Machine Learning as a Means towards Precision Diagnostics and Prognostics

Aristeidis Sotiras, Bilwaj Gaonkar, Harini Eavani, Nicolas Honnorat, Erdem Varol, Aoyan Dong and Christos Davatzikos

Section of Biomedical Image Analysis, Center for Biomedical Image Computing and Analytics, University of Pennsylvania, Philadelphia PA, USA

Abstract

Machine learning plays an essential role in the medical imaging field. High-dimensional pattern analysis techniques have been developed that can identify, and quantify, subtle and spatially complex patterns of structural and functional changes in the brain that are induced by brain diseases despite the presence of confounding statistical noise and inter-individual anatomical and functional variability. Sophisticated pattern analysis of structural and functional images can detect early signs of diseases, which otherwise would remain undetectable using conventional methods. As a consequence, pattern analysis techniques have been used to construct sensitive biomarkers that can identify disease, or risk for developing it, and characterize future clinical progression on an individual patient basis. This has lead to led to them becoming an indispensable part for the growing need for personalized, predictive medicine. However, despite important advances and successes, there remains a number of challenges to be addressed before gaining widespread acceptance as tools for precision diagnostics and prognostics in clinical practice. Some of the most important challenges include: i) feature extraction and dimensionality reduction; ii) readily interpreting complex multivariate models; and iii) elucidating disease heterogeneity. In this chapter, we describe these challenges, putting emphasis on possible solutions and present evidence of the usefulness of machine learning techniques at the clinical and research level.

Keywords: Multivariate pattern analysis, high-dimensional pattern classification, regression, Support Vector Machines, statistical inference, heterogeneity, clustering, Markov Random Fields, Alzheimer's Disease, genetics, structural MRI, fMRI, Diagnosis of AD

1. Introduction

The advent of new imaging modalities providing high resolution depictions of the anatomy and function of the brain in disease and health has resulted in medical imaging becoming increasing 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, image registration, image fusion and computer aided diagnosis.

Among the reasons behind the success of machine learning in medical imaging is 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 region of interest based analyses. Moreover, contrary to conventional automated approaches such as mass univariate analyses, high dimensional

multivariate pattern analysis (MPVA) driven ones 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 along 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. 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 [1–11], to differentially distinguish between Alzheimer's Disease and frontotemporal dementia (FTD) [12], or to predict clinical progression of patients [13, 14]. Mental disorders have also provided fertile ground for 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 [15, 16], psychosis [17] or depression [18].

However, despite important advances and successes, there remains 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 regards one of the fundamental problems 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 multi-parametric imaging, which further increase the dimensionality and complexity of the available data. The main challenge is to summarize 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 regards 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 mutlivariate 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 regards 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 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 Sec. 2 we present clustering and statistical-based approached for dimensionality reduction of both structural and functional data. In Sec. 3 we detail an efficient technique for deriving statistical significance maps in classification tasks using Support Vector Machines, while in Sec. 4 we present a palette of techniques to tackle disease heterogeneity under different methodological assumptions. In Sec. 5 we provide evidence of the usefulness of machine learning techniques at the clinical and research level, while Sec. 6 concludes the chapter.

2. Dimensionality Reduction

During the past decades, the advent of high-resolution imaging techniques has given rise to high dimensional, complex clinical data sets consisting of hundred of patient scans that comprise millions of voxels [19, 20]. The high dimensionality of the data along with the relative small sample size that is typically available pose 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 multi-parametric imaging data, which results in additionally increasing both the dimensionality and the 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 emphasize the aforementioned challenge but also call for tailored analysis tools.

Towards addressing the previous 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, the extracted 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 for reducing 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.

2.1. Dimensionality reduction through spatial grouping

Methods of this class typically formulate the problem as a clustering one 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 a computing a single average measure per estimated cluster, while clusters are obtained by segmenting the brain into contiguous regions that encompass elements whose imaging measurements 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. Following, we summarize two such algorithms for structural Magnetic Resonance Imaging (MRI) scans and resting-state functional MRI (rs-fMRI), respectively.

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 type of 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 (*i.e.*, 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 [21] 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 the subjects' diagnosis. The two types of information are integrated at each image location p in a multiplicative fashion though the score $s(p) = P(p)C(p)$, 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:

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

where $\rho(p,i)$ denotes the Pearson correlation measured between the imaging signal at p and the classification labels when excluding the i-th subject/sample. The consistency $C(p)$ is the intra-class coefficient measuring the proportion of neighboring features variance that is explained by the inter-subject variability [21, 22]. 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 (*i.e.*, highly correlated or anti-correlated with patient diagnosis).

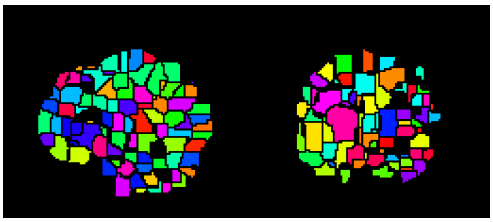


Figure 1: Coronal and sagittal cross-sectional views of a watershed segmentation generated by COMPARE.

This score map is subsequently smoothed, and its gradient is used in conjunction with a watershed segmentation algorithm [23] to partition the brain into different regions (Fig. 1 presents brain regions generated by watershed form white matter tissue density maps of demented and normally aging subjects [21, 24]). 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. This approach extracts a single connected component per watershed region. Each component comprises highly discriminative elements whose average imaging signal may be used as feature for training a classifier such as a Support Vector Machine [25].

The efficiency of this supervised dimensionality reduction scheme was demonstrated in classifying demented and normal patients as well as distinguishing between schizophrenic patients

and normal controls [21]. The previous supervised technique is generic and can be straightforwardly extended to incorporate different forms of prior information such as the ones provided in regression and multi-class classification settings.

2.2.1. Spatial grouping of rs-fMRI

Functional MRI is an imaging technique that tracks neural activity in the whole brain by detecting changes in oxygen consumption. Resting-state fMRI describes a large part of the brain networks [26] 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 due to that 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 to 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 [27].

Clustering approaches typically aim to divide the brain into spatially smooth regions 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 regions 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 the partial correlation [28]. 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 [29], normalized clustering [30, 31], k-means [32], region growing [33, 34] and Markov Random Fields (MRFs) [35–37]. Different methods exhibit distinct advantages and disadvantages. Generally, many of the above methods are either initialization dependent (*e.g.*, region growing [33, 34] and k-means [32]), or rely on complex models that involve a large number of parameters [36]. As a result, they are sensitive to initialization and suffer from limitations related to the employed heuristics (*e.g.*, hierarchical clustering may lead to the apparition of bad parcels at coarse scale [29, 37]) and the large number of inferred parameters that may negatively impact the quality of the locally optimal solution that is obtained [36]. Moreover, not all methods produce contiguous parcels.

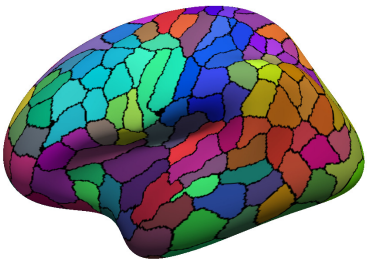


Figure 2: Functional parcellation of the left hemisphere of the brain, projected on an inflated brain surface.

In order to address the aforementioned concerns, a discrete MRF approach, termed GraSP, was recently introduced in [37]. This approach adopts an *exemplar-based clustering* 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 thought the encoding of the parcels with their functional center. Only

one parameter need to be chosen by the user; the label cost K that corresponds to the cost of introducing a new parcel into the clustering result, and thus indirectly determines the size of the produced parcels [38]. Contrary to other MRF clustering methods [35], these parcels are connected (Fig. 2 presents a functional parcellation that was produced for reducing the dimension of rs-fMRI scans from a neurodevelopmental study [20]). Parcel connectedness is promoted without any spatial smoothing by the inclusion of a shape prior term into the MRF energy formulation [39, 40]. Lastly, the energy is optimized in a single step, thus removing the need for initialization and specifying 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 a 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 [38] 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 result produced by hierarchical clustering and normalized cuts [37].

2.3. Statistically driven dimensionality reduction

The second family of dimensionality reduction methods is based on exploiting statistical procedures to project the data is 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) [41–43] 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 the basis vector (or principal components) are ordered in descending order according to the amount of the variance they explain. ICA [44–46], 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 cancellation 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 to 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 data sets.

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 threefold: i) sparse methods achieve a higher degree of specificity than conventional multivariate analysis methods [47]; ii) they show improved generalizability [48], while iii) sparsity is an important property for effective tools in brain modeling and analysis [49]. Sparsity is introduced in a tailored way, taking into account the specific properties of different imaging modalities.

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) [47, 50]. Non-Negative Matrix Factorization was proposed as an analytical and interpretive tool in structural neuroimaging in [50].

Non-Negative Matrix Factorization 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:

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

where $\mathbf{B} = [\mathbf{b}_1, \ldots, \mathbf{b}_K]$, $\mathbf{b}_i \in \mathbb{R}^D$ and $\mathbf{C} = [\mathbf{c}_1, \ldots, \mathbf{c}_N], \mathbf{c}_i \in \mathbb{R}^K$. The non-negativity constraints lead to a sparse, parts-based representation [47]. 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 summarize 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 [50]. 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 component ones.

Typical components estimated by NNMF are shown in Fig. 3. Note that the representation amounts to a soft clustering that segments the brain to 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

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 solely statistical construct, but highly correspond to known structural and functional networks of the brain, or in some cases reflect underlying pathological processes.

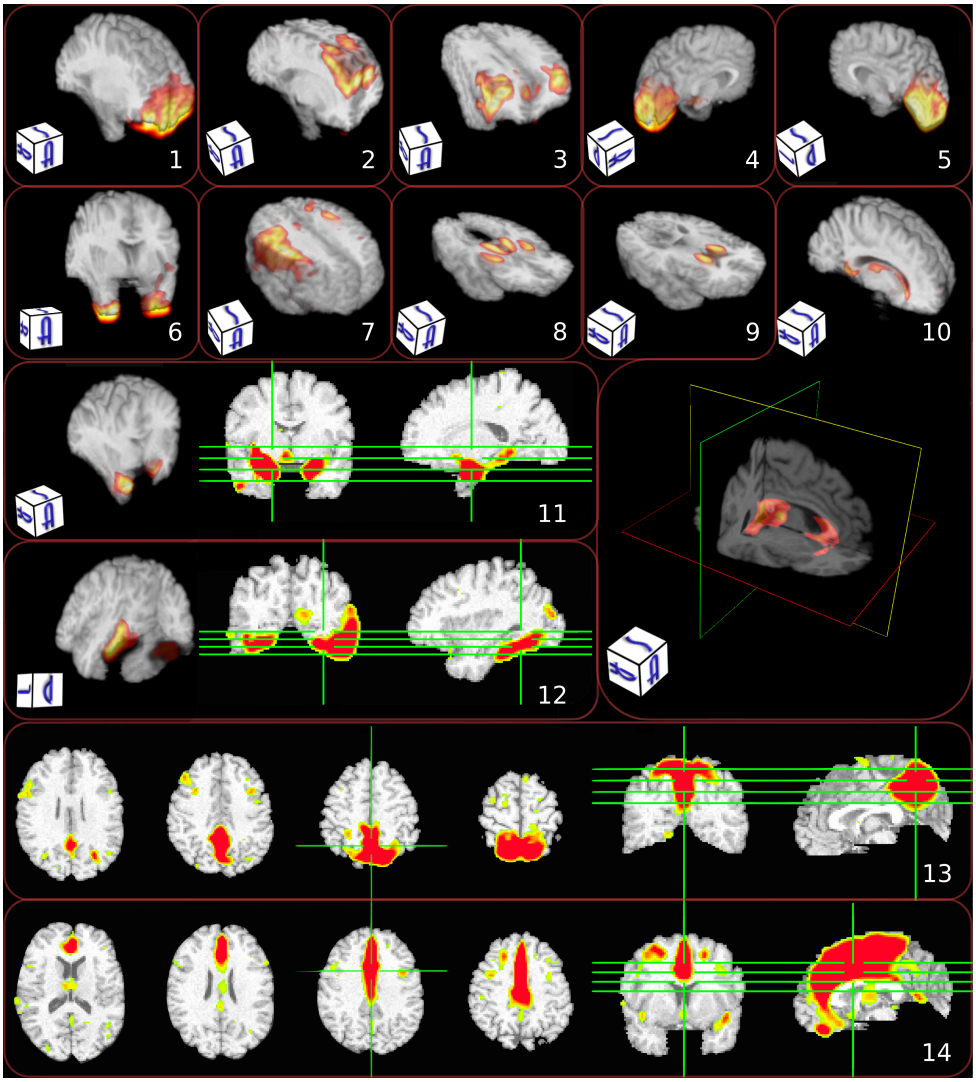


Figure 3: Characteristic components estimated by NNMF. Different visualization strategies were used in order to enhance the visual perception of the components (note that the 2D images use radiographic convention). Warmer colors correspond to higher values. Note the alignment with anatomical regions: 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) peri-ventricular structures; 11) amygdala and hippocampus; 12) fusiform; 13) medial parietal including precuneus; 14) anterior and middle cingulate. The figure is reprinted from [50].

2.3.2. Statistically driven dimensionality reduction of functional MRI

Resting-state functional MRI is typically used to analyze regional interactions, aiming to reveal brain's functional organization. Resting state functional connectivity aims to reveal functional networks that can be found consistently in healthy populations by examining the connectivity between all pairs of regions in the brain. The Pearson correlation is typically use to measure the connectivity between different brain regions due to its simplicity and robustness [28, 51]. The resulting functional connectivity data tends to be 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 model 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 [52]. In this approach, each subject specific correlation matrix $\mathbf{\Sigma}_n$ is approximated by a non-negative sum of sparse rank one matrices $\mathbf{b}_k\mathbf{b}_k^T$. 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 $\mathbf{\Sigma}_n$:

$$
\begin{aligned}
&\underset{\mathbf{B},\mathbf{C}}{\text{minimize}} \sum_{n=1}^{N} \left|\left|\mathbf{\Sigma}_n - \mathbf{B}\,\text{diag}(\mathbf{c}_n)\,\mathbf{B}^T\right|\right|_F^2 \\
&\text{subject to} \\
&\qquad\qquad ||\mathbf{b}_k||_1 \leq \lambda, \quad k = 1,\ldots,K, \\
&\qquad -1 \leq \mathbf{b}_k(i) \leq 1, \quad \max_i |\mathbf{b}_k(i)| = 1, \quad i = 1,\ldots,P \\
&\qquad\qquad \mathbf{c}_n \geq 0, \qquad\quad n = 1,\ldots,N
\end{aligned}
$$

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 [53, 54]. 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. Figure 4 shows a large primary SCP with contributions from the operculum and anti-correlated 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.

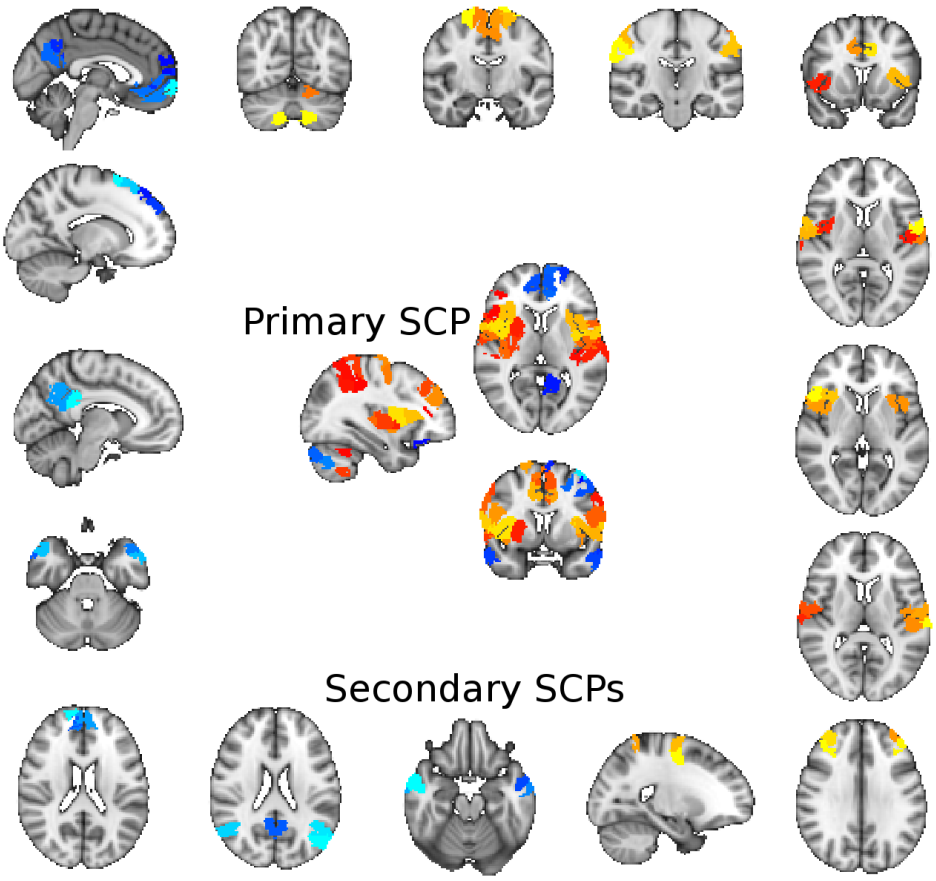


Figure 4: Primary SCP (middle) showing the cingulum, operculum (red-yellow) and anti-correlated with the default mode (blue-light blue). Sixteen of its associated secondary SCPs are shown around it.

3. Model Interpretation: From Classification to Statistical Significance maps

Given an appropriate set of features, 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.

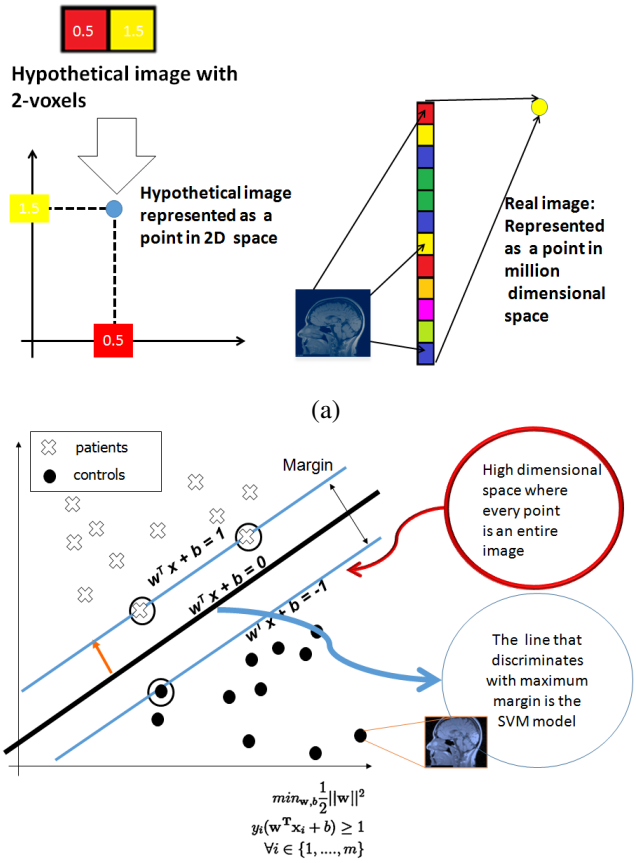


Figure 5: The concept of imaging-based diagnosis using SVMs: (a) Images are treated as points located in a high-dimensional space; (b) The maximum margin principle of classification used in SVMs. Dots and crosses represent imaging scans taken from two groups. Even though the two groups can not be separated 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.

In this section, we present such a framework for Support Vector Machines (SVMs) [25, 55]. Support vector machines enjoy significant popularity in neuroimaging [2, 11, 21, 56–58], mainly due to their simplicity and the fact that the resulting problem is convex, allowing for efficient and globally optimal solutions. The support vector machine operates by constructing a hyperplane in a high dimensional space that separates samples from two classes (*e.g.*, disease group vs. healthy controls) by the largest possible margin (see Fig. 5 for an illustration of the principle). The hyperplane coefficients denoted by $\mathbf{w}^*$ and b^* are estimated by solving the following optimization problem:

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

where $\mathbf{x}_i \in \mathbb{R}^D$ denotes the vectorized image of the i-th subject of the study, $y_i \in \{+1,-1\}$ denotes its respective binary label and ξ_i denotes the slack variable that accounts for the case that the classes are not separable. The weight vector $\mathbf{w}^* \in \mathbb{R}^D$ describes the combination of all imaging elements that 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 [59] 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. 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 estimating of statistical significance. The number of permutations determines the minimal 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 terms requires a considerable 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.

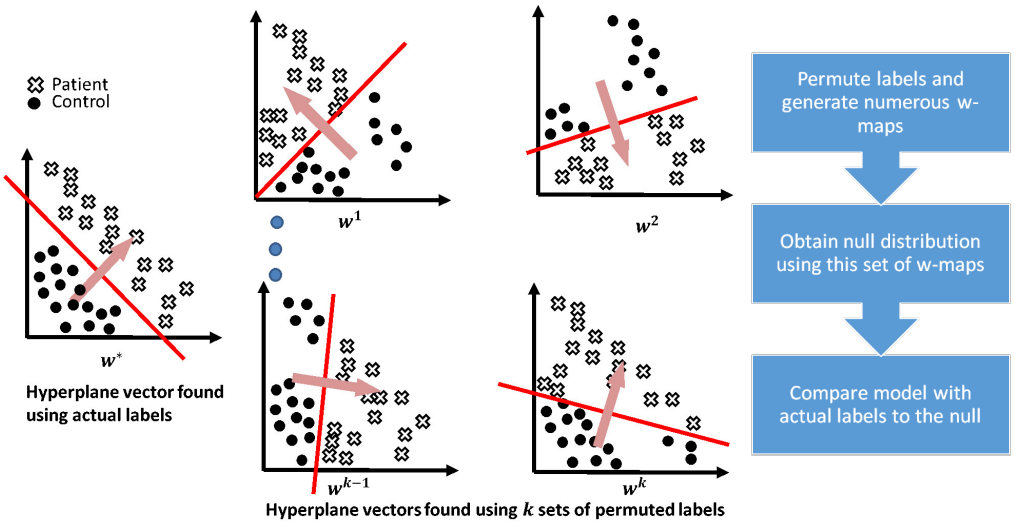


Figure 6: Illustration of the permutation testing procedure. This figure is reprinted from [56].

Such a theoretical framework that describes an analytic alternative to permutation testing was introduced in [56, 60]. 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:

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

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

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

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 Lyapunov central limit theorem, it was demonstrated in [56, 60] that the distribution of the individual components of $\mathbf{w}$ can be approximated using the normal distribution for 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. 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. Figure 8 shows an interpretative statistical atlas obtained using [60] from a model that was used to classify Alzheimer's disease patients from controls. Note that the hippocampal complex along with parahippocampal regions and amygdala are clearly highlighted.

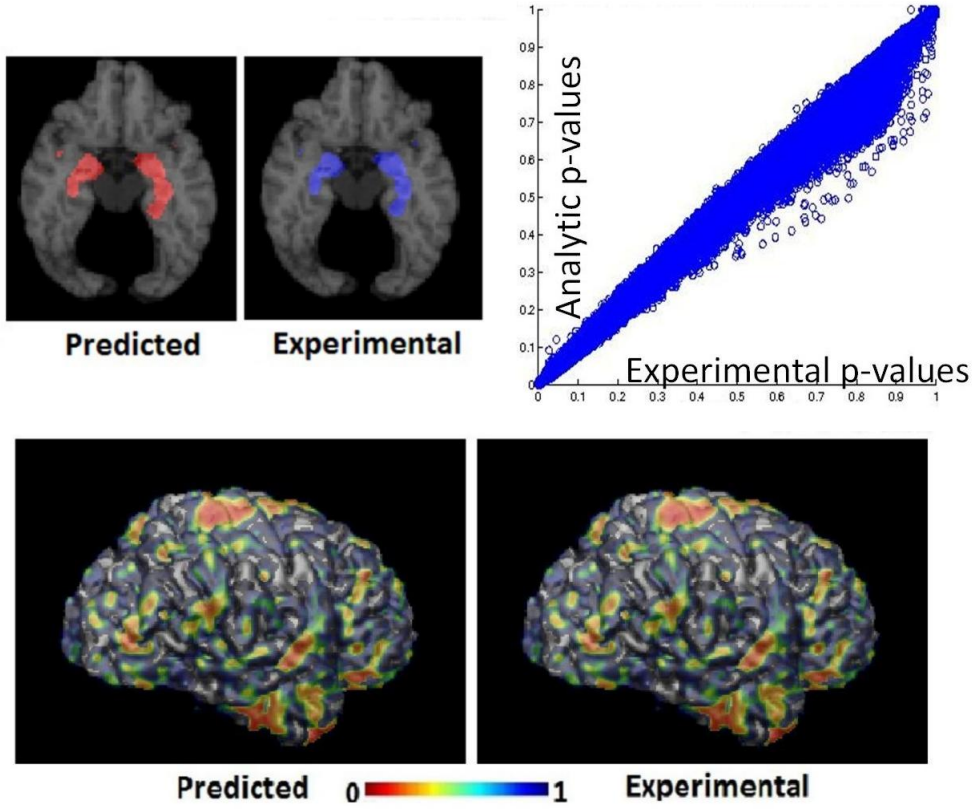


Figure 7: (Top left) Analytic and experimental p-value maps thresholded at 0.01 are overlaid on the template brain. (Top right) A scatter plot of p-values comparing experimental and analytical p-values. (Bottom) A 3D rendering representing the predicted and experimental p-value maps. This figure is reprinted from [56].

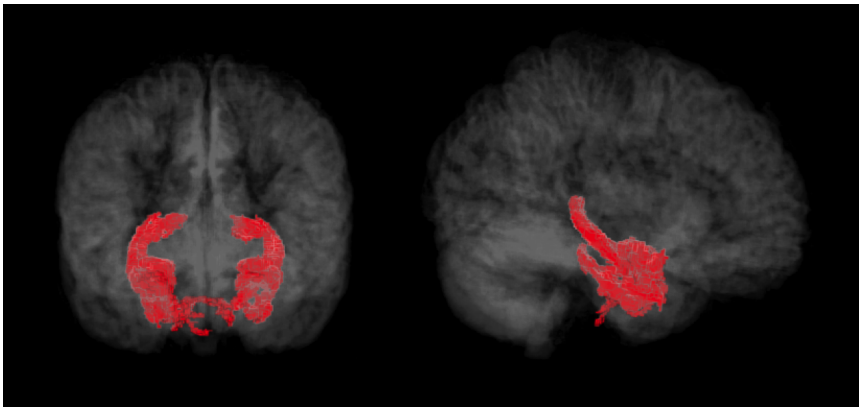
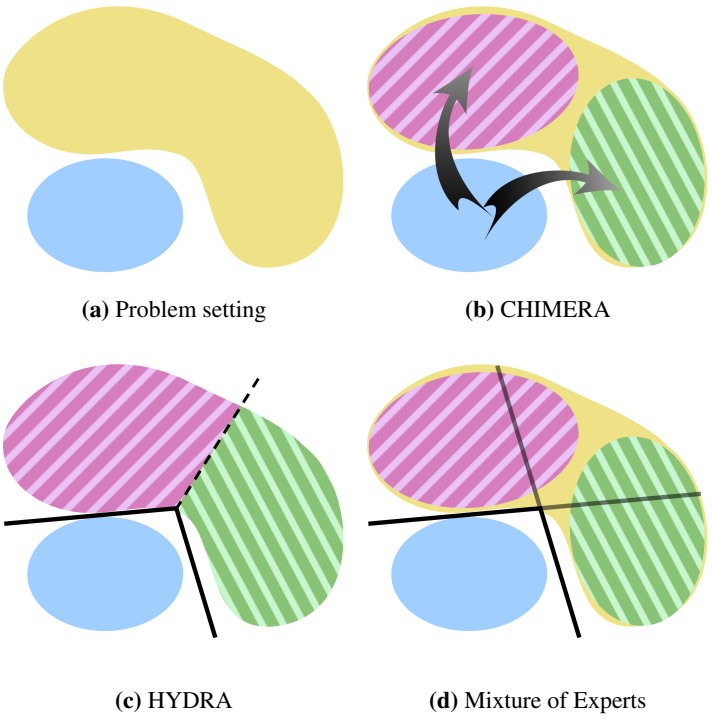


Figure 8: 3D Views of the hippocampal and parahippocampal regions used by the SVM ($\alpha \leq 0.01$ FDR corrected).

4. Heterogeneity



(a) Problem setting **(b)** CHIMERA

(c) HYDRA **(d)** Mixture of Experts

Figure 9: Heterogeneity problem setting and different methods.

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 [61, 62], Schizophrenia [63–65], Parkinson [66, 67], Alzheimer's Disease [68, 69] or Mild Cognitive Impairment (MCI) [70, 71] are all characterized by clinical heterogeneity (see Fig. 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 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.

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 could 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 [72] considers the disease effect to be a transformation from the normal control distribution to the patient distribution (see Fig. 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 $\mathscr{E}$:

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

where Θ denotes the parameters of the model, such as that transformations that are applied to $\mathbf{X}$ in order to generate $\mathbf{Y}$; $\mathscr{L}$ is the log-likelihood of the distributions $\mathbf{X}$ and $\mathbf{Y}$ given the parameters; and $\mathscr{R}$ 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 [73]. Each transformed normal control point is considered as a centroid of a spherical Gaussian cluster, and patient points are treated as i.i.d. 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 multi-kernel setting. These model assumptions lead to the log-likelihood term $\mathscr{L}$ to be:

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

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.

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

Energy objective $\mathscr{E}$ is optimized with an Expectation-Maximization[74] approach. The heterogeneous disease subgroups of patients are further clustered by the estimated transformations.

This method was applied to an Alzheimer's Disease 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 was used as covariate features. With the cross-validated parameters, two subgroups were discovered. Voxel-Based Morphometry (VBM) [75] 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. Subgroup 1 has more gray matter atrophy in limbic lobe and frontal insular regions, and 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.

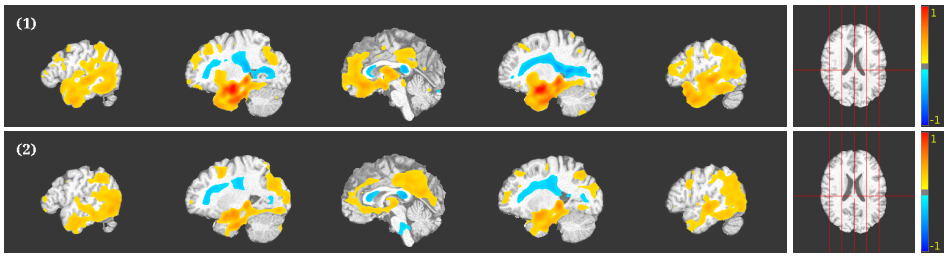


Figure 10: VBM performed on gray matter RAVENS [76] maps between (1) Subgroup 1 and Control group; (2) Subgroup 2 and Control group. The results were thresholded by FDR adjusted p-value at level of 0.01 are presented, overlaid on the registration template image. Color-maps indicate the scale of the t-statistic. Warmer colors indicate volume loss, while colder colors indicate volume increase.

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 non-linear classifiers, allowing for larger margins

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

and thus, better generalization. However, while kernel methods, such as Gaussian kernel SVM, provide non-linearity, they lack interpretability when aiming to characterize heterogeneity.

In order to tackle the aforementioned limitations, a novel maximum margin non-linear learning algorithm for simultaneous binary classification and subtype identification, termed HYDRA (Heterogeneity through Discriminative Analysis) was introduced in [77, 78]. 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 (*e.g.*, healthy controls from pathologic samples); each face of this polytope effectively defines a disease subtype (see Fig. 9c for a graphical illustration).

More formally, let us assume an imaging (or genetic) dataset consisting of n binary labeled d-dimensional data points ($\mathscr{D} = (\mathbf{x}_i, y_i)_{i=1}^n, \mathbf{x}_i \in \mathbb{R}^d$ 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:

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

The first term encourages maximum average margin across all K faces 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 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 Alzheimer's disease (AD) patients and 68 cognitively normal (CN) older adults (see demographic information in Table 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 [79] 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 [80]. This analysis suggested that 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. 11A and 11B). Correction for multiple comparisons was performed using FDR. The results were thresholded at $q < 0.05$. It can be observed that at the $K = 2$ cluster level (see Fig. 11), the

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

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

Table 1: Demographic and clinical characteristics of healthy controls, AD patients (left) and the estimated genetic-driven subtypes of AD (right). [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.

estimated subgroups were associated with distinct patterns of structural brain alterations. The first subgroup had increased temporal lobe atrophy (see Fig. 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. 11B), including periventricular white matter lesions.

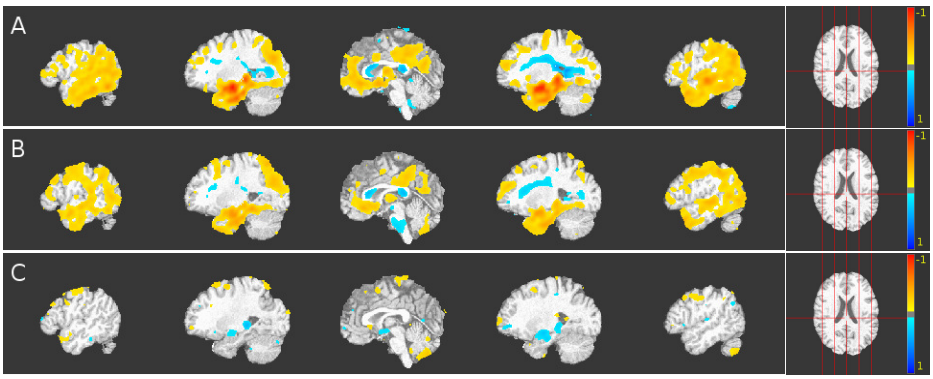


Figure 11: The anatomic differences between the two genetic subtypes of AD: Axial views of gray matter group comparisons of (A) Controls vs. first AD subgroup; (B) Controls vs. second AD subgroup; and (C) first AD subgroup vs. second AD subgroup are visualized. For (A) and (B), colder colors indicate 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.

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 to 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.

4.3. Generative Discriminative Framework

The last approach that aims to identify heterogeneous sub-groups in patient populations is based upon a Mixture-of-Experts (MOE) framework. The MOE framework was initially proposed for vowel discrimination within speech recognition [81] and later on, as a fast and efficient alternative to "kernel" SVMs [82, 83]. While kernel SVMs can successfully model non-linear 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 non-linear 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 [84]. Using this combination of unsupervised clustering (mixture) with supervised classification/regression (expert), it approximates the non-linear boundary that separates the two classes with a piece-wise linear separating boundary, providing us the identification of the sub-groups as well as the multivariate patterns that discriminate each sub-group from the reference group (see Fig. 9d for a graphical illustration). The data is modeled using a mixture of distributions, such as Fuzzy C-Means, which assigns a soft sub-group membership to each subject in the affected group. The linear boundary between each affected sub-group 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 l2-loss linear SVMs, the authors in [84] found heterogeneity in the manner in which normal older individuals age in terms of functional connectivity. Of the two sub-groups 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.

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.

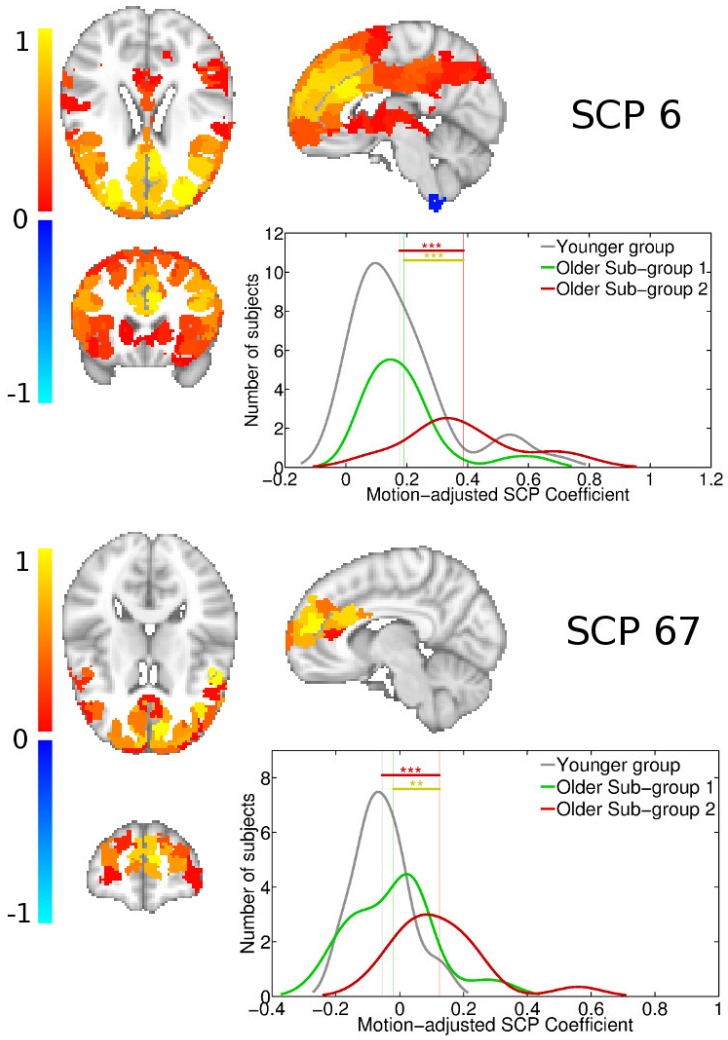


Figure 12: Plot showing primary SCP 6, and its associated secondary SCP 67, whose average connectivity is increased in the second older subgroup, but not the first. SCP 6 highlights most of the pre-frontal cortex. SCP 67 captures the bilateral para-cingulate gyrus and inferior temporal gyrus. The distribution fit of the underlying SCP coefficient histograms are also shown, for each SCP and for each subgroup. Significance levels are indicated as follows: '***' for p-value < 0.001, '**' for p-value < 0.01 and '*' for p-value < 0.05. The figure is reprinted from [84].

5. Applications

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 on an individual person basis has gained a great deal of attention, 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[1, 2] and schizophrenia [15]. We now summarize our groups work on deriving the SPARE-AD index, an index that measures the presence of AD-like patterns of brain atrophy from brain MRI.

5.2. MRI-based diagnosis of AD: the SPARE-AD

In [13], 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 15 indicates the regions with the most significant brain atrophy and ventricular expansion.

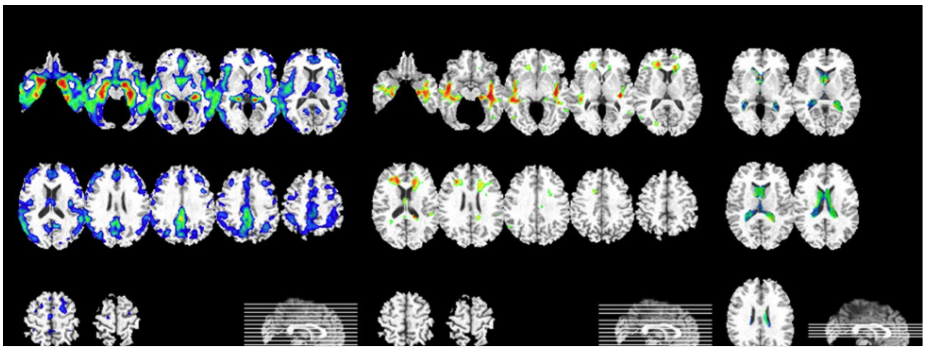


Figure 13: From left to right, group comparison results on GM, WM, and CSF are shown. The color-maps indicate the scale for the t-statistic. Images are displayed in radiological convention. Images reprinted with permission from [13].

The histograms of the (cross-validated) SPARE-AD scores achieved in this classification are shown in Fig. 15, 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.

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 pre-clinically. These are the stages where standard clinical evaluations might be less effective and hence likely to benefit from imaging-based biomarkers. Along this vein, the SPARE-AD index was examined in individuals with mild cognitive impairment in [14, 85], and it was found to have predict to a large extent an individuals future progression to dementia. Fig. 15 shows survival curves obtained from baseline measures in 432 MCI patients of the ADNI1 study.

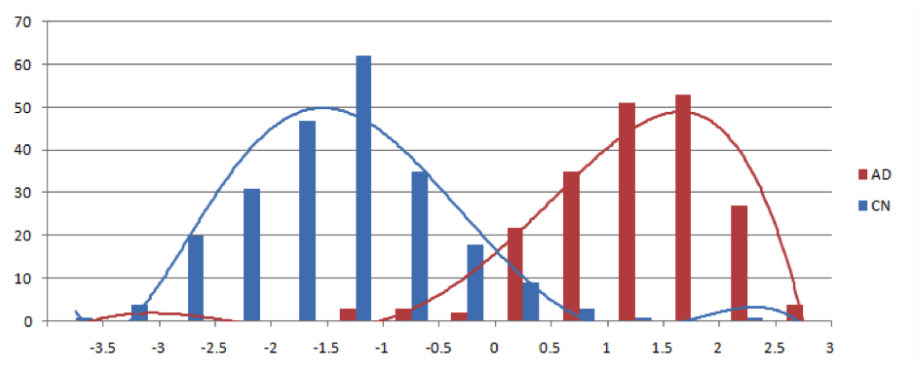


Figure 14: Histograms of SPARE-AD scores obtained via cross validation from the ADNI1 sample from CN and AD individuals.

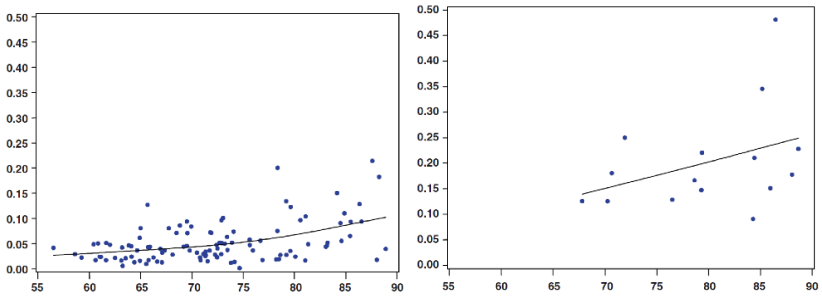


Figure 16: Annual rates of SPARE-AD change at the BLSA longitudinal study. People who remained stable are shown on the left, and people who converted to MCI are shown on the right displaying markedly higher rates of SPARE-AD change prior to cognitive decline.

Looking at even earlier stages of the progression of patterns of brain atrophy evaluated via machine learning, the study in [86] 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. 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 (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.

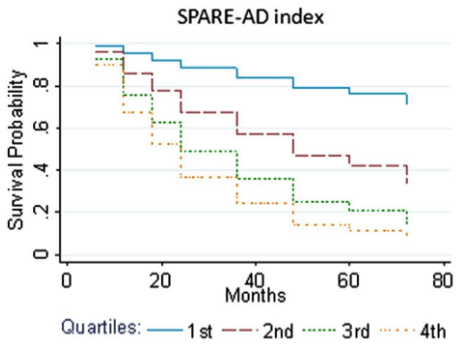


Figure 15: Survival curves showing predictive value of MRI-derived patterns of atrophy that were evaluated using machine learning (the SPARE-AD index).

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 multi-faceted, 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/sites. This can be a 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

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

[2] S. Klöppel, C. M. Stonnington, C. Chu, B. Draganski, R. I. Scahill, J. D. Rohrer, N. C. Fox, C. R. Jack, J. Ashburner, R. S. J. Frackowiak, Automatic classification of MR scans in Alzheimer's disease, Brain 131 (3) (2008) 681–689.

[3] P. Vemuri, J. L. Gunter, M. L. Senjem, J. L. Whitwell, K. Kantarci, D. S. Knopman, B. F. Boeve, R. C. Petersen, C. R. Jack, Alzheimer's disease diagnosis in individual subjects using structural MR images: validation studies., NeuroImage 39 (3) (2008) 1186–97.

[4] S. Duchesne, a. Caroli, C. Geroldi, C. Barillot, G. B. Frisoni, D. L. Collins, MRI-based automated computer classification of probable AD versus normal controls., IEEE transactions on medical imaging 27 (4) (2008) 509–20.

[5] Y. Fan, S. M. Resnick, X. Wu, C. Davatzikos, Structural and functional biomarkers of prodromal Alzheimer's disease: a high-dimensional pattern classification study., NeuroImage 41 (2) (2008) 277–85.

[6] L. K. McEvoy, C. Fennema-Notestine, J. C. Roddey, D. J. Hagler, D. Holland, D. S. Karow, C. J. Pung, J. B. Brewer, A. M. Dale, Alzheimer Disease: Quantitative Structural Neuroimaging for Detection and Prediction of Clinical and Structural Changes in Mild Cognitive Impairment, Radiology 251 (1) (2009) 195–205.

[7] C. Hinrichs, V. Singh, L. Mukherjee, G. Xu, M. K. Chung, S. C. Johnson, Spatially augmented LPboosting for AD classification with evaluations on the ADNI dataset, NeuroImage 48 (1) (2009) 138–149.

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

[9] C. Hinrichs, V. Singh, G. Xu, S. C. Johnson, Predictive markers for AD in a multi-modality framework: an analysis of MCI progression in the ADNI population., NeuroImage 55 (2) (2011) 574–89.

[10] D. Zhang, Y. Wang, L. Zhou, H. Yuan, D. Shen, Multimodal classification of Alzheimer's disease and mild cognitive impairment., NeuroImage 55 (3) (2011) 856–67.

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

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

[13] Y. Fan, N. Batmanghelich, C. M. Clark, C. Davatzikos, Spatial patterns of brain atrophy in MCI patients, identified via high-dimensional pattern classification, predict subsequent cognitive decline., NeuroImage 39 (4) (2008) 1731–43.

[14] C. Davatzikos, P. Bhatt, L. M. Shaw, K. N. Batmanghelich, J. Q. Trojanowski, Prediction of MCI to AD conversion, via MRI, CSF biomarkers, and pattern classification., Neurobiology of aging 32 (12) (2011) 2322.e19–27.

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

[16] N. Koutsouleris, E. M. Meisenzahl, S. Borgwardt, A. Riecher-Rossler, T. Frodl, J. Kambeitz, Y. Kohler, P. Falkai, H.-J. Moller, M. Reiser, C. Davatzikos, Individualized differential diagnosis of schizophrenia and mood disorders using neuroanatomical biomarkers, Brain.
[17] N. Koutsouleris, C. Davatzikos, R. Bottlender, K. Patschurek-Kliche, J. Scheuerecker, P. Decker, C. Gaser, H.-J. Moller, E. M. Meisenzahl, Early Recognition and Disease Prediction in the At-Risk Mental States for Psychosis Using Neurocognitive Pattern Classification, Schizophrenia Bulletin 38 (6) (2012) 1200–1215.
[18] J. Mourão-Miranda, D. R. Hardoon, T. Hahn, A. F. Marquand, S. C. R. Williams, J. Shawe-Taylor, M. Brammer, Patient classification as an outlier detection problem: an application of the One-Class Support Vector Machine., NeuroImage 58 (3) (2011) 793–804.
[19] D. C. Van Essen, S. Smith, D. Barch, T. Behrens, E. Yacoub, K. Ugurbil for the WU-Minn HCP Consortium., The WU-Minn human connectome project: An overview., NeuroImage 80 (2013) 62 – 79.
[20] T. Satterthwaite, M. Elliott, K. Ruparel, J. Loughead, K. Prabhakaran, M. Calkins, R. Hopson, C. Jackson, J. Keefe, M. Riley, F. Mentch, P. Sleiman, R. Verma, C. Davatzikos, H. Hakonarson, R. Gur, R. Gur, Neuroimaging of the Philadelphia neurodevelopmental cohort., Neuroimage. 86 (2014) 544 – 553.
[21] Y. Fan, D. Shen, R. Gur, R. Gur, C. Davatzikos, COMPARE: classification of morphological patterns using adaptive regional elements, IEEE Transaction on Medical Imaging (2007) 93 – 105.
[22] K. McGraw, S. Wong, Forming inferences about some intraclass correlation coefficients, Psychological Meth. 1 (1996) 30 – 46.
[23] L. Vincent, P. Soille, Watersheds in digital spaces: An efficient algorithm based on immersion simulations, IEEE Trans. Pattern Anal. Mach. Intell. 13 (6) (1991) 583 – 589.
[24] D. Shen, C. Davatzikos, HAMMER: Hierarchical attribute matching mechanism for elastic registration, IEEE Trans. Med. Imag. 21 (11) (2002) 1421 – 1439.
[25] V. N. Vapnik, The Nature of Statistical Learning Theory, Springer New York, 2000.
[26] B. B. Biswal, Resting state fmri: A personal history, Neuroimage 62 (2) (2012) 938 – 944.
[27] G. Tononi, O. Sporns, G. Edelman, A measure for brain complexity: Relating functional segregation and integration in the nervous system, Proceedings of the National Academy of Sciences of the United States of America (PNAS) 91 (1994) 5033 – 5037.
[28] S. M. Smith, K. L. Miller, G. Salimi-Khorshidi, M. Webster, C. F. Beckmann, T. E. Nichols, J. D. Ramsey, M. W. Woolrich, Network modelling methods for fmri, NeuroImage 54 (2011) 875–891.
[29] D. Cordes, V. Haughton, J. D. Carew, K. Arfanakis, K. Maravilla, Hierarchical clustering to measure connectivity in fmri resting-state data., Magnetic Resonance Imaging 20 (2002) 305–317.
[30] X. Shen, X. Papademetris, R. T. Constable, Graph-theory based parcellation of functional subunits in the brain from resting-state fmri data, NeuroImage 50 (2010) 1027 – 1035.
[31] R. Craddock, G. James, P. I. Holtzheimer, X. Hu, H. Mayberg, A whole brain fMRI atlas generated via spatially constrained spectral clustering, Human Brain Mapping 33 (8) (2012) 1914 – 1928.
[32] P. Bellec, P. Rosa-Neto, O. C. Lyttelton, H. Benali, A. C. Evans, Multi-level bootstrap analysis of stable clusters in resting-state fm, NeuroImage 51 (2010) 1126 – 1139.
[33] T. Blumensath, S. Jbabdi, M. F. Glasser, D. C. Van Essen, K. Ugurbil, T. E. Behrens, S. M. Smith, Spatially constrained hierarchical parcellation of the brain with resting-state fmri, NeuroImage 76 (2013) 313 – 324.
[34] R. Heller, D. Stanley, D. Yekutieli, N. Rubin, Y. Benjamini, Cluster-based analysis of fmri data., NeuroImage 33 (2) (2006) 599 – 608.
[35] S. Ryali, T. Chen, K. Supekar, V. Menon, A parcellation scheme based on von mises-fisher distributions and markov random fields for segmenting brain regions using resting-state fMRI, NeuroImage 65 (0) (2013) 83 – 96.
[36] P. Golland, Y. Golland, R. Malach, 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 Heidelberg, 2007, pp. 110–118.
[37] N. Honnorat, H. Eavani, T. Satterthwaite, R. Gur, R. Gur, Davatzikos.C., Grasp: Geodesic graph-based segmentation with shape priors for the functional parcellation of the cortex, NeuroImage 106, 207-221 106 (2015) 207 – 211.
[38] A. Delong, A. Osokin, H. Isack, Y. Boykov, Fast approximate energy minimization with label costs, International Journal of Computer Vision 96 (2012) 1 – 27.
[39] O. Veksler, Star shape prior for graph-cut image segmentation, in: IEEE European Conference on Computer Vision (ECCV), 2008, pp. 454 – 467.
[40] V. Gulshan, C. Rother, A. Criminisi, A. Blake, A. Zisserman, Geodesic star convexity for interactive image segmentation, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2010, pp. 3129 – 3136.
[41] K. J. Friston, C. D. Frith, P. F. Liddle, R. S. Frackowiak, Functional connectivity: the principal-component analysis of large (PET) data sets., Journal of Cerebral Blood Flow and Metabolism 13 (1) (1993) 5–14.
[42] S. C. Strother, J. R. Anderson, K. A. Schaper, J. J. Sidtis, J. S. Liow, R. P. Woods, D. A. Rottenberg, Principal component analysis and the scaled subprofile model compared to intersubject averaging and statistical paramet-

ric mapping: I. "Functional connectivity" of the human motor system studied with [15O]water PET., Journal of Cerebral Blood Flow and Metabolism 15 (5) (1995) 738–53.
[43] L. K. Hansen, J. Larsen, F. A. Nielsen, S. C. Strother, E. Rostrup, R. Savoy, N. Lange, J. Sidtis, C. Svarer, O. B. Paulson, Generalizable patterns in neuroimaging: how many principal components?, NeuroImage 9 (5) (1999) 534–44.
[44] M. J. McKeown, S. Makeig, G. G. Brown, T. P. Jung, S. S. Kindermann, a. J. Bell, T. J. Sejnowski, Analysis of fMRI data by blind separation into independent spatial components., Human brain mapping 6 (3) (1998) 160–88.
[45] V. Calhoun, G. Adali, and Pearlson, J. Pekar, A method for making group inferences from functional mri data using independent component analysis, Human Brain Mapping 14 (2001) 140 – 151.
[46] C. F. Beckmann, S. M. Smith, Probabilistic independent component analysis for functional magnetic resonance imaging., IEEE transactions on medical imaging 23 (2) (2004) 137–52.
[47] D. D. Lee, H. S. Seung, Learning the parts of objects by non-negative matrix factorization., Nature 401 (6755) (1999) 788–91.
[48] B. B. Avants, P. A. Cook, L. Ungar, J. C. Gee, M. Grossman, Dementia induces correlated reductions in white matter integrity and cortical thickness: a multivariate neuroimaging study with sparse canonical correlation analysis., NeuroImage 50 (3) (2010) 1004–16.
[49] I. Daubechies, E. Roussos, S. Takerkart, M. Benharrosh, C. Golden, K. D. Ardenne, W. Richter, J. D. Cohen, J. Haxby, Independent component analysis for brain fMRI does not select for independence, Proceedings of the National Academy of Sciences 106 (26) (2009) 10415–422.
[50] A. Sotiras, S. M. Resnick, C. Davatzikos, Finding imaging patterns of structural covariance via Non-Negative Matrix Factorization, NeuroImage 108 (2015) 1–16.
[51] D. Lashkari, E. Vul, N. Kanwisher, P. Golland, Discovering structure in the space of fMRI selectivity profiles, Neuroimage 50 (3) (2010) 1085 – 1098.
[52] H. Eavani, T. D. Satterthwaite, R. Filipovych, R. E. Gur, R. C. Gur, C. Davatzikos, Identifying sparse connectivity patterns in the brain using resting-state fmri, NeuroImage 105 (2015) 286–299.
[53] V. Calhoun, T. Adali, L. Hansen, J. Larsen, J. Pekar, Ica of functional mri data: an overview, in: in Proceedings of the International Workshop on Independent Component Analysis and Blind Signal Separation, 2003, pp. 281–288.
[54] S. M. Smith, K. L. Miller, S. Moeller, J. Xu, E. J. Auerbach, M. W. Woolrich, C. F. Beckmann, M. Jenkinson, J. Andersson, M. F. Glasser, D. C. Van Essen, D. A. Feinberg, E. S. Yacoub, K. Ugurbil, Temporally-independent functional modes of spontaneous brain activity, Proceedings of the National Academy of Sciences of the United States of America (PNAS) 109 (8) (2012) 3131 – 3136.
[55] C. J. Burges, A tutorial on support vector machines for pattern recognition, Data mining and knowledge discovery 2 (2) (1998) 121–167.
[56] B. Gaonkar, C. Davatzikos, Analytic estimation of statistical significance maps for support vector machine based multi-variate image analysis and classification, NeuroImage 78 (2013) 270–283.
[57] N. K. Batmanghelich, B. Taskar, C. Davatzikos, Generative-discriminative basis learning for medical imaging, Medical Imaging, IEEE Transactions on 31 (1) (2012) 51–69.
[58] E. Varol, B. Gaonkar, C. Davatzikos, Classifying medical images using morphological appearance manifolds, in: Biomedical Imaging (ISBI), 2013 IEEE 10th International Symposium on, IEEE, 2013, pp. 744–747.
[59] S. Haufe, F. Meinecke, K. Görgen, S. Dähne, J. D. Haynes, B. Blankertz, F. Bießmann, On the interpretation of weight vectors of linear models in multivariate neuroimaging, NeuroImage 87 (2014) 96–110.
[60] B. Gaonkar, R. T. Shinohara, C. Davatzikos, A. D. N. Initiative, et al., Interpreting support vector machine models for multivariate group wise analysis in neuroimaging, Medical image analysis 24 (1) (2015) 190–204.
[61] D. H. Geschwind, P. Levitt, Autism spectrum disorders: developmental disconnection syndromes., Current opinion in neurobiology 17 (1) (2007) 103–11.
[62] S. S. Jeste, D. H. Geschwind, Disentangling the heterogeneity of autism spectrum disorder through genetic findings., Nature reviews. Neurology 10 (2) (2014) 74–81.
[63] R. W. Buchanan, W. T. Carpenter, Domains of psychopathology: an approach to the reduction of heterogeneity in schizophrenia, The Journal of nervous and mental disease 182 (4) (1994) 193–204.
[64] N. Koutsouleris, C. Gaser, M. Jäger, R. Bottlender, T. Frodl, S. Holzinger, G. J. E. Schmitt, T. Zetzsche, B. Burgermeister, J. Scheuerecker, C. Born, M. Reiser, H.-J. Möller, E. M. Meisenzahl, Structural correlates of psychopathological symptom dimensions in schizophrenia: a voxel-based morphometric study., NeuroImage 39 (4) (2008) 1600–12.
[65] T. Zhang, N. Koutsouleris, E. Meisenzahl, C. Davatzikos, Heterogeneity of Structural Brain Changes in Subtypes of Schizophrenia Revealed Using Magnetic Resonance Imaging Pattern Analysis, Schizophrenia Bulletin 41 (1) (2015) 74–84.
[66] J. M. Graham, H. J. Sagar, A data-driven approach to the study of heterogeneity in idiopathic parkinson's disease: identification of three distinct subtypes, Movement Disorders 14 (1) (1999) 10–20.
[67] S. J. G. Lewis, T. Foltynie, A. D. Blackwell, T. W. Robbins, A. M. Owen, R. A. Barker, Heterogeneity of Parkin-

son's disease in the early clinical stages using a data driven approach., Journal of neurology, neurosurgery, and psychiatry 76 (3) (2005) 343–8.

[68] M. E. Murray, N. R. Graff-Radford, O. A. Ross, R. C. Petersen, R. Duara, D. W. Dickson, Neuropathologically defined subtypes of Alzheimer's disease with distinct clinical characteristics: a retrospective study., The Lancet. Neurology 10 (9) (2011) 785–96.

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

[70] C. Huang, L. O. Wahlund, O. Almkvist, D. Elehu, L. Svensson, T. Jonsson, B. Winblad, P. Julin, Voxel- and VOI-based analysis of SPECT CBF in relation to clinical and psychological heterogeneity of mild cognitive impairment, NeuroImage 19 (3) (2003) 1137–1144.

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

[72] A. Dong, N. Honnorat, B. Gaonkar, C. Davatzikos, CHIMERA: Clustering of heterogeneous disease effects via distribution matching of imaging patterns, Medical Imaging, IEEE Transactions on PP (99) (2015) 1–1.

[73] A. Myronenko, X. Song, Point set registration: Coherent point drift, PAMI 32 (2010) 2262–2275.

[74] T. K. Moon, The expectation-maximization algorithm, Signal processing magazine, IEEE 13 (6) (1996) 47–60.

[75] J. Ashburner, K. J. Friston, Voxel-based morphometry the methods, Neuroimage 11 (6) (2000) 805–821.

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

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

[78] E. Varol, A. Sotiras, C. Davatzikos, HYDRA: revealing heterogeneity of imaging and genetic patterns through a multiple max-margin discriminative analysis framework, Neuroimage – (2015) –.

[79] J.-C. Lambert, C. A. Ibrahim-Verbaas, D. Harold, A. C. Naj, R. Sims, C. Bellenguez, G. Jun, A. L. DeStefano, J. C. Bis, G. W. Beecham, et al., Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for alzheimer's disease, Nature genetics 45 (12) (2013) 1452–1458.

[80] L. Hubert, P. Arabie, Comparing partitions, Journal of classification 2 (1) (1985) 193–218.

[81] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, G. E. Hinton, Adaptive mixtures of local experts, Neural computation 3 (1) (1991) 79–87.

[82] L. Ladicky, P. Torr, Locally linear support vector machines, in: Proceedings of the 28th International Conference on Machine Learning (ICML-11), 2011, pp. 985–992.

[83] Z. Fu, A. Robles-Kelly, J. Zhou, Mixing linear svms for nonlinear classification, Neural Networks, IEEE Transactions on 21 (12) (2010) 1963–1975.

[84] H. Eavani, M. K. Hsieh, Y. An, G. Erus, L. Beason-Held, S. Resnick, C. Davatzikos, Capturing heterogeneous group differences using mixture-of-experts: Application to a study of aging, NeuroImage.

[85] X. Da, J. B. Toledo, J. Zee, D. A. Wolk, S. X. Xie, Y. Ou, A. Shacklett, P. Parmpi, L. Shaw, J. Q. Trojanowski, C. Davatzikos, 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 (2014) 164–173.

[86] C. Davatzikos, F. Xu, Y. An, Y. Fan, S. M. Resnick, Longitudinal progression of Alzheimer's-like patterns of atrophy in normal older adults: the SPARE-AD index., Brain 132 (Pt 8) (2009) 2026–35.