-wide association study (Lambert et al., 2013) were kept. This resulted in a reduced set of 18 SNPs that were represented by using two binary variables that encode the presence of major-major or major-minor alleles, thus raising the total number of features to 36.

In order to estimate the optimal number of clusters, a reproducibility analysis was performed. The reproducibility of the clustering was evaluated at $K = 1, \ldots, 9$ by using the Adjusted Rand Index (Hubert and Arabie, 1985). This analysis suggested that

[2] http://adni.loni.usc.edu/data-samples/genetic-data/

Table 10.1 Demographic and Clinical Characteristics of Healthy Controls, AD Patients (Left) and the Estimated Genetic-Driven Subtypes of AD (Right)

Genetic Heterogeneity in Alzheimer's Disease						
	AD vs. CN ($n = 121$)			**AD Subgroups ($n = 53$)**		
	CN ($n = 68$)	AD ($n = 53$)	p-value[b]	Group 1 ($n = 34$)	Group 2 ($n = 19$)	p-value[c]
Age (years)	76.08 ± 4.672	76.08 ± 7.188	0.9944	75.27 ± 5.981	77.43 ± 8.872	0.3184
Sex (female), n (%)	33 (50)	25 (52.08)	0.828	15 (50)	10 (55.56)	0.7163
MMSE	28.44 ± 2.367	19.06 ± 5.05	1.228e-24	18.77 ± 5.71	19.56 ± 3.807	0.6057
Apoε-4 genotype[a], n (%)	20 (30.3)	31 (64.58)	0.0002108	29 (96.67)	2 (11.11)	1.901e-15

[a]Denotes subjects with at least one Apoε-4 allele present. [b]p-value estimated using two-tailed t-test to compare AD with CN. [c] p-value estimated using analysis of variance (ANOVA) to compare the two estimated AD subgroups.

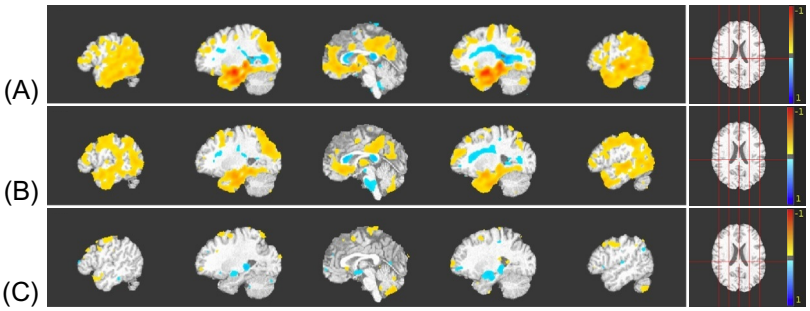

FIG. 10.11

The anatomic differences between the two genetic subtypes of AD: Axial views of gray matter group comparisons of (A) Controls vs. first AD subgroup; (B) Controls vs. second AD subgroup; and (C) first AD subgroup vs. second AD subgroup are visualized. For (A) and (B), colder colors indicate relative GM volume increases (CN < AD subgroups), while warmer colors correspond to relative GM volume decreases (CN > AD subgroups). Similarly for (C), colder colors indicate relative GM volume increases (first AD subgroup < second AD subgroup), while warmer colors correspond to relative GM volume decreases (first AD subgroup > second AD subgroup). Both groups exhibit atrophy in the temporal lobe and posterior medial cortex, while white matter lesions are present in the periventricular area. However, the first AD subgroup, which mainly comprises Apo-ε4 carriers, is characterized by significantly more hippocampal and entorhinal cortex atrophy.

two clusters were appropriate for capturing the intrinsic dimensionality of the genetic heterogeneity associated with AD. The optimal genotype clustering is visualized by contrasting the imaging phenotypes of the estimated subgroups against the healthy control population through morphometric analysis using RAVENS (see Fig. 10.11A and B). Correction for multiple comparisons was performed by controlling for False Discovery Rate (FDR). The results were thresholded at $q < 0.05$. It can be observed that at the $K = 2$ cluster level (see Fig. 10.11), the estimated subgroups were associated with distinct patterns of structural brain alterations. The first subgroup had increased temporal lobe atrophy (see Fig. 10.11A), including focal atrophy in the hippocampus and entorhinal cortex, as well as increased white matter lesion load. The second subgroup was characterized by diffuse temporal lobe atrophy (see Fig. 10.11B), including periventricular white matter lesions.

In summary, HYDRA seamlessly integrates clustering and discrimination in a coherent framework by solving a piecewise linear classifier that bears common geometric properties with convex polytopes. Discrimination is achieved by constraining one class in the interior of the polytope, while at the same time maximizing the margin between examples and class boundary. On the other hand, clustering is performed by associating disease samples with different faces of the polytope, and hence to different disease processes. Thus, each face of the polytope informs

us about the distinct foci of disease effects that distinguish the patients from the healthy control subjects. This coupling between clustering and classification allows for segregating patients based on disease effects rather than global anatomy.

10.4.3 GENERATIVE DISCRIMINATIVE FRAMEWORK

The last approach that aims to identify heterogeneous subgroups in patient populations is based upon a mixture-of-experts (MOE) framework. The MOE framework was initially proposed for vowel discrimination within speech recognition (Jacobs et al., 1991) and, later on, as a fast and efficient alternative to "kernel" SVMs (Ladicky and Torr, 2011; Fu et al., 2010). While kernel SVMs can successfully model nonlinear separation boundaries between groups, they suffer from a major limitation in neuroimaging applications, namely the lack of interpretability of the results. In a kernel-based method, the data is projected into a higher dimensional space prior to being classified and the nonlinear separating boundary in the original feature space is not explicitly computed.

The presented joint generative-discriminative approach tackles this shortcoming by combining a generative clustering model with a discriminative classification/regression model (Eavani et al., 2016). Using this combination of unsupervised clustering (mixture) with supervised classification/regression (expert), it approximates the nonlinear boundary that separates the two classes with a piecewise linear separating boundary, providing us the identification of the subgroups as well as the multivariate patterns that discriminate each subgroup from the reference group (see Fig. 10.9d for a graphical illustration). The data is modeled using a mixture of distributions, such as fuzzy c-means, which assigns a soft subgroup membership to each subject in the affected group. The linear boundary between each affected subgroup and the reference group can be found using a linear classifier, such as a linear SVM.

This is a general framework that can be applied to any dataset, using any appropriate mixture model and expert classifier. Using a combination of fuzzy c-means and ℓ_2-loss linear SVMs, Eavani et al. (2016) found heterogeneity in the manner in which normal older individuals age in terms of functional connectivity. Of the two subgroups that were found within the older individuals (relative to a reference group of younger individuals), the authors found that one set of individuals had increased functional connectivity between the bilateral frontal and insula regions. Upon further investigation, the same set of individuals were found to have specific cognitive abilities (executive function and visual processing) comparable to that of the younger group, while the rest had worse cognition than the younger group, as expected due to aging. It is possible that the increased bilateral connectivity in the subset of older people acts as a compensatory mechanism, resulting in better cognitive performance for their age (Fig. 10.12).

These results produced using MOE have significant clinical implications in terms of identifying functional bio-markers of resilient aging, which is a very active topic of research in brain aging. These results provide important biological clues to the wide variation in cognitive performance that is normally seen in older individuals.

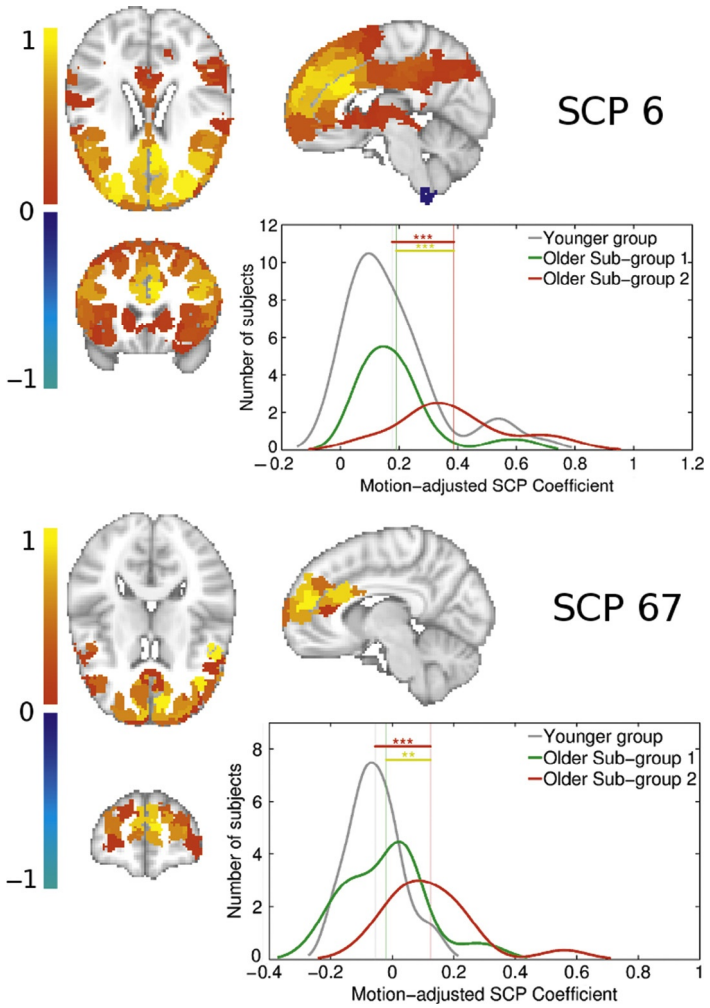

FIG. 10.12

Plot showing primary SCP 6, and its associated secondary SCP 67, whose average connectivity is increased in the second older subgroup, but not the first. SCP 6 highlights most of the prefrontal cortex. SCP 67 captures the bilateral paracingulate gyrus and inferior temporal gyrus. The distribution fit of the underlying SCP coefficient histograms are also shown, for each SCP and for each subgroup. Significance levels are indicated as follows: $***p$-value < 0.001, $**p$-value < 0.01, and $*p$-value < 0.05.

Source: Reprinted from Eavani, H., Hsieh, M.K., An, Y., Erus, G., Beason-Held, L., Resnick, S., Davatzikos, C., 2016. Capturing heterogeneous group differences using mixture-of-experts: application to a study of aging. NeuroImage 125, 498–514.

10.5 APPLICATIONS

In this section, we present applications of machine learning tools towards tackling clinically relevant problems.

10.5.1 INDIVIDUALIZED DIAGNOSTIC INDICES USING MRI

The past 20 years have seen a wide acceptance of pattern analysis methods in neuroimaging as a means for capturing spatial patterns of morphological, functional, and pathologic signals. However, the vast majority of methods investigating disease effects on the brain have relied on voxel-based analysis (VBA) methods, which apply mass-univariate tests on a voxel-by-voxel basis in an attempt to elucidate the spatial patterns of imaging differences between patients and healthy controls. During the past decade, the use of machine learning to integrate and synthesize these patterns into indices of diagnostic and predictive value for each individual has gained a great deal of attention. This is due to its significance beyond understanding disease effects and into deriving individualized clinical indices of disease. Such machine learning-derived indices have been used in several diseases, including AD (Davatzikos et al., 2008a; Klöppel et al., 2008) and schizophrenia (Davatzikos et al., 2005). We now summarize our group's work on deriving the SPARE-AD index, an index that measures the presence of AD-like patterns of brain atrophy from brain MRI.

10.5.2 MRI-BASED DIAGNOSIS OF AD: THE SPARE-AD

In Fan et al. (2008a), the COMPARE algorithm was used on 122 MRI scans of cognitively normal (CN) older adults and AD patients, and the SPARE-AD index was derived: positive values reflect the presence of AD-like patterns of brain atrophy, and negative values indicate CN-like brain anatomy. The patterns used by the COMPARE algorithm to build the SPARE-AD score were fairly complex and distributed over several brain regions of gray matter (GM), white matter (WM), and cerebrospinal fluid (CSF). Fig. 10.13 indicates the regions with the most significant brain atrophy and ventricular expansion.

The histograms of the (cross-validated) SPARE-AD scores achieved in this classification are shown in Fig. 10.14, indicating excellent discrimination between CN individuals and AD patients. The SPARE-AD index is therefore an index that offers promise as a clinical score derived from sMRI and measuring the presence of AD patterns of brain atrophy.

10.5.3 INDIVIDUALIZED EARLY PREDICTIONS

As individualized diagnostic indices, like the SPARE-AD, are developed based on machine learning approaches, it is perhaps of greater interest to evaluate the predictive value of these indices at early disease stages or even preclinically. These are the stages where standard clinical evaluations might be less effective and hence

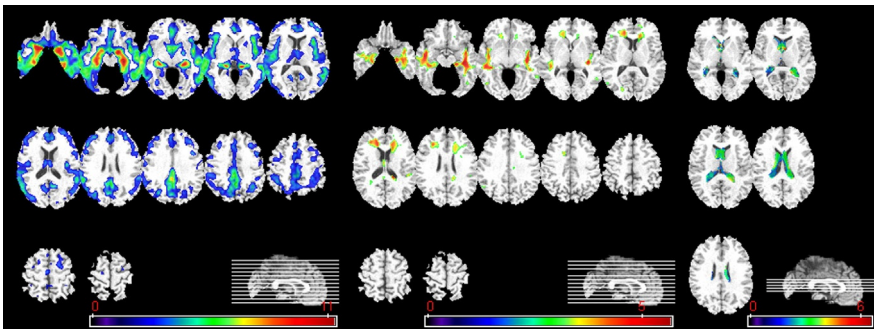

FIG. 10.13

From left to right, group comparison results on GM, WM, and CSF are shown. The color maps indicate the scale for the *t*-statistic. Images are displayed in radiological convention.

Source: Reprinted with permission from Fan, Y., Batmanghelich, N., Clark, C.M., Davatzikos, C., 2008a. Spatial patterns of brain atrophy in MCI patients, identified via high-dimensional pattern classification, predict subsequent cognitive decline. NeuroImage 39(4), 1731–1743.

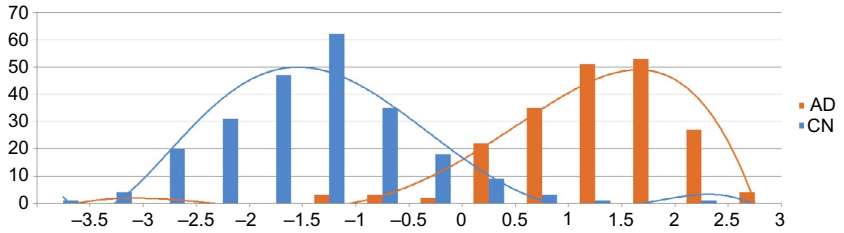

FIG. 10.14

Histograms of SPARE-AD scores obtained via cross-validation from the ADNI1 sample comprising CN and AD individuals.

likely to benefit from imaging-based biomarkers. In this vein, the SPARE-AD index was examined in individuals with mild cognitive impairment in Da et al. (2014) and Davatzikos et al. (2011), and it was found to predict, to a large extent, an individual's future progression to dementia. Fig. 10.15 shows survival curves obtained from baseline measures in 432 MCI patients of the ADNI1 study.

Looking at even earlier stages of the progression of patterns of brain atrophy evaluated via machine learning, the study in Davatzikos et al. (2009) investigated the predictive value of SPARE-AD in preclinical stages of cognitively normal aging. It was found that patterns of brain change at those stages are quite predictive of future cognitive decline. Fig. 10.16 shows the rates of SPARE-AD change for people who remained cognitively stable (left), and for people who progressed to MCI over an 8-year period; since conversion from MCI to AD also takes additional time

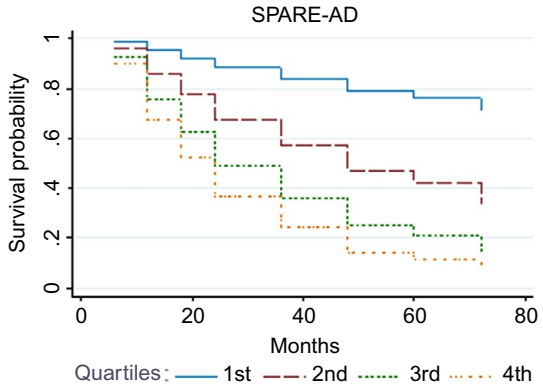

FIG. 10.15

Survival curves showing the predictive value of MRI-derived patterns of atrophy that were evaluated using machine learning (the SPARE-AD index).

Source: Reprinted with permission from Da, X., Toledo, J.B., Zee, J., Wolk, D.A., Xie, S.X., Ou, Y., Shacklett, A., Parmpi, P., Shaw, L., Trojanowski, J.Q., Davatzikos, C., 2014. Integration and relative value of biomarkers for prediction of MCI to AD progression: spatial patterns of brain atrophy, cognitive scores, APOE genotype and CSF biomarkers. NeuroImage: Clinical 4, 164–173.

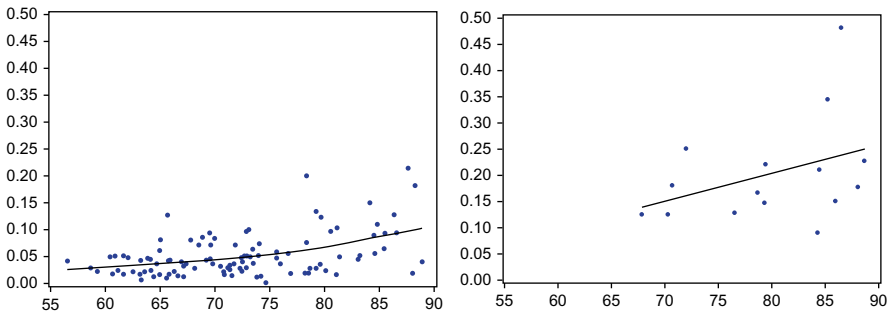

FIG. 10.16

Annual rates of SPARE-AD change at the Baltimore Longitudinal Study of Aging (BLSA). People who remained stable are shown on the left, and people who converted to MCI are shown on the right, displaying markedly higher rates of SPARE-AD change prior to cognitive decline.

Source: Adapted with permission from Davatzikos, C., Xu, F., An, Y., Fan, Y., Resnick, S.M., 2009. Longitudinal progression of Alzheimer's-like patterns of atrophy in normal older adults: the SPARE-AD index. Brain 132 (Pt 8), 2026–2035.

(conversion rate is about 15% annually), these studies indicate that patterns of brain atrophy captured by these machine learning approaches can evolve a decade or longer before dementia. The availability of such an early time window can prove critical for the success of future treatments.

10.6 CONCLUSION

In summary, machine learning approaches offer great promise in clinical research as a means for integrating complex imaging data into personalized indices of diagnostic and prognostic value. As imaging (and genomic) data becomes increasingly complex and multifaceted, such approaches promise to help reduce otherwise unmanageable data volumes down to relatively few clinically informed indices. One of the challenges faced ahead is the need to prove the generalization of these approaches in large samples of data obtained across different studies, scanners, or sites. This can be particularly challenging, in part due to the very ability of these methods to find subtle patterns. If these patterns become too specific to one type of data, then they might be less likely to generalize well across different clinics. Good imaging harmonization across clinics is essential, as is the need to regularize and cross-test machine learning methods sufficiently, to avoid data overfitting.

REFERENCES

Ashburner, J., Friston, K.J., 2000. Voxel-based morphometry—the methods. NeuroImage 11 (6), 805–821.

Atlas, S.W., 2009. Magnetic Resonance Imaging of the Brain and Spine. Lippincott Williams & Wilkins, Baltimore, 2256 .

Avants, B.B., Cook, P.A., Ungar, L., Gee, J.C., Grossman, M., 2010. Dementia induces correlated reductions in white matter integrity and cortical thickness: a multivariate neuroimaging study with sparse canonical correlation analysis. NeuroImage 50 (3), 1004–1016.

Batmanghelich, N.K., Taskar, B., Davatzikos, C., 2012. Generative-discriminative basis learning for medical imaging. IEEE Trans. Med. Imaging 31 (1), 51–69.

Baxton, R.B., 2009. Introduction to Functional Magnetic Resonance Imaging. Cambridge University Press, Cambridge, UK.

Beckmann, C.F., Smith, S.M., 2004. Probabilistic independent component analysis for functional magnetic resonance imaging. IEEE Trans. Med. Imaging 23 (2), 137–152.

Bellec, P., Rosa-Neto, P., Lyttelton, O.C., Benali, H., Evans, A.C., 2010. Multi-level bootstrap analysis of stable clusters in resting-state fMRI. NeuroImage 51, 1126–1139.

Biswal, B.B., 2012. Resting state fMRI: a personal history. NeuroImage 62 (2), 938–944.

Blumensath, T., Jbabdi, S., Glasser, M.F., Van Essen, D.C., Ugurbil, K., Behrens, T.E.J., Smith, S.M., 2013. Spatially constrained hierarchical parcellation of the brain with resting-state fMRI. NeuroImage 76, 313–324.

Buchanan, R.W., Carpenter, W.T., 1994. Domains of psychopathology: an approach to the reduction of heterogeneity in schizophrenia. J. Nerv. Ment. Dis. 182 (4), 193–204.

Burges, C.J., 1998. A tutorial on support vector machines for pattern recognition. Data Min. Knowl. Disc. 2 (2), 121–167.

Calhoun, V.D., Adali, and Pearlson, G.D., Pekar, J.J., 2001. A method for making group inferences from functional MRI data using independent component analysis. Hum. Brain Map. 14, 140–151.

Calhoun, V.D., Adali, T., Hansen, L.K., Larsen, J., Pekar, J.J., 2003. ICA of functional MRI data: an overview. In: Proceedings of the International Workshop on Independent Component Analysis and Blind Signal Separation, pp. 281–288.

Cordes, D., Haughton, V., Carew, J.D., Arfanakis, K., Maravilla, K., 2002. Hierarchical clustering to measure connectivity in fMRI resting-state data. Mag. Reson. Imaging 20, 305–317.

Craddock, R.C., James, G.A., Holtzheimer, P.E.I., Hu, X.P., Mayberg, H.S., 2012. A whole brain fMRI atlas generated via spatially constrained spectral clustering. Hum. Brain Map. 33 (8), 1914–1928.

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.

Da, X., Toledo, J.B., Zee, J., Wolk, D.A., Xie, S.X., Ou, Y., Shacklett, A., Parmpi, P., Shaw, L., Trojanowski, J.Q., Davatzikos, C., 2014. Integration and relative value of biomarkers for prediction of MCI to AD progression: spatial patterns of brain atrophy, cognitive scores, APOE genotype and CSF biomarkers. NeuroImage: Clinical 4, 164–173.

Daubechies, I., Roussos, E., Takerkart, S., Benharrosh, M., Golden, C., Ardenne, K.D., Richter, W., Cohen, J.D., Haxby, J., 2009. Independent component analysis for brain fMRI does not select for independence. Proc. Natl. Acad. Sci. USA 106 (26), 10415–10422.

Davatzikos, C., 1998. Mapping image data to stereotaxic spaces: applications to brain mapping. Hum. Brain Map. 6 (5-6), 334–338.

Davatzikos, C., Shen, D., Gur, R.C., Wu, X., Liu, D., Fan, Y., Hughett, P., Turetsky, B.I., Gur, R.E., 2005. Whole-brain morphometric study of schizophrenia revealing a spatially complex set of focal abnormalities. Arch. Gen. Psychiat. 62 (11), 1218–1227.

Davatzikos, C., Fan, Y., Wu, X., Shen, D., Resnick, S.M., 2008a. Detection of prodromal Alzheimer's disease via pattern classification of magnetic resonance imaging. Neurobiol. Aging 29 (4), 514–523.

Davatzikos, C., Resnick, S.M., Wu, X., Parmpi, P., Clark, C.M., 2008b. Individual patient diagnosis of AD and FTD via high-dimensional pattern classification of MRI. NeuroImage 41 (4), 1220–1227.

Davatzikos, C., Xu, F., An, Y., Fan, Y., Resnick, S.M., 2009. Longitudinal progression of Alzheimer's-like patterns of atrophy in normal older adults: the SPARE-AD index. Brain 132 (Pt 8), 2026–2035.

Davatzikos, C., Bhatt, P., Shaw, L.M., Batmanghelich, K.N., Trojanowski, J.Q., 2011. Prediction of MCI to AD conversion, via MRI, CSF biomarkers, and pattern classification. Neurobiol. Aging 32 (12), 2322.e19–2322.e27.

Delong, A., Osokin, A., Isack, H.N., Boykov, Y., 2012. Fast approximate energy minimization with label costs. Int. J. Comput. Vision 96, 1–27.

Detre, J.A., Leigh, J.S., Williams, D.S., Koretsky, A.P., 1992. Perfusion imaging. Magn. Reson. Med. 23 (1), 37–45.

Dong, A., Honnorat, N., Gaonkar, B., Davatzikos, C., 2016. CHIMERA: clustering of heterogeneous disease effects via distribution matching of imaging patterns. IEEE Trans. Med. Imaging 35 (2), 612–621.

Duchesne, S., Caroli, a., Geroldi, C., Barillot, C., Frisoni, G.B., Collins, D.L., 2008. MRI-based automated computer classification of probable AD versus normal controls. IEEE Trans. Med. Imaging 27 (4), 509–520.

Eavani, H., Satterthwaite, T.D., Filipovych, R., Gur, R.E., Gur, R.C., Davatzikos, C., 2015. Identifying sparse connectivity patterns in the brain using resting-state fMRI. NeuroImage 105, 286–299.

Eavani, H., Hsieh, M.K., An, Y., Erus, G., Beason-Held, L., Resnick, S., Davatzikos, C., 2016. Capturing heterogeneous group differences using mixture-of-experts: application to a study of aging. NeuroImage 125, 498–514.

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. Imaging 26 (1), 93–105.

Fan, Y., Batmanghelich, N., Clark, C.M., Davatzikos, C., 2008a. 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., Resnick, S.M., Wu, X., Davatzikos, C., 2008b, Structural and functional biomarkers of prodromal Alzheimer's disease: a high-dimensional pattern classification study. NeuroImage 41 (2), 277–285.

Friston, K.J., Frith, C.D., Liddle, P.F., Frackowiak, R.S., 1993. Functional connectivity: the principal-component analysis of large (PET) data sets. J. Cerebral Blood Flow Metabol. 13 (1), 5–14.

Fu, Z., Robles-Kelly, A., Zhou, J., 2010. Mixing linear SVMS for nonlinear classification. IEEE Trans. Neural Netw. 21 (12), 1963–1975.

Gaonkar, B., Davatzikos, C., 2013. Analytic estimation of statistical significance maps for support vector machine based multi-variate image analysis and classification. NeuroImage 78, 270–283.

Gaonkar, B., Shinohara, R.T., Davatzikos, C., Initiative, A.D.N., et al., 2015. Interpreting support vector machine models for multivariate group wise analysis in neuroimaging. Med. Image Anal. 24 (1), 190–204.

Gerardin, E., Chételat, G., Chupin, M., Cuingnet, R., Desgranges, B., Kim, H.S., Niethammer, M., Dubois, B., Lehéricy, S., Garnero, L., Eustache, F., Colliot, O., 2009. Multidimensional classification of hippocampal shape features discriminates Alzheimer's disease and mild cognitive impairment from normal aging. NeuroImage 47 (4), 1476–1486.

Geschwind, D.H., Levitt, P., 2007. Autism spectrum disorders: developmental disconnection syndromes. Curr. Opin. Neurobiol. 17 (1), 103–111.

Golland, P., Golland, Y., Malach, R., 2007. Detection of spatial activation patterns as unsupervised segmentation of fMRI data. In: Medical Image Computing and Computer-Assisted Intervention MICCAI 2007, vol. 4791. Springer, Berlin, pp. 110–118.

Graham, J.M., Sagar, H.J., 1999. A data-driven approach to the study of heterogeneity in idiopathic Parkinson's disease: identification of three distinct subtypes. Move. Disord. 14 (1), 10–20.

Gulshan, V., Rother, C., Criminisi, A., Blake, A., Zisserman, A., 2010. Geodesic star convexity for interactive image segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3129–3136.

Hansen, L.K., Larsen, J., Nielsen, F.A., Strother, S.C., Rostrup, E., Savoy, R., Lange, N., Sidtis, J., Svarer, C., Paulson, O.B., 1999. Generalizable patterns in neuroimaging: how many principal components? NeuroImage 9 (5), 534–544.

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.

Heimann, T., Meinzer, H.P., 2009. Statistical shape models for 3d medical image segmentation: a review. Med. Image Anal. 13 (4), 543–563.
Heller, R., Stanley, D., Yekutieli, D., Rubin, N., Benjamini, Y., 2006. Cluster-based analysis of fMRI data. NeuroImage 33 (2), 599–608.
Hinrichs, C., Singh, V., Mukherjee, L., Xu, G., Chung, M.K., Johnson, S.C., 2009. Spatially augmented LPboosting for AD classification with evaluations on the ADNI dataset. NeuroImage 48 (1), 138–149.
Hinrichs, C., Singh, V., Xu, G., Johnson, S.C., 2011. Predictive markers for AD in a multi-modality framework: an analysis of MCI progression in the ADNI population. NeuroImage 55 (2), 574–589.
Honnorat, N., Eavani, H., Satterthwaite, T.D., Gur, R.E., Gur, R.C., Davatzikos.C., 2015. GraSP: geodesic graph-based segmentation with shape priors for the functional parcellation of the cortex. NeuroImage 106, 207–211.
Hsieh, J., 2009. Computed tomography: principles, design, artifacts, and recent advances. SPIE, Bellingham, WA.
Huang, C., Wahlund, L.O., Almkvist, O., Elehu, D., Svensson, L., Jonsson, T., Winblad, B., Julin, P., 2003. Voxel- and VOI-based analysis of SPECT CBF in relation to clinical and psychological heterogeneity of mild cognitive impairment. NeuroImage 19 (3), 1137–1144.
Hubert, L., Arabie, P., 1985. Comparing partitions. J. Class. 2 (1), 193–218.
Jacobs, R.A., Jordan, M.I., Nowlan, S.J., Hinton, G.E., 1991. Adaptive mixtures of local experts. Neural Comput. 3 (1), 79–87.
Jeste, S.S., Geschwind, D.H., 2014. Disentangling the heterogeneity of autism spectrum disorder through genetic findings. Nat. Rev. Neurol. 10 (2), 74–81.
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.
Koutsouleris, N., Gaser, C., Jäger, M., Bottlender, R., Frodl, T., Holzinger, S., Schmitt, G.J.E., Zetzsche, T., Burgermeister, B., Scheuerecker, J., Born, C., Reiser, M., Möller, H.J., Meisenzahl, E.M., 2008. Structural correlates of psychopathological symptom dimensions in schizophrenia: a voxel-based morphometric study. NeuroImage 39 (4), 1600–1612.
Koutsouleris, N., Davatzikos, C., Bottlender, R., Patschurek-Kliche, K., Scheuerecker, J., Decker, P., Gaser, C., Moller, H.J., Meisenzahl, E.M., 2012. Early Recognition and Disease Prediction in the At-Risk Mental States for Psychosis Using Neurocognitive Pattern Classification. Schizophrenia Bull. 38 (6), 1200–1215.
Koutsouleris, N., Meisenzahl, E.M., Borgwardt, S., Riecher-Rossler, A., Frodl, T., Kambeitz, J., Kohler, Y., Falkai, P., Moller, H.J., Reiser, M., Davatzikos, C., 2015. Individualized differential diagnosis of schizophrenia and mood disorders using neuroanatomical biomarkers. Brain 138 (7), 2059–2073.
Ladicky, L., Torr, P., 2011. Locally linear support vector machines. In: Proceedings of the 28th International Conference on Machine Learning (ICML-11), pp. 985–992.
Lambert, J.C., Ibrahim-Verbaas, C.A., Harold, D., Naj, A.C., Sims, R., Bellenguez, C., Jun, G., DeStefano, A.L., Bis, J.C., Beecham, G.W., et al., 2013. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. Nat. Genet. 45 (12), 1452–1458.
Lashkari, D., Vul, E., Kanwisher, N., Golland, P., 2010. Discovering structure in the space of fMRI selectivity profiles. NeuroImage 50 (3), 1085–1098.

Lee, D.D., Seung, H.S., 1999. Learning the parts of objects by non-negative matrix factorization. Nature 401 (6755), 788–791.

Lewis, S.J.G., Foltynie, T., Blackwell, A.D., Robbins, T.W., Owen, A.M., Barker, R.A., 2005. Heterogeneity of Parkinson's disease in the early clinical stages using a data driven approach. J. Neurol. Neurosurg. Psychiat. 76 (3), 343–348.

Liang, Z.P., Lauterbur, P.C., 2000. Principles of Magnetic Resonance Imaging. SPIE Optical Engineering Press, Bellingham, WA.

Maintz, J.B.A., Viergever, M.A., 1998. A survey of medical image registration. Med. Image Anal. 2 (1), 1–36.

McEvoy, L.K., Fennema-Notestine, C., Roddey, J.C., Hagler, D.J., Holland, D., Karow, D.S., Pung, C.J., Brewer, J.B., Dale, A.M., 2009. Alzheimer disease: quantitative structural neuroimaging for detection and prediction of clinical and structural changes in mild cognitive impairment. Radiology 251 (1), 195–205.

McGraw, K.O., Wong, S.P., 1996. Forming inferences about some intraclass correlation coefficients. Psychol. Meth. 1, 30–46.

McIntosh, A.R., Mišić, B., 2013. Multivariate statistical analyses for neuroimaging data. Ann. Rev. Psychol. 64, 499–525.

McKeown, M.J., Makeig, S., Brown, G.G., Jung, T.P., Kindermann, S.S., Bell, A.J., Sejnowski, T.J., 1998. Analysis of fMRI data by blind separation into independent spatial components. Hum. Brain Map. 6 (3), 160–188.

Moon, T.K., 1996. The expectation-maximization algorithm. IEEE Signal Process. Mag. 13 (6), 47–60.

Mour ao-Miranda, J., Bokde, A.L.W., Born, C., Hampel, H., Stetter, M., 2005. Classifying brain states and determining the discriminating activation patterns: support Vector Machine on functional MRI data. NeuroImage 28 (4), 980–995.

Mour ao-Miranda, J., Hardoon, D.R., Hahn, T., Marquand, A.F., Williams, S.C.R., Shawe–Taylor, J., Brammer, M., 2011. Patient classification as an outlier detection problem: an application of the One-Class Support Vector Machine. NeuroImage 58 (3), 793–804.

Murray, M.E., Graff-Radford, N.R., Ross, O.A., Petersen, R.C., Duara, R., Dickson, D.W., 38 (6), 1200–1215 2011. Neuropathologically defined subtypes of Alzheimer's disease with distinct clinical characteristics: a retrospective study. Lancet Neurol. 10 (9), 785–796.

Myronenko, A., Song, X., 2010. Point set registration: coherent point drift. IEEE Trans. Pattern Anal. Mach. Intell. 32, 2262–2275.

Noh, Y., Jeon, S., Lee, J.M., Seo, S.W., Kim, G.H., Cho, H., Ye, B.S., Yoon, C.W., Kim, H.J., Chin, J., et al., 2014. Anatomical heterogeneity of Alzheimer disease based on cortical thickness on MRIs. Neurology 83 (21), 1936–1944.

Norman, K.A., Polyn, S.M., Detre, G.J., Haxby, J.V., 2006. Beyond mind-reading: multi-voxel pattern analysis of fMRI data. Trends Cogn. Sci. 10 (9), 424–430.

Pham, D.L., Xu, C., Prince, J.L., 2000. Current Methods in Medical Image Segmentation. Ann. Rev. Biomed. Eng. 2 (1), 315–337.

Phelps, M.E., 2000. Positron emission tomography provides molecular imaging of biological processes. Proc. Natl. Acad. Sci. USA 97 (16), 9226–9233.

Ryali, S., Chen, T., Supekar, K., Menon, V., 2013. A parcellation scheme based on von Mises-Fisher distributions and Markov random fields for segmenting brain regions using resting-state fMRI. NeuroImage 65 (0), 83–96.

Sajda, P., 2006. Machine Learning for Detection and Diagnosis of Disease. Ann. Rev. Biomed. Eng. 8 (1), 537–565.

Satterthwaite, T., Elliott, M.A., Ruparel, K., Loughead, J., Prabhakaran, K., Calkins, M.E., Hopson, R., Jackson, C., Keefe, J., Riley, M., Mentch, F.D., Sleiman, P., Verma, R., Davatzikos, C., Hakonarson, H., Gur, R.C., Gur, R.E., 2014. Neuroimaging of the Philadelphia neurodevelopmental cohort. NeuroImage 86, 544–553.

Shen, D., Davatzikos, C., 2002. HAMMER: hierarchical attribute matching mechanism for elastic registration. IEEE Trans. Med. Imaging 21 (11), 1421–1439.

Shen, X., Papademetris, X., Constable, R.T., 2010. Graph-theory based parcellation of functional subunits in the brain from resting-state fMRI data. NeuroImage 50, 1027–1035.

Smith, S.M., Miller, K.L., Salimi-Khorshidi, G., Webster, M., Beckmann, C.F., Nichols, T.E., Ramsey, J.D., Woolrich, M.W., 2011. Network modelling methods for fMRI. NeuroImage 54, 875–891.

Smith, S.M., Miller, K.L., Moeller, S., Xu, J., Auerbach, E.J., Woolrich, M.W., Beckmann, C.F., Jenkinson, M., Andersson, J., Glasser, M.F., Van Essen, D.C., Feinberg, D.A., Yacoub, E.S., Ugurbil, K., 2012. Temporally-independent functional modes of spontaneous brain activity. Proc. Natl. Acad. Sci. USA 109 (8), 3131–3136.

Sotiras, A., Davatzikos, C., Paragios, N., 2013. Deformable medical image registration: a survey. IEEE Trans. Med. Imaging 32 (7), 1153–1190.

Sotiras, A., Resnick, S.M., Davatzikos, C., 2015. Finding imaging patterns of structural covariance via Non-Negative Matrix Factorization. NeuroImage 108, 1–16.

Strother, S.C., Anderson, J.R., Schaper, K.A., Sidtis, J.J., Liow, J.S., Woods, R.P., Rottenberg, D.A., 1995. Principal component analysis and the scaled subprofile model compared to intersubject averaging and statistical parametric mapping: I. "Functional connectivity" of the human motor system studied with [15O]water PET. J. Cerebral Blood Flow Metabol. 15 (5), 738–753.

Sun, D., van Erp, T.G., Thompson, P.M., Bearden, C.E., Daley, M., Kushan, L., Hardt, M.E., Nuechterlein, K.H., Toga, A.W., Cannon, T.D., 2009. Elucidating a magnetic resonance imaging-based neuroanatomic biomarker for psychosis: classification analysis using probabilistic brain atlas and machine learning algorithms. Biol. Psychiat. 66 (11), 1055–1060.

Tononi, G., Sporns, O., Edelman, G.M., 1994. A measure for brain complexity: relating functional segregation and integration in the nervous system. Proc. Natl. Acad. Sci. USA 91, 5033–5037.

Van Essen, D.C., Smith, S.M., Barch, D.M., Behrens, T.E.J., Yacoub, E., Ugurbil for the WU-Minn HCP Consortium., K., 2013. The WU-Minn human connectome project: an overview. NeuroImage 80, 62–79.

Vapnik, V.N., 2000. The Nature of Statistical Learning Theory. Springer New York. 315.

Varol, E., Gaonkar, B., Davatzikos, C., 2013. Classifying medical images using morphological appearance manifolds. In: IEEE 10th International Symposium on Biomedical Imaging (ISBI), pp. 744–747.

Varol, E., Sotiras, A., Davatzikos, C., 2015. Disentangling disease heterogeneity with max–margin multiple hyperplane classifier. In: Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015. Springer, pp. 702–709.

Varol, E., Sotiras, A., Davatzikos, C., 2016. HYDRA: revealing heterogeneity of imaging and genetic patterns through a multiple max-margin discriminative analysis framework. NeuroImage. ISSN: 1053–8119, doi: http://dx.doi.org/10.1016/j.neuroimage.2016.02.041, http://www.sciencedirect.com/science/article/pii/S1053811916001506.

Veksler, O., 2008. Star shape prior for graph-cut image segmentation. In: IEEE European Conference on Computer Vision (ECCV), pp. 454–467.

Vemuri, P., Gunter, J.L., Senjem, M.L., Whitwell, J.L., Kantarci, K., Knopman, D.S., Boeve, B.F., Petersen, R.C., Jack, C.R., 2008. Alzheimer's disease diagnosis in individual subjects using structural MR images: validation studies. NeuroImage 39 (3), 1186–1197.

Vincent, L., Soille, P., 1991. Watersheds in digital spaces: an efficient algorithm based on immersion simulations. IEEE Trans. Pattern Anal. Mach. Intell. 13 (6), 583–589.

Whitwell, J.L., Petersen, R.C., Negash, S., Weigand, S.D., Kantarci, K., Ivnik, R.J., Knopman, D.S., Boeve, B.F., Smith, G.E., Jack, C.R., 2007. Patterns of atrophy differ among specific subtypes of mild cognitive impairment. Arch. Neurol. 64 (8), 1130–1138.

Zeng, L.L., Shen, H., Liu, L., Wang, L., Li, B., Fang, P., Zhou, Z., Li, Y., Hu, D., 2012. Identifying major depression using whole-brain functional connectivity: a multivariate pattern analysis. Brain 135 (5), 1498–1507.

Zhang, D., Wang, Y., Zhou, L., Yuan, H., Shen, D., 2011. Multimodal classification of Alzheimer's disease and mild cognitive impairment. NeuroImage 55 (3), 856–867.

Zhang, T., Koutsouleris, N., Meisenzahl, E., Davatzikos, C., 2015. Heterogeneity of Structural Brain Changes in Subtypes of Schizophrenia Revealed Using Magnetic Resonance Imaging Pattern Analysis. Schizophrenia Bull. 41 (1), 74–84.

Zhu, C.Z., Zang, Y.F., Cao, Q.J., Yan, C.G., He, Y., Jiang, T.Z., Sui, M.Q., Wang, Y.F., 2008. Fisher discriminative analysis of resting-state brain function for attention-deficit/hyperactivity disorder. NeuroImage 40 (1), 110–120.