Chapter 10
Analysis of Brain Magnetic Resonance (MR) Scans for the Diagnosis of Mental Illness

Aydın Ulaş, Umberto Castellani, Manuele Bicego, Vittorio Murino, Marcella Bellani, Michele Tansella, and Paolo Brambilla

Abstract We address the problem of schizophrenia detection by analyzing magnetic resonance imaging (MRI). In general, mental illness like schizophrenia or bipolar disorders are traditionally diagnosed by self-reports and behavioral observations. A new trend in neuroanatomical research consists of using MRI images to find possible connections between cognitive impairments and neuro-physiological abnormalities. Indeed, brain imaging techniques are appealing to provide a non-invasive diagnostic tool for mass analyses and early diagnoses. The problem is challenging due to the heterogeneous behavior of the disease and up to now, although the literature is large in this field, there is not a consolidated framework to deal with it. In this context, advanced pattern recognition and machine learning techniques can

A. Ulaş · U. Castellani · M. Bicego (✉) · V. Murino
Departimento di Informatica, University of Verona, Verona, Italy
e-mail: manuele.bicego@univr.it

A. Ulaş
e-mail: mehmetaydin.ulas@univr.it

U. Castellani
e-mail: umberto.castellani@univr.it

V. Murino
e-mail: vittorio.murino@univr.it

V. Murino
Istituto Italiano di Tecnologia (IIT), Genova, Italy

M. Bellani · M. Tansella · P. Brambilla
Department of Public Health and Community Medicine, Section of Psychiatry and Clinical Psychology, Inter-University Centre for Behavioural Neurosciences, University of Verona, Verona, Italy

M. Bellani
e-mail: marcella.bellani@univr.it

M. Tansella
e-mail: michele.tansella@univr.it

P. Brambilla
e-mail: paolo.brambilla@univr.it

M. Pelillo (ed.), *Similarity-Based Pattern Analysis and Recognition*,
Advances in Computer Vision and Pattern Recognition,
DOI 10.1007/978-1-4471-5628-4_10, © Springer-Verlag London 2013

be useful to improve the automatization of the involved procedures and the characterization of mental illnesses with specific and detectable brain abnormalities. In this book, we have exploited similarity-based pattern recognition techniques to further improve brain classification problem by employing the algorithms developed in the other chapters of this book. (This chapter is based on previous works (Castellani et al. in Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, MICCAI'11, vol. 6892, pp. 426–433, 2011; Gönen et al. in Proceedings of the International Workshop on Similarity-Based Pattern Analysis, SIMBAD'11, vol. 7005, pp. 250–260, 2011; Ulaş et al. in Proceedings of the Iberoamerican Congress on Pattern Recognition, CIARP'11, vol. 7042, pp. 491–498, 2011; in IAPR International Conference on Pattern Recognition in Bioinformatics, PRIB'11, vol. 7036, pp. 306–317, 2011; and in Int. J. Imaging Syst. Technol. 21(2):179–192, 2011) by the authors and contains text, equations and experimental results taken from these papers.)

10.1 Introduction

Brain analysis techniques using Magnetic Resonance Imaging (MRI) are playing an increasingly important role in understanding pathological structural alterations of the brain [33, 70]. The ultimate goal is to identify structural brain abnormalities by comparing normal subjects with patients affected by a certain disease. Here, we focus on schizophrenia. Schizophrenia is a heterogeneous psychiatric disorder characterized by several symptoms such as hallucinations, delusions, cognitive and thought disorders [8]. Although genetic and environmental factors play a role in the disorder, its etiology remains unknown and substantial body of research has demonstrated numerous structural and functional brain abnormalities in patients with both chronic and acute forms of the disorder [66, 70].

Our main contribution here is to deal with schizophrenia detection as a binary classification problem—we have to distinguish between normal subjects and patients affected by schizophrenia [25]—by applying advanced pattern recognition techniques by exploiting the capability of similarity-based methods mentioned in the other chapters of this book to this problem.

We highlight that the problem of schizophrenia detection is very complex since the symptoms of the disease are different and related to different properties of the brain. Thus, although the literature has shown a large amount of promising methodological procedures to address this disease, up to now a consolidate framework is not available.

In this chapter, we have exploited different approaches to address schizophrenia detection. We have defined a general working pipeline composed of the four main steps: (i) data acquisition, (ii) region selection, (iii) data description, and (iv) classification. Each step may be instantiated in different ways, each one having pros and cons. Here, for each stage, we summarize the possible choices we adapted. Figure 10.1 shows the proposed overall scheme of the working pipeline and the involved possibilities. In summary:

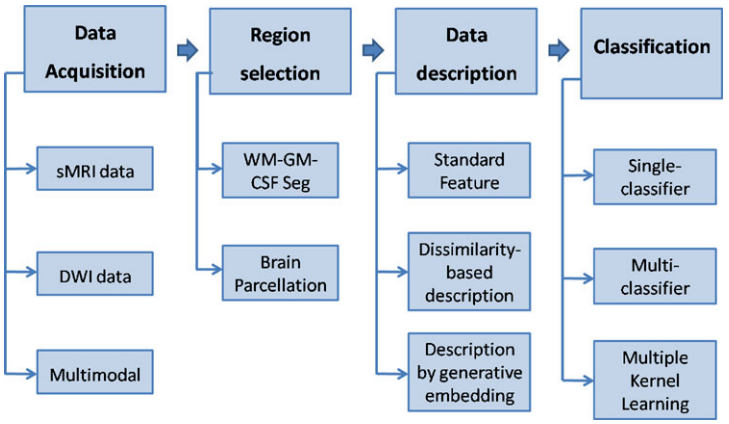

Fig. 10.1 Overall scheme of the proposed working pipeline

- *Data acquisition* regards the imaging technique employed to acquire data. Different acquisition modalities are encoding different brain information. We use Structural MRI to deal with morphological properties, and Diffusion Weighted Imaging (DWI) to evaluate functional aspects of the brain. Moreover, in order to integrate different sources of information, a multimodal approach is exploited.
- *Region selection* is necessary to focus the analysis on brain subparts. A common approach is to segment the whole brain among *White Matter* (WM), *Gray Matter* (GM), and *Cerebro-Spinal Fluid* (CSF). Another approach consists in extracting one or more Regions of Interest (ROIs) which are strictly related to the analyzed disease. The brain segmentation in ROIs is in general called *brain parcellation*.
- *Data description* aims at extracting the most useful information for the involved task, in our case brain classification. The standard approach consists in using *features*. According to the overall aim of this book, we exploited the possibility to go beyond features. Indeed, we have investigated two paradigms, derived from other chapters of the book, a dissimilarity-based description (Chap. 2) and description by generative embeddings (Chap. 4).
- *Classification* is the last step of the proposed pipeline. As simplest approach a single classifier has been employed. In order to integrate different sources of information at classification stage, we exploited two paradigms, multi-classifier approach and multiple kernel learning.

Roadmap The chapter is organized as follows: In Sect. 10.2, we present the state-of-the-art in schizophrenia detection. In Sects. 10.3 and 10.4, we introduce data acquisition and region selection, respectively. Then, data description phase is split into standard features (Sect. 10.5.1), dissimilarity-based description (Sect. 10.5.2), and description by generative embeddings (Sect. 10.5.3). We define our approaches of classification using ensembles and Multiple Kernel Learning in Sect. 10.6. We explain three case studies which utilize the working pipeline in Sects. 10.7, 10.8, and 10.9; and conclude in Sect. 10.10.

10.2 Related Work

Several works have been proposed for human brain classification in the context of schizophrenia research [70]. In the following, we have organized the state-of-the-art in (i) *shape-based* techniques and (ii) *classification-based* techniques.

10.2.1 Shape-Based Techniques

Standard approaches are based on detecting morphological differences on certain brain regions, namely Region Of Interests (ROIs). Usually, the aim is the observation of volume variations [7, 59, 70]. In general, ROI-based techniques require the manual tracing of brain subparts. In order to avoid such an expensive procedure, Voxel Based Morphometry (VBM) techniques have been introduced [4, 41] for which the entire brain is transformed onto a template, namely the stereotaxic space. In this fashion, a voxel-by-voxel correspondence is available for comparison purposes. In [41], a multivariate Voxel Based Morphometry approach method is proposed to differentiate schizophrenic patients from normal controls. Inferences about the structural relevance of gray matter distribution are carried out on several brain sub-regions. In [85], cortical changes in adolescent on-set schizophrenic patients are analyzed by combining Voxel-Based with Surface-Based Morphometry (SBM). A different approach consists in encoding the shape by a *global* region descriptor [32, 63, 76]. In [76], a new morphological descriptor is introduced by properly encoding both the displacement fields and the distance maps for amygdala and hippocampus. In [32], a ROI-based morphometric analysis is introduced by defining spherical harmonics and 3D skeleton as shape descriptors. Improvement of such a shape-descriptor-based approach with respect to classical volumetric techniques is shown experimentally. Although results are interesting, the method is not invariant to surface deformations and therefore it requires shape registration and data resampling. This pre-processing is avoided in [63], where the so-called Shape-DNA signature has been introduced by taking the eigenvalues of the Laplace–Beltrami operator as region descriptor for both the external surface and the volume. Although *global* methods can be satisfying for some classification tasks, they do not provide information about the localization of the morphological anomalies. To this aim, *local* methods have been proposed. In [77], the so called *feature*-based morphometry (FBM) approach is introduced. Taking inspiration from feature-based techniques proposed in computer vision, FBM identifies a subset of features corresponding to anatomical brain structures that can be used as disease biomarkers.

10.2.2 Classification-Based Techniques

In order to improve the capability of distinguishing between healthy and non-healthy subjects, learning-by-example techniques [27] are applied (see, for example, [25]).

Usually, geometric signatures extracted from the MRI data are used as feature vectors for classification purposes [30, 58, 87]. In [87], a support vector machine (SVM) has been employed to classify cortical thickness which has been measured by calculating the Euclidean distance between linked vertices on the inner and outer cortical surfaces. In [30], a new approach has been defined by combining deformation-based morphometry with SVM. In this fashion, multivariate relationships among various anatomical regions have been captured to characterize more effectively the group differences. Finally, in [58], a unified framework is proposed to combine advanced probabilistic registration techniques with SVM. The local spatial warps parameters are also used to identify the discriminative warp that best differentiates the two groups. It is worth to note that in most of the mentioned works, the involved classifier was a Support Vector Machine, but more general approaches are also proposed, see, e.g., [51]. Here, a set of image features which encode both general statistical properties and Law's texture features from the whole brain are analyzed. Such features are concatenated onto a very high dimensional vector which represents the input for a classic learning-by-example classification approach. Several classifiers are then evaluated such as decision trees or decision graphs. In [15], the authors proposed a neural network to measure the relevance of thalamic subregions implicated in schizophrenia. The study is based on the metabolite N-acetylaspartate (NAA) using in vivo proton magnetic resonance spectroscopic imaging. The diffusion of water in the brain characterized by its apparent diffusion coefficient (ADC), which represents the mean diffusivity of water along all directions gives potential information about the size, orientation, and tortuosity of both intracellular and extracellular spaces, providing evidence of disruption when increased [64]. DWI has been shown to be keen in exploring the microstructural organization of white matter, therefore providing intriguing information on brain connectivity [13, 78].

10.3 Data Acquisition

The data set involves a 124 subject database cared by the Research Unit on Brain Imaging and Neuropsychology (RUBIN) at the Department of Medicine and Public Health-Section of Psychiatry and Clinical Psychology of the University of Verona. The data set is composed of MRI brain scans of 64 patients recruited from the area of South Verona (i.e., 100,000 inhabitants) through the South Verona Psychiatric Case Register [2, 3, 82]. Additionally, 60 individuals without schizophrenia (control subjects) were also recruited.

10.3.1 MRI Data

MRI scans were acquired with a 1.5 T Magnetom Symphony Maestro Class Syngo MR 2002B (Siemens), and in total, it took about 19 minutes to complete an MR

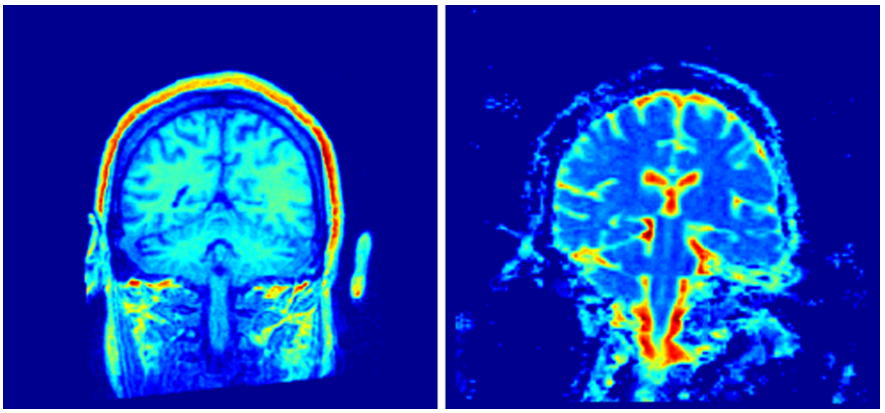

Fig. 10.2 Slices acquired by 3D Morphological technique (*left*) and Diffusion Weighting Imaging technique (*right*)

session. A standard head coil was used for radio frequency transmission and reception of the MR signal, and restraining foam pads were used to minimize head motion. T1-weighted images were first obtained to verify the participants head position and image quality (TR = 450 ms, TE = 14 ms, flip angle = 90°, FOV = 230 × 230, 18 slices, slice thickness = 5 mm, matrix size = 384 × 512, NEX = 2). Proton density (PD)/T2-weighted images were then acquired (TR = 2500 ms, TE = 24/121 ms, flip angle = 180°, FOV = 230 × 230, 20 slices, slice thickness = 5 mm, matrix size = 410 × 512, NEX = 2) according to an axial plane running parallel to the anterior–posterior (AC–PC) commissures to exclude focal lesions. Subsequently, a coronal 3-dimensional magnetization prepared rapid gradient echo (MP-RAGE) sequence was acquired (TR = 2060 ms, TE = 3.9 ms, flip angle = 15°, FOV = 176 × 235, slice thickness = 1.25 mm, matrix size = 270 × 512, inversion time = 1100) to obtain 144 images covering the entire brain. In Fig. 10.2 (left), we can see a slice of a subject acquired by using MRI.

10.3.2 DWI Data

Diffusion-weighted imaging (DWI) investigates molecular water mobility within the local tissue environment, providing information on tissue microstructural integrity. The diffusion of water in the brain is characterized by its apparent diffusion coefficient (ADC), which represents the mean diffusivity of water along all directions [75]. Thus, ADC gives potential information about the size, orientation, and tortuosity of both intracellular and extracellular spaces, providing evidence of disruption when increased [64]. ADC has also been used to explore regional grey matter microstructure, being higher in the case of potential neuron density alterations or volume deficit [62].

Diffusion-weighted echoplanar images in the axial plane parallel to the AC–PC line (TR = 3200 ms, TE = 94 ms, FOV = 230 × 230, 20 slices, slice thickness = 5 mm with 1.5-mm gap, matrix size = 128 × 128, echo-train length = 5; these parameters were the same for $b = 0$, $b = 1000$, and the ADC maps). Specifically, three gradients were acquired in three orthogonal directions. ADC maps (denoted by D_{ADC}) were obtained from the diffusion images with $b = 1000$, according to the following equation:

$$-bD_{\mathrm{ADC}} = \ln\left[A(b)/A(0)\right],$$

where $A(b)$ is the measured echo magnitude, b is the measure of diffusion weighting, and $A(0)$ is the echo magnitude without diffusion gradient applied. In Fig. 10.2 (right), we can see a slice of a subject acquired by using DWI.

10.3.3 Multimodal Approach

A multimodal approach can be applied when different kinds of acquisition procedures are used for the same subject. As can be seen in Fig. 10.2, while MRI images are more reliable, DWI resolution is very low and it's hard to segment ROIs from these DWI images. In order to integrate such data, a *co-registration* procedure is necessary.

The co-registration consists in matching high-resolution (also known as T1-w) and DWI images defined in different coordinate systems. Open source libraries of National Library of Medicine *Insight Segmentation and Registration Toolkit* are adapted for the co-registration procedure, while Tcl/Tk code and VTK open source libraries are chosen for the graphic interface. Digital Imaging and Communications in Medicine format (DICOM) tag parameters necessary for the co-registration are: Image Origin, Image Spacing, Patient Image Orientation, and Frame of Reference.

Assuming the same anatomy topology for different studies, a Mutual Information technique based on Mattes algorithm is applied. An in-house software for multimodal registration was developed. The program 3D Slicer,[1] a free open source software for visualization and image computing, is employed for the graphic interface. The process was performed in several steps.

The source DWI study (*moving image*, see Fig. 10.3) is aligned through a roto-translational matrix with the T1-w data (*fixed image*); the two studies are acquired in straight succession with the same MR unit without patient repositioning; the parameters related to algorithm implementation are automatically defined; then, by applying a multi-resolution pyramid, we are able to reach a registration within eight iterations avoiding local minimal solution.

The results of the registration are visually inspected in a checkerboard, where each block alternately displayed data from each study, verifying alignment of

[1] http://www.slicer.org/.

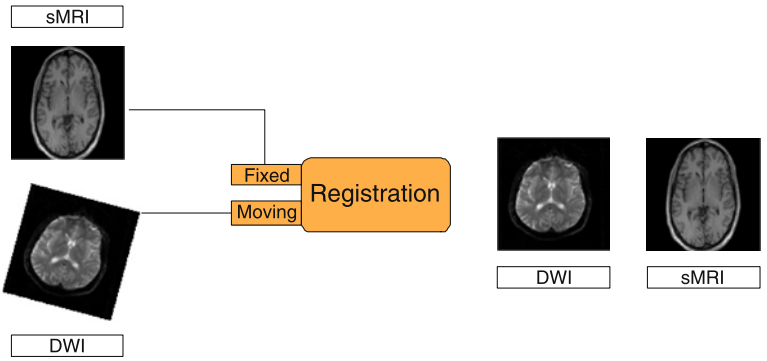

Fig. 10.3 Registration of sMRI to DWI

anatomical landmarks (ventricles, etc.) for confirmation. This procedure is needed because sMRI images have better resolution and the anatomy can better be seen for manual ROI segmentation. We use this procedure to extract ADC values for each of the ROIs instead of the whole image. Once the co-registration is carried out, a direct voxel-by-voxel comparison between the two data modalities becomes feasible and therefore any joint feature can be extracted.

10.4 Region Selection

The brain is a complex organ composed of different kinds of tissues related to different physiological properties of brain matter. Moreover, the brain can be segmented into well defined anatomical structures which are associated to specific functions of the brain. In order to improve the search of brain abnormalities, it is important to take into account of such kind of brain subdivisions. Two main paradigms are in general defined: (i) White matter (WM), Gray matter (GM), and Cerebrospinal Fluid (CSF) segmentation, and (ii) brain parcellation.

10.4.1 WM–GM–CSF Segmentation

WM–GM–CSF segmentation aims at decomposing the brain into its main kinds of tissues (see Fig. 10.4). In particular, white matter encloses mainly the axons which connect different parts of the brain, while gray matter contains neural cell bodies. Cerebrospinal fluid is a clear, colorless bodily fluid that occupies the ventricular system around and inside the brain and sulci.

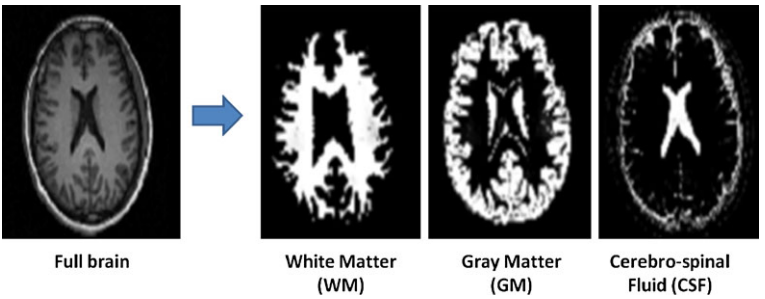

Fig. 10.4 Example of brain segmentation among White Matter (WM), Gray Matter (GM), and Cerebrospinal Fluid (CSF)

10.4.2 Brain Parcellation

The raw images acquired using a 1.5 T MRI machine have 144 slices and 384 × 512 resolution. These images are then transferred to PC workstations in order to be processed for ROI *tracing* which we adapted. Based on manual identification of landmarks, these slices are resampled and realigned by the medical personnel using the Brains2[2] software. The same software is used to manually trace the ROIs by drawing contours enclosing the intended region. This was carried out by a trained expert following a specific protocol for each ROI [7] without knowledge of the class labels. There are methods which automatically segment the ROIs, but their accuracy is lower than the manual methods so manual segmentation was preferred. The ROIs traced are 7 pairs (for the left and the right hemisphere, respectively) of disconnected image areas:

- Amygdala (*lamyg* and *ramyg*, in short)
- Dorso-lateral PreFrontal Cortex (*ldlpfc* and *rdlpfc*)
- Entorhinal Cortex (*lec* and *rec*)
- Heschl's Gyrus (*lhg* and *rhg*)
- Hippocampus (*lhippo* and *rhippo*)
- Superior Temporal Gyrus (*lstg* and *rstg*)
- Thalamus (*lthal* and *rthal*)

We select these ROIs because they have consistently been found to be impaired in schizophrenia and in a recent work, some of them have been found to support a specific altered neural network [21]. The Inter Rater Reliability (IRR) values for each brain hemisphere and ROI can be seen in Table 10.1 which shows us the reliability of the segmentation. Higher value means the segmentation is more reliable.

Additionally, another important ROI that is traced is the *intracranial volume* (ICV), that is the volume occupied by the brain in the cranial cavity leaving out the brainstem and the cerebellum. This information is extremely useful for normalizing volume values against differing overall brain sizes.

[2] http://www.psychiatry.uiowa.edu/mhcrc/IPLpages/BRAINS.htm.

Table 10.1 IRR values for ROI segmentation

ROI	left	right
amyg	0.91	0.98
dlpfc	0.93	0.98
ec	0.94	0.96
hg	0.96	0.98
hippo	0.96	0.96
stg	0.93	0.99
thal	0.95	0.96

10.5 Data Description

In this section, we show how we describe the data to be used in classification.

10.5.1 Standard Features

In order to encode useful information in a compact representation, data descriptors are employed. The overall idea consists of representing the brain with a signature which summarizes brain characteristics, and using such signature for comparison purposes. We exploited several kinds of brain characteristics; each of them focusing on a specific aspect of the brain. In particular, we have employed histogram of image *intensities* to encode tissue characteristics, and *geometric* features to concentrate the analysis on shape properties of brain structures. We highlight that, according to standard feature-based approach, such descriptors could be directly used for brain classification. Since we aim at going beyond features in this book, we have exploited the new paradigm to deal with such brain characteristics by proposing new approaches for data description (as we will explain in Sects. 10.5.2 and 10.5.3).

In the following, we introduce (i) Intensity Histograms of sMRI, (ii) Histograms of Apparent Diffusion Coefficient values, (iii) basic geometric shape descriptors, and (iv) spectral shape descriptor.

10.5.1.1 Intensity Histograms of Structural MRI Images

From MRI data we compute scaled histograms of image intensities. In particular, we compute a histogram for each ROI. A major disadvantage of MRI compared to other imaging techniques is the fact that its intensities are not standardized. Even MR images taken for the same patient on the same scanner with the same protocol at different times may differ in content due to a variety of machine-dependent reasons, therefore, image intensities do not have a fixed meaning [54]. This implies a significant effect on the accuracy and precision of the following image processing, analysis, segmentation, and registration methods relying on intensity similarity.

A successful technique used to calibrate MR signal characteristics at the time of acquisition employs *phantoms* [29], by placing physical objects with known attributes within the scanning frame. Unfortunately, this technique is not always exploited, which is our present case. Alternatively, it is possible to apply bias correction (using software like SPM[3] or FSL[4]) for the image intensities, and apply intensity rescaling afterwards. Here, we rescale intensities based on landmark matching from the ICV histograms [54] because it is easier to identify landmarks on the histograms that match the canonical subdivision of intracranial tissue into white matter, gray matter and cerebrospinal fluid. We select a rescaling mapping that conserves most of the signal in the gray matter—white matter area, corresponding to the two highest bumps in the range 60–90, since ROIs primarily contain those kinds of tissue.

10.5.1.2 Histograms of Apparent Diffusion Coefficient values

Although we don't have manually segmented ROIs for DWI images, we used a co-registration procedure to segment DWI images into ROIs. For this purpose, every subject's DWI image was registered into the corresponding structural MRI image. Then Apparent Diffusion Coefficient (ADC) values are calculated using these images. We form the histograms of ADC values and use them in our experiments. Since the ADC values are already normalized, we don't need to do another step of normalization on ADC histograms.

10.5.1.3 Basic Geometric Shape Descriptors

From the set of 2D ROIs of the shapes (slices) the 3D surface is computed as triangle mesh using marching cubes. A minimal smoothing operation is applied to remove noise and voxelization effect. We encode geometric properties of the surface using the *Shape Index* [44], which is defined as:

$$\mathrm{si} = -\frac{2}{\pi}\arctan\left(\frac{k_1 + k_2}{k_1 - k_2}\right), \quad k_1 > k_2,$$

where k_1, k_2 are the principal curvatures of a generic surface point. The Shape Index varies in $[-1, 1]$ and provides a local categorization of the shape into primitive forms such as spherical cap and cup, rut, ridge, trough, or saddle [44]. Shape index is pose and scale invariant [44] and it has already been successfully employed in biomedical domain [5]. The shape index is computed at each vertex of the extracted mesh. Then, all the values are quantized and a histogram of occurrences is computed. Such histograms represent the descriptor of a given subject and it basically encodes the brain local geometry of a subject, disregarding the spatial relationships.

[3] http://www.fil.ion.ucl.ac.uk/spm/.

[4] http://www.fmrib.ox.ac.uk/fsl/.

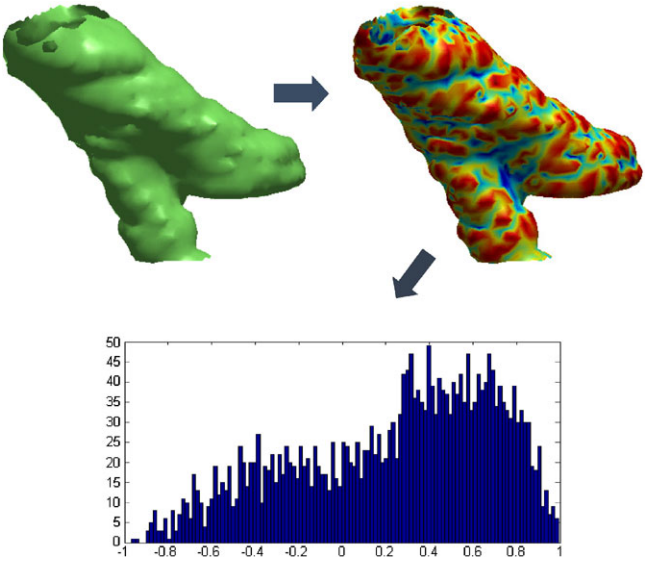

Fig. 10.5 Geometric feature extraction: 3D surface of the thalamus (*left*), the surface colored according with Shape Index values (*right*), and the histogram of Shape Index occurrences (*bottom*)

Figure 10.5 shows the 3D surface of the left-Thalamus (left), the surface colored according with Shape Index values (right), and the histogram of Shape Index occurrences (bottom). It is worth noting that convex regions (in blue) are clearly distinguished from concave regions (in red) by the Shape Index values. As a further step we also calculate the mean curvature using the same methodology:

$$m = \frac{k_1 + k_2}{2}.$$

10.5.1.4 Spectral Shape Descriptor

In this section, we describe a new shape descriptor, which is based on advanced *diffusion* geometry techniques. Considering a shape M as a compact Riemannian manifold [14], the heat diffusion on shape[5] is defined by the *heat* equation:

$$\left(\Delta_M + \frac{\partial}{\partial t}\right) u(t, \boldsymbol{m}) = 0, \tag{10.1}$$

where u is the distribution of heat on the surface, $\boldsymbol{m} \in M$, Δ_M is the *Laplace–Beltrami* operator which, for compact spaces, has discrete eigendecomposition of

[5] In this section, we borrow the notation from [14, 73].

the form $\Delta_M \phi_i = \lambda_i \phi_i$. In this way, the *heat kernel* has the following eigendecomposition:

$$h_t(\boldsymbol{m}, \boldsymbol{m}') = \sum_{i=0}^{\infty} e^{-\lambda_i t} \phi_i(\boldsymbol{m}) \phi_i(\boldsymbol{m}'), \tag{10.2}$$

where λ_i and ϕ_i are the ith eigenvalue and the ith eigenfunction of the Laplace–Beltrami operator, respectively. The heat kernel $h_t(\boldsymbol{m}, \boldsymbol{m}')$ is the solution of the heat equation with initial point heat source in $\boldsymbol{m}$ at time $t = 0$, and heat value in ending point $\boldsymbol{m}' \in M$ after time t. The heat kernel is *isometrically invariant*, it is *informative*, and *stable* [73].

In the case of volumetric representations, the volume is sampled by a regular Cartesian grid composed by voxels, which allows the use of standard Laplacian in $\mathbb{R}^3$ as the Laplace–Beltrami operator. We use finite differences to evaluate the second derivative in each direction of the volume. The heat kernel on volumes is invariant to volume isometries, in which shortest paths between points inside the shape do not change. Note that in real applications, exact volume isometries are limited to the set of rigid transformations [61], however, also non-rigid deformations can faithfully be modeled as approximated volume isometries in practice. It is also worth noting that, as observed in [61, 73], for small t the autodiffusion heat kernel $h_t(\boldsymbol{m}, \boldsymbol{m})$ of a point $\boldsymbol{m}$ with itself is directly related to the *scalar* curvature $s(\boldsymbol{m})$ [61]. More formally,

$$h_t(\boldsymbol{m}, \boldsymbol{m}) = (4\pi t)^{-3/2} \left(1 + \frac{1}{6} s(\boldsymbol{m})\right). \tag{10.3}$$

In practice, Eq. (10.3) states that the heat tends to diffuse slower at points with positive curvature, and vice-versa. This gives an intuitive explanation about the geometric properties of $h_t(\boldsymbol{m}, \boldsymbol{m})$, and suggests the idea of using it to build a shape descriptor [73].

Global Heat Kernel Signature Once data are collected, a strategy to encode the most informative properties of the shape M can be devised. To this end, a global shape descriptor is proposed, which is inspired by the so-called *Heat Kernel Signature* (HKS) defined as:

$$\mathrm{HKS}(x) = \left[h_{t_0}(x, x), \ldots, h_{t_n}(x, x)\right], \tag{10.4}$$

where x is a point of the shape (i.e., a vertex of a mesh or a voxel) and $(t_0, t_1, \ldots, t_n)$ are n time values. To extend this approach to the whole shape, we introduce the following global shape descriptor:

$$\mathrm{GHKS}(M) = \left[hist\left(H_{t_0}(M)\right), \ldots, hist\left(H_{t_n}(M)\right)\right], \tag{10.5}$$

where $H_{t_i}(M) = \{h_{t_i}(x, x), \forall x \in M\}$, and $hist(\cdot)$ is the histogram operator. Note that our approach combines the advantages of [14, 61] since it encodes the distribution of local heat kernel values and it works at multiscales. GHKS allows for shape

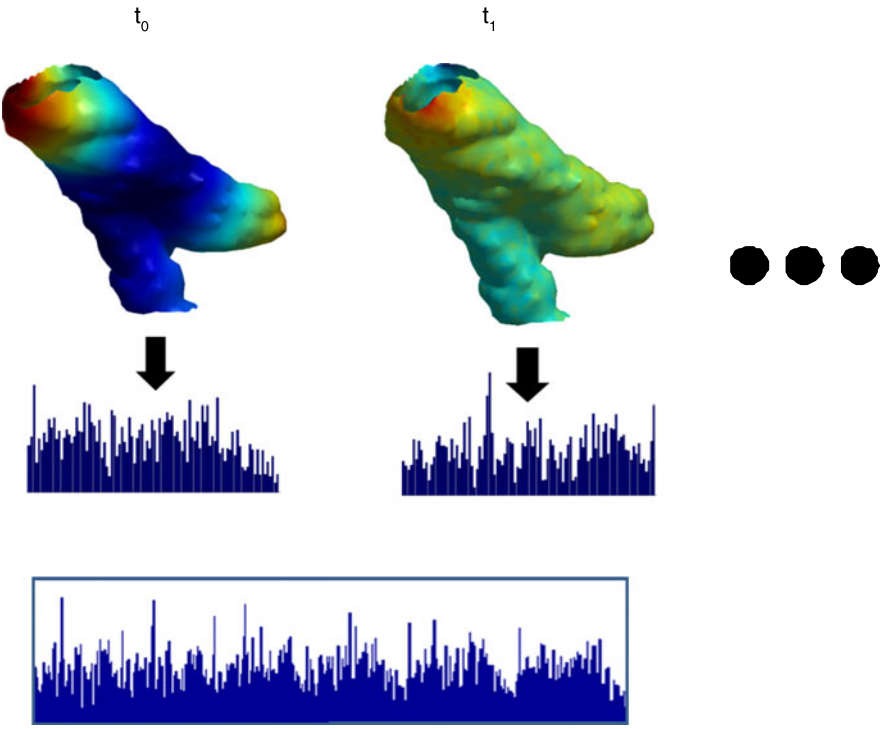

Fig. 10.6 GHKS: Each point of the shape is colored according to $h_{t_i}(x, x)$. Such values are collected into a histogram for each scale t_i. Finally, histograms are concatenated leading to the global signature

comparisons using minimal shape preprocessing, in particular, no registration, mapping, or remeshing is necessary. GHKS is robust to noise since it implicitly employs surface smoothing by neglecting higher frequencies of the shape. Finally, GHKS is able to encode isometric invariance properties of the shape [73] which are crucial to deal with shape deformations. Figure 10.6 shows a scheme of the proposed descriptor. Each point of the shape is colored according to $h_{t_i}(x, x)$. Such values are collected into a histogram for each scale t_i. Finally, histograms are concatenated leading to the global signature.

10.5.2 Descriptors on Dissimilarity Space

In this section, we describe data descriptors generated by employing similarity-based approach. In general, similarity-based approach aims at exploiting the discriminative properties of similarity measures per se, as opposed to standard feature-based approach. In fact, the similarity-based paradigm differs from typical pattern

recognition approaches where objects to be classified are represented by feature vectors. Devising pattern recognition techniques starting from similarity measures is a real challenge, and the main idea of this book. Among the different proposed techniques, in this work we investigated the use of the dissimilarity-based representation paradigm, introduced by Pekalska and Duin [55] and described in Chap. 2. Within this approach, objects are described using pairwise (dis)similarities to a representation set of objects. This offers the analyst a different way to express application background knowledge as compared to features. In a second step, the dissimilarity representation is transformed into a vector space in which traditional statistical classifiers can be employed. Unlike the related kernel approach, whose application is often restrained by technicalities like fulfilling Mercer's condition, basically any dissimilarity measure can be used.

Similarity-based approach is more versatile in dealing with different data representations (i.e., images, MRI volume, graphs, DNA strings, and so on) since for each kind of data the most suitable (dis)similarity measure can be chosen. In the following, we introduce several dissimilarity measures and define the dissimilarity space.

10.5.2.1 Dissimilarity Measures

Up to this level of the pipeline, data are characterized by histograms. Therefore, we can use histograms to devise similarity measures to be employed in the dissimilarity-based representation scheme. There are various dissimilarity measures that can be applied to measure the dissimilarities between histograms [18, 68]. Moreover, histograms can be converted to pdfs and dissimilarity measures between two discrete distributions can be used as well. All in all, we decided to study measures below.

Given two histograms S and M with n bins, we define the number of elements in S and M as $|S|$ and $|M|$, respectively.

Histogram Intersection It measures the number of intersecting values in each bin [74]:

$$\mathrm{sim}(S, M) = \frac{\sum_{i=1}^{n} \min(S_i, M_i)}{\min(|S|, |M|)}.$$

Since this is a similarity measure, we convert it to a dissimilarity using $D = \min(|M|, |S|) \times (1 - \mathrm{sim}(S, M))$.

Diffusion Distance In diffusion distance [50], the distance between two histograms is defined as a temperature field $T(x, t)$ with $T(x, 0) = S(x) - M(x)$. Using the heat diffusion equation $\frac{\partial T}{\partial t} = \frac{\partial^2 T}{\partial x^2}$ which has a unique solution $T(x, t) = T(x, 0) \times \phi(x, t)$ where $\phi(x, t) = \frac{1}{(2\phi)^{1/2} t} \exp\{-\frac{x^2}{2t^2}\}$, we can compute D as:

$$D = \int_0^r k\big(\big|T(x, t)\big|\big)\, dt,$$

where $k(\cdot)$ is the L_1-norm.

χ^2 Distance This metric is based on the χ^2 test for testing the similarity between histograms. It is defined as

$$D = \sum_{i=1}^{n} \frac{(S_i - M_i)^2}{S_i + M_i}.$$

It is a standard measure for histograms.

Earth Mover's Distance This distance was originally proposed by Rubner et al. [65]. It is basically defined as the cost to transform one distribution into another. It is calculated using linear optimization by defining the problem as a transportation problem. For 1D histograms, it reduces to a simple calculation [18] which was implemented in this study.

$$C_i = \left| \sum_{j=1}^{i} (S_j - M_j) \right|, \qquad D = \sum_{i=1}^{n} C_i.$$

Similarly, we have considered the following dissimilarities between pdfs:

Bhattacharyya It is used to measure the similarity of discrete probability distributions p and q. It is defined as

$$D(p, q) = -\log \mathrm{BC}(p, q),$$

where

$$\mathrm{BC}(p, q) = \sum_{x \in X} \sqrt{p(x)q(x)}.$$

Kullback–Leibler (KL) Divergence Kullback–Leibler divergence is defined as

$$D(p, q) = \sum_{i=1}^{n} q_i \log \frac{q_i}{p_i}.$$

This measure is not a distance metric but a relative entropy since $D(p, q) \neq D(q, p)$, i.e., the dissimilarity matrix is not symmetric. There are various ways to symmetrize this dissimilarity. We simply used $D = D(p, q) + D(q, p)$ and the so-called Jensen–Shannon divergence: $D = \frac{1}{2}D(p, r) + \frac{1}{2}D(q, r)$, where r is the average of p and q.

10.5.2.2 Dissimilarity Space

Suppose that we have n objects and we have a dissimilarity matrix D of size $n \times n$. And suppose that the dissimilarity between two objects o and $\hat{o}$ are denoted by $D(o, \hat{o})$. There are several ways to transform an $n \times n$ dissimilarity matrix D with elements $D(o, \hat{o})$ into a vector space with objects represented by vectors $X = \{x'_1, \ldots, x'_o, \ldots, x'_{\hat{o}}, \ldots, x'_n\}$ [55]. Classical scaling (for proper Euclidean dissimilarities) and pseudo-Euclidean embedding (for arbitrary symmetric dissimilarities) yield vector spaces in which vector dissimilarities can be defined that produce the given dissimilarities D. As almost all dissimilarity measures studied here are non-Euclidean, classification procedures for these pseudo-Euclidean spaces are ill-defined, as, for instance, the corresponding kernels are indefinite.

A more general solution is to work directly in the *dissimilarity space*. It postulates an Euclidean vector space using the given dissimilarities to a representation set as features. As opposed to the previously mentioned techniques, it is not true anymore that dissimilarities in this space are identical to the given dissimilarities, but this is an advantage in case it is doubtful whether they really represent dissimilarities between the physical objects. As this holds in our case we constructed such a dissimilarity space using all available objects by taking X equal to D. In the dissimilarity space, basically any traditional classifier can be used. The number of dimensions, however, equals the number of objects in the representation set. Many classifiers will need dimension reduction techniques or regularization to work properly in this space.

A further refining of the scheme can be obtained by considering at the same time different dissimilarities (we have many, linked to different modalities, different zones of the brain or different methods to compute them), trying to combine them in a single dissimilarity space. Combined dissimilarity spaces can be constructed by combining dissimilarity representations. As in normal classifier combination [42, 45], a simple and often effective way is using an (weighted) average of the given dissimilarity measures:

$$D_{\text{combined}} = \frac{\sum \alpha_i D_i}{\sum \alpha_i}. \tag{10.6}$$

It is related to the sum-rule in the area of combining classifiers. The weights can be optimized for some overall performance criterion, or determined from the properties of the dissimilarity matrix D_i itself, e.g., its maximum or average dissimilarity. Here, we used equal weights while combining multiple dissimilarity matrices and all the dissimilarity matrices are scaled such that the average dissimilarity is one, i.e.,

$$\frac{D(o, \hat{o})}{\frac{1}{n(n-1)} \sum_{o,\hat{o}} D(o, \hat{o})} = 1. \tag{10.7}$$

This is done to assure that the results are comparable over the dissimilarities as we deal with dissimilarity data in various ranges and scales. Such scaled dissimilarities

are denoted as $\tilde{D}$. In addition, we assume here that the dissimilarities are symmetric. So, every dissimilarity $\tilde{D}(i, j)$ has been transformed by

$$\tilde{D}(i, j) := \frac{\tilde{D}(i, j) + \tilde{D}(j, i)}{2}. \tag{10.8}$$

10.5.3 Descriptors by Generative Embeddings

In this section, we define a new class of data descriptors based on generative embedding procedure (see Chap. 4). The overall idea consists in fitting a generative model on training data and using the generative process to define new data-dependent representations. Then, such representations can be plugged into a standard discriminative classifier for classification purposes. This approach is pursued by hybrid architectures of discriminative and generative classifiers which is currently one of the most interesting, useful, and difficult challenges for Machine Learning. The underlying motivation is the proved complementariness of discriminative and generative estimations: asymptotically (in the number of labeled training examples), classification error of discriminative methods is lower than that of generative ones [53]. On the other side, generative counterparts are effective with less, possibly unlabeled, data; further, they provide intuitive mappings among structure of the model and data features. Among these hybrid generative–discriminative methods, "generative embeddings" (also called generative score space) have grown in importance in the literature [11, 12, 39, 49, 56, 71, 72, 79].

Generative score space framework consists of two steps: first, one or a set of generative models are learned from the data; then a score (namely a vector of features) is extracted from it, to be used in a discriminative scenario. The idea is to extract fixed dimension feature vectors from observations by subsuming the process of data generation, projecting them in highly informative spaces called score spaces. In this way, standard discriminative classifiers such as support vector machines, or logistic regressors have achieved higher performances than a solely generative or discriminative approach.

Using the notation of [56, 71], such spaces can be built from data by mapping each observation x to the fixed-length score vector $\varphi^{f}_{\hat{F}}(x)$,

$$\varphi^{f}_{\hat{F}}(x) = \varphi_{\hat{F}} f\left(\left\{P_i(x \mid \theta_i)\right\}\right), \tag{10.9}$$

where $P_i(x \mid \theta_i)$ represents the family of generative models learned from the data, f is the function of the set of probability densities under the different models, and $\hat{F}$ is some operator applied to it. In general, the generative score-space approaches help to distill the relationship between a model's parameters θ and the particular data sample.

Generative score-space approaches are strictly linked to generative kernels family, namely kernels which compute similarity between points through a generative

model—the most famous example being the Fisher Kernel [39]. Typically, a generative kernel is obtained by defining a similarity measure in the score space, e.g., the inner product. Score spaces are also called model dependent feature extractors, since they extract features from a generative model.

In order to apply the generative embedding scheme to the MRI data, we should define a generative model able to explain and model what we have. Here, we adapted as generative model the probabilistic Latent Semantic Analysis (pLSA—[38]), a tool widely applied in the linguistic and in the computer vision community.

In the following, we first describe the basics of the pLSA, then explain how this model can be applied to our problem, finally describing the kind of generative embeddings we exploited.

10.5.3.1 Probabilistic Latent Semantic Analysis

In the Probabilistic Latent Semantic Analysis (PLSA—[38]), the input is a set of D documents, each one containing a set of words taken from a vocabulary of cardinality W. The documents are summarized by an occurrence matrix of size $W \times D$, where $n(w_j, d_i)$ indicates the number of occurrences of the word w_j in the document d_i. In PLSA, the presence of a word w_j in the document d_i is mediated by a latent *topic* variable, $z \in Z = \{z_1, \ldots, z_Z\}$, also called *aspect* class, i.e.,

$$P(w_j, d_i) = \sum_{k=1}^{Z} P(z_k) P(w_j \mid z_k) P(d_i \mid z_k). \tag{10.10}$$

In practice, the topic z_k is a probabilistic co-occurrence of words encoded by the distribution $\beta(w) = p(w \mid z_k)$, $w = \{w_1, \ldots, w_N\}$, and each document d_i is compactly (usually, $Z < W$) modeled as a probability distribution over the topics, i.e., $p(z \mid d_i)$, $z = \{z_1, \ldots, z_Z\}$ (note that this formulation, derived from $p(d_i \mid z)$, provides an immediate interpretation).

The hidden distributions of the model, $p(w \mid z)$, $p(d \mid z)$ and $p(z)$, are learned using Expectation Maximization (EM) [26], maximizing the model data log-likelihood $\mathcal{L}$:

$$\mathcal{L} = \prod_{j=1}^{W} \prod_{i=1}^{D} n(w_j, d_i) \log\big(p(w_j, d_i)\big). \tag{10.11}$$

The E-step computes the posterior over the topics, $p(z \mid w, d)$, and the M-step updates the hidden distributions. Even if pLSA is a model for documents, it has been largely applied in other contexts, especially in computer vision [12, 23] but also in the medical informatics domain [9, 10, 17].

The idea under its application to the MRI domain is straightforward. In particular, we can assume that a given brain (or the particular ROI) represents the documents d, whereas the words w_j are the local features previously described. With such a point of view, the extracted histograms represent the counting vectors, able to describe

how much a visual feature (namely a word) is present in a given image (namely a document).

10.5.3.2 PLSA-Based Generative Embeddings

Once a generative model is trained, different spaces can be obtained. Generally speaking, we can divide them into two families: parameter-based and hidden variable-based. The former class derives the features on the basis of differential operations linked to the parameters of the probabilistic model, while the latter seeks to derive feature maps on the basis of the log-likelihood function of a model, focusing on the random variables rather than on the parameters.

Parameter-Based Score Space These methods derive the features on the basis of differential operations linked to the parameters of the probabilistic model.

The Fisher Score The Fisher score for the PLSA model has been introduced in [37], starting from the asymmetric formulation of PLSA. In this case, the log-probability of a document d_i is defined by

$$\begin{aligned} l(d_i) &= \frac{\log P(d_i, w)}{\sum_m n(d_i, w_m)} \\ &= \sum_{j=1}^{W} \hat{P}(w_j \mid d_i) \log \sum_{k=1}^{Z} P(w_j \mid z_k) P(d_i \mid z_k) P(z_k), \end{aligned} \tag{10.12}$$

where $\hat{P}(w_j \mid d_i) \equiv n(d_i, w_j) / \sum_m n(d_i, w_m)$ and where $l(d_i)$ represents the probability of all the word occurrences in d_i normalized by document length.

Differentiating Eq. (10.12) with respect to $P(z)$ and $P(w \mid z)$, the pLSA model parameters, we can compute the score. The samples are mapped in a space of dimension $W \times Z + Z$. The Fisher kernel is defined as the inner product in this space. We will refer to it as FSH.

TOP Kernel Scores Top Kernel and the tangent vector of posterior log-odds score space were introduced in [79]. Whereas the Fisher score is calculated from the marginal log-likelihood, TOP kernel is derived from Tangent vectors Of Posterior log-odds. Therefore, the two score spaces have the same score function (i.e., the gradient) but different score arguments, which, for TOP kernel $f(p(x \mid \theta)) = \log P(c = +1 \mid x, \theta) - \log P(c = -1 \mid x, \theta)$ where c is the class label. We will refer to it as TOP.

Log-Likelihood Ratio Score Space The log-likelihood ratio score space is introduced in [72]. Its dimensions are similar as for the Fisher score, except that the procedure is repeated for each class: a model θ_c per class is learned and the gradient is applied to each $\log p(x \mid \theta_c)$. The dimensionality of the resulting space is $C \times$ the dimensionality of the original Fisher score. We will refer to it as LLR.

Random Variable Based Methods These methods, starting from considerations in [56], seek to derive feature maps on the basis of the log-likelihood function of a model, focusing on the random variables rather than on the parameters in their derivation (as done in the parameter-based score spaces).

Free Energy Score Space (FESS) In the Free Energy Score Space [56], the score function is the free energy while the score argument is its unique decomposition into the terms that compose it.[6] Free energy is a popular score function representing a lower bound of the negative log-likelihood of the visible variables used in the variational learning. For pLSA it is defined by the following equation:

$$\begin{aligned}\mathcal{F}(d_i) = &\sum_w n(d_i, w) \sum_z P(z \mid d, w) \log P(z \mid d, w) \\ &- \sum_w n(d_i, w) \sum_z P(z \mid d, w) \log P(d, w \mid z) P(z), \end{aligned} \tag{10.13}$$

where the first term represents the entropy of the posterior distribution and the second term is the cross-entropy. For further details on the free energy and on variational learning, see [31]; for the pLSA's free energy, see [38].

For pLSA this results in a space of dimension equal to $C \times 2 \times Z \times W$. In [56], the authors point out that, if the dimensionality is too high, some of the sums can be carried out to reduce the dimensionality of the score vector before learning the weights. The choice of the term to optimize is intuitive but guided by the particular application. In our case, as previously done in [49, 57], we perform the sums over the word indices, optimizing the contributing topics. The resulting score space has dimension equal to $C \times 2 \times Z$; we will refer to this score space as FESS.

Posterior Divergence Posterior Divergence score space is described in [49]. Like FESS, it takes into account how well a sample fits the model (cross-entropy terms in FESS) and how uncertain the fitting is (entropy terms in FESS, Eq. (10.13)) but it also assesses the change in model parameters brought by the input sample, i.e., how much a sample affects the model. These three measures are not simply stacked together, but they are derived from the incremental EM algorithm which, in the E-step only, looks at one or a few selected samples to update the model at each iteration. Details on posterior divergence score vector for pLSA and on its relationships with FESS case can be found in [49]. We will refer to this score space as PD.

Classifying with the Mixture of Topics of a Document Very recently, pLSA has been used as a dimensionality reduction method in several fields, like computer vision, bioinformatics, and medicine [10, 12, 17]. The idea is to learn a pLSA model to capture the co-occurrence between visual words [12, 17], or gene expressions [10], which represent the (usually) high-dimensional data description; co-occurrences are

[6] This is true once a family for the posterior distribution is given. See the original paper for details.

captured by the topics. Subsequently, the classification is performed using the topic distribution that defines a document as sample descriptor.

Since we are extracting features from a generative model, we are defining a score space which is the Z-dimensional simplex. In this case, the score argument f, a function of the generative model, is the topic distribution $P(z \mid d)$ (using Bayes' formula, one can easily derive $P(z \mid d)$ starting from $P(d \mid z)$), while the score function is the identity. We will refer to this score space as TPM.

In our experiments, for the two score spaces FESS and TPM, we include two versions. The first version is where we train one pLSA per class and concatenate the resulting feature vectors (we will refer these as FESS-1 and TPM-1), the second one is where we train a pLSA for the whole data without looking at the class label (we will refer these as FESS-2 and TPM-2). All in all, we have eight different score spaces: TPM-1, TPM-2, FESS-1, FESS-2, LLR, FSH, TOP, PD.

10.6 Classification

After data description step, a learning-by-example procedure is employed for brain classification in order to discriminate between healthy subjects and patients affected by schizophrenia. As a basic approach, when a single source of information is considered, a standard single classifier can be employed. From the medical point of view, this means that the relevance of a particular source of information is considered to characterize the brain abnormality. On the other hand, when several factors can be the possible cause of the disease, a multi-source classification strategy may be employed. Here, we have exploited two paradigms: (i) multi-classification, and (ii) Multiple Kernel Learning (MKL).

10.6.1 Multi-classifier

It is a well-known fact that there is no single most accurate classification algorithm, so methods have been proposed to combine classifiers based on different assumptions [1, 45]. Classifier combination (also called ensemble construction) can be done at different levels and in different ways: (i) sensor fusion, (ii) representation fusion, (iii) algorithm fusion, (iv) decision fusion, and others. Each classifier (⟨algorithm/parameter set/data representation⟩ triplet) makes a different assumption about the data and makes errors on different instances and by combining multiple classifiers; the overall error can be decreased. Classifiers' making different errors on different parts of the space is called "diversity" (in a broad definition), and to achieve diversity different (i) learning algorithms, (ii) hyperparameters, (iii) input features, and (iv) training sets have been used [45, 83] .

There are various methods to combine classifiers; the simple method is to use voting [42] (or take an average over the outputs) which corresponds to fixed rules

which we applied when the classifiers created posterior probability outputs, i.e., $P(C_k \mid \mathbf{x}, E) = \sum_{i=1}^{L} P(C_k \mid \mathbf{x}, M_i)$, where E denotes the ensemble, $P(C_k \mid \mathbf{x}, E)$ is the posterior of the ensemble for class C_k, L is the number of classifiers to combine, $M_i, i = 1, \ldots, L$ are the individual classifiers to combine, and $P(C_k \mid \mathbf{x}, M_i)$ is the posterior of classifier M_i. Voting does not require any parameter to be optimized and is simple. Other methods such as weighted averaging or more advanced methods require the estimation of other parameters. In previous works [20, 80], we used single classifier and multi-classifier approaches to schizophrenia detection with correlation analysis which serve as a baseline for our dissimilarity based analysis.

10.6.2 Multiple Kernel Learning (MKL)

The main idea behind SVMs [84] is to transform the input feature space to another space (possibly with a greater dimension) where the classes are linearly separable. After training, the discriminant function of SVM becomes $f(\boldsymbol{x}) = \langle \boldsymbol{w}, \Phi(\boldsymbol{x}) \rangle + b$, where $\boldsymbol{w}$ is the vector of weights, b is the threshold, and $\Phi(\cdot)$ is the mapping function. Using the dual formulation and the kernel trick, one does not have to define this mapping function explicitly and the discriminant function can be written as

$$f(\boldsymbol{x}) = \sum_{i=1}^{N} \alpha_i y_i k(\boldsymbol{x}_i, \boldsymbol{x}) + b,$$

where $k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \langle \Phi(\boldsymbol{x}_i), \Phi(\boldsymbol{x}_j) \rangle$ is the kernel function that calculates a similarity metric between data instances. Selecting the kernel function is the most important issue in the training phase; it is generally handled by choosing the best-performing kernel function among a set of kernel functions on a separate validation set.

In recent years, MKL methods have been proposed [6, 46] (for a review see [35]), for learning a combination k_η of multiple kernels instead of selecting only one:

$$k_\eta(\boldsymbol{x}_i, \boldsymbol{x}_j; \boldsymbol{\eta}) = f_\eta\left(\left\{k_m\left(\boldsymbol{x}_i^m, \boldsymbol{x}_j^m\right)_{m=1}^{P}\right\}; \boldsymbol{\eta}\right), \tag{10.14}$$

where the combination function f_η forms a single kernel from P base kernels using the parameters $\boldsymbol{\eta}$. Different kernels correspond to different notions of similarity and instead of searching which works best, the MKL method does the picking for us, or may use a combination of kernels. MKL also allows us to combine different representations possibly coming from different sources or modalities.

There is significant work on the theory and application of MKL, and most of the proposed algorithms use a linear combination function such as convex sum or conic sum. Fixed rules use the combination function in (10.14) as a fixed function of the kernels, without any training. Once we calculate the combined kernel, we train a single kernel machine using this kernel. For example, we can obtain a valid kernel by taking the mean of the combined kernels.

Instead of using a fixed combination function, we can also have a function parameterized by a set of parameters and then we have a learning procedure to optimize these parameters as well. The simplest case is to parameterize the sum rule as a weighted sum:

$$k_\eta(\boldsymbol{x}_i, \boldsymbol{x}_j; \boldsymbol{\eta}) = \sum_{m=1}^{P} \eta_m k_m\left(\boldsymbol{x}_i^m, \boldsymbol{x}_j^m\right)$$

with $\eta_m \in \mathbb{R}$. Different versions of this approach differ in the way they put restrictions on the kernel weights [6, 46, 60]. For example, we can use arbitrary weights (i.e., linear combination), nonnegative kernel weights (i.e., conic combination), or weights on a simplex (i.e., convex combination). A linear combination may be restrictive, and nonlinear combinations are also possible [22, 34, 36, 48].

10.7 Case Study 1: Brain Classification on Dissimilarity Space

After presenting all the possible choices we made in the different parts of the pipeline, let us present some concrete systems. In particular, here we describe a method based on the dissimilarity-representation paradigm, whereas in Sect. 10.8 a method based on the generative embeddings is presented.

Concerning the taxonomies presented in Fig. 10.1, here we are using both sMRI and DWI approaches (namely a multimodal scheme), starting from the brain parcellation, employing the dissimilarity-based description and dissimilarity combination by classifying with a single classifier.

In particular, in these experiments [82], we use a 114 subject subset of the original data set (59 patients, 55 healthy controls). We used the intensity histograms from sMRI images (SMRI), ADC histograms from DWI images (ADC), and two geometric shape descriptors, shape index (SH) and mean curvature (MCUR). We used all the ROIs and used the dissimilarity space by computing the dissimilarities between the histograms and their corresponding pdfs. In summary, for each ROI and representation we use the following 13 measures:

- *hist-euclid*—Euclidean distance between histograms
- *hist-l1*—L_1-distance between histograms
- *hist-intersect*—Intersection between histograms
- *hist-diffusion*—Diffusion distance between histograms
- *hist-chi*—χ^2-distance between histograms
- *hist-emd*—Earth mover's distance between histograms
- *pdf-euclid*—Euclidean distance between pdfs
- *pdf-l1*—L_1-distance between pdfs
- *pdf-emd*—Earth mover's distance between pdfs
- *pdf-bs*—Bhattacharyya distance between pdfs
- *pdf-kl*—Symmetrized KL divergence between pdfs
- *pdf-kl-orig*—Original, asymmetric KL divergence

- *pdf-js*—Jensen–Shannon divergence between pdfs

All in all, there are 14 ROIs and 13 different dissimilarity measures per modality, which yields a total of 182×4 dissimilarity matrices. In addition to these, we propose to merge the different dissimilarity matrices into one overall dissimilarity matrix per modality, potentially exploiting complementary information useful to improve the classification accuracy. We also test the accuracy of these combinations against combining classifiers in the original feature space (histograms and pdfs for each of the four modalities). For each test we evaluated the leave-one-out error. All differences in accuracy reported in this case study are significant at $p = 0.05$ using the paired t-test. The dissimilarity spaces have been built in a transductive way by using all available subjects for dissimilarity (of course, labels are ignored in this phase). Three classifiers are considered to compare the performances:

- Linear SVM classifier on the original feature space (called *svm*)
- The 1-Nearest Neighbor (NN) rule on the dissimilarity matrices (called *1nn*)
- Linear SVM classifier on the dissimilarity space (called *sv0*)

The linear SVM in dissimilarity space avoids complications that could arise from the dissimilarity measures being non-Euclidean because we treat the dissimilarities as features in this new space. While combining dissimilarities, we use for α_i in (Eq. (10.6)) the reciprocal of the number of dissimilarity matrices to be combined [47]. On the original feature space, the SVM classifiers produce posterior probability outputs, and these outputs are combined using the SUM rule [42]. So, on the original feature space, we combine after training the classifiers, whereas on the dissimilarity space, we combine before we do classification. The experiments are carried out using the Matlab package PRTools [28]. We designed three experiments to show the improvements of dissimilarity-based pattern recognition techniques and combination of dissimilarities using multiple ROIs and modalities:

1. ROI-based classification—For each modality, we report the highest accuracy that a classifier reaches without combination (on the original feature space and on the dissimilarity space). We use these results as a baseline for comparison.
2. Multi-ROI classification—In this set of experiments, for each modality, we fix the dissimilarity measure and combine all ROIs using this dissimilarity measure.
3. Multimodal classification—In this experiment, we go one step further to combine information coming from different sources by combining different dissimilarity matrices from different modalities.

We note that, throughout this section, we will use the following notation: every dissimilarity representation will be referred to as MODALITY-*roi-dissimilarity* (i.e., SMRI-*ldlpfc-pdf-js* shows the dissimilarity matrix for the structural MRI of ROI *ldlpfc* using the dissimilarity measure of Jensen–Shannon divergence). The modality, ROI, or the dissimilarity measure will be omitted when it's clear from the context.

Table 10.2 Best accuracies for each dissimilarity measure combining all ROIs on all modalities

Modality	SMRI			ADC			SH			MCUR		
ROI	*svm*	*1nn*	*sv0*	*svm*	*1nn*	*sv0*	*svm*	*1nn*	*sv0*	*svm*	*1nn*	*sv0*
lamyg	68.42	64.04	**78.07**	64.04	57.89	62.28	45.61	**68.42**	64.91	43.86	61.40	57.02
ramyg	54.39	65.79	66.67	54.39	56.14	59.65	49.12	53.51	57.89	46.49	58.77	57.89
ldlpfc	60.53	62.28	76.32	56.14	51.75	61.40	52.63	62.28	57.89	49.12	53.51	53.51
rdlpfc	64.04	57.89	68.42	54.39	56.14	65.79	54.39	59.65	60.53	56.14	57.02	60.53
lec	64.04	56.14	64.91	53.51	62.28	61.40	46.49	51.75	54.39	47.37	53.51	53.51
rec	64.91	65.79	71.05	64.04	58.77	60.53	52.63	60.53	57.02	52.63	65.79	63.16
lhg	51.75	60.53	63.16	55.26	61.40	54.39	47.37	54.39	65.79	61.40	56.14	62.28
rhg	50.00	63.16	59.65	50.88	58.77	58.77	43.86	55.26	55.26	57.89	64.04	61.40
lhippo	63.16	60.53	72.81	52.63	57.02	59.65	55.26	50.00	57.02	53.51	58.77	62.28
rhippo	60.53	64.04	66.67	48.25	55.26	52.63	47.37	59.65	57.02	54.39	55.26	58.77
lstg	55.26	59.65	69.30	54.39	56.14	60.53	40.35	58.77	52.63	50.00	50.00	51.75
rstg	63.16	57.02	64.91	64.04	60.53	**70.18**	53.51	55.26	57.89	41.23	63.16	57.89
lthal	58.77	64.91	67.54	57.89	57.89	58.77	46.49	53.51	57.89	53.51	56.14	58.77
rthal	64.91	59.65	64.04	53.51	60.53	59.65	54.39	57.89	59.65	48.25	54.39	**67.54**

10.7.1 ROI-Based Classification

We evaluate the classification accuracies for each of the original dissimilarity matrices. Table 10.2 summarizes the results for all modalities. For each ROI the best performance is reported with respect to various dissimilarity measures (for details see [82]). The first column for each modality reports the accuracy estimates for *svm* using the original feature space (histograms and pdfs). The second column reports the maximum accuracy of *1nn* on different dissimilarity measures. The third column reports the leave-one-out accuracy estimates of the linear SVM in dissimilarity space. For SMRI, it shows clearly the improvements of our dissimilarity-based approach. Except for two ROIs (*rhg* and *rthal*), SVM classifier in the dissimilarity space is always better than classifiers in the standard space. While the best accuracy of standard approaches is 68.42 %, we can reach 78.07 % accuracy on dissimilarity space and the dissimilarity space accuracies are more stable.

For the other modalities, the results are similar. From the results for the ADC values extracted from DWI images, we can again see that when we switch to dissimilarity based classification, we get better accuracies (either *1nn* or *sv0*) except for two ROIs (*lamyg* and *rec*). We can again see that with a single ROI and dissimilarity measure, we can reach 70.18 % whereas the highest accuracy we can obtain in the original space is 64.04 %. The same pattern can be observed when we investigate SH and MCUR. Also in these modalities, the best accuracy can be achieved using dissimilarities. We can see that on SH, we reach 68.42 % using *1nn* and 65.79 % using *sv0*. The best accuracy using the features on the original space is 55.26 %. Also on MCUR, best accuracy is reached using *sv0*.

Table 10.3 Best accuracies for each dissimilarity measure combining all ROIs on all modalities

Measure	SMRI		ADC		SH		MCUR	
	1nn	*sv0*	*1nn*	*sv0*	*1nn*	*sv0*	*1nn*	*sv0*
hist-l2	62.28	71.05	50.00	60.53	57.89	57.89	49.12	50.88
hist-l1	62.28	74.56	46.49	**64.91**	58.77	**60.53**	50.00	51.75
hist-intersect	66.67	68.42	43.86	61.40	40.35	53.51	53.51	50.88
hist-diffusion	62.28	74.56	46.49	**64.91**	58.77	**60.53**	50.00	51.75
hist-chi	57.02	71.05	50.88	55.26	59.65	57.02	**55.26**	48.25
hist-emd	52.63	58.77	58.77	51.75	55.26	56.14	43.86	53.51
pdf-l2	57.02	74.56	57.02	60.53	50.88	55.26	57.02	51.75
pdf-l1	60.53	**76.32**	54.39	61.40	50.88	58.77	54.39	46.49
pdf-emd	59.65	75.44	57.89	53.51	**60.53**	**60.53**	50.00	52.63
pdf-bc	65.79	69.30	53.51	53.51	48.25	57.89	44.74	52.63
pdf-kl	66.67	70.18	55.26	48.25	52.63	59.65	48.25	49.12
pdf-kl-orig	64.04	64.91	49.12	51.75	57.02	59.65	**55.26**	46.49
pdf-js	65.79	71.93	52.63	54.39	48.25	**60.53**	53.51	48.25
average	60.53	**76.32**	51.75	60.53	54.39	**60.53**	54.39	49.12
svm	71.93		63.16		51.75		47.37	

10.7.2 Multi-ROI Classification

In this section, we will show our experiments where we combine multiple ROIs, fixing the modality and distance measure. We also conducted experiments by fixing the ROIs and combining multiple dissimilarity matrices using the same ROI. We see that the accuracy does not increase as compared to combining ROIs with fixed dissimilarity measure. This conforms to our previous studies, therefore here, we do not report on combination of distance measures with fixed ROI.

In this experiment, a multi-ROI approach is adapted to use all ROIs at the same time. All the dissimilarity matrices for each ROI are combined by averaging the normalized dissimilarity matrices. The second and third columns of Table 10.3 report the results on intensity histograms, using 1-NN rule on the dissimilarity matrices and the support vector classifiers in the dissimilarity spaces. Also in this case, the classification on the dissimilarity space clearly outperforms the standard approach. Moreover, the multi-ROI approach brings an improvement by confirming the complementary information enclosed onto the different brain subparts when we use *sv0* on the dissimilarity space. In most of the cases, the results from the averaged similarity matrices are better than the respective best single ROI results. The row average in Table 10.3 reports the error estimates computed on the overall dissimilarity matrix (combining all the measures and ROIs), which has the highest accuracy 76.32 % (same as combining all ROIs for *pdf-l1*) for both the standard approach and dissimilarity-based approach, respectively. The last row reports the accuracy of

combining all SVM classifiers in the original feature space. When we combine all the SVM classifiers on the original space, we get 71.93 % accuracy. This shows us that the dissimilarity space produces better results also when we consider classifier combination. We repeated the same experiments also with the other modalities. In Table 10.3, we can also see the results using the other modalities. We observe that again we get the most accurate results when we combine ROIs in the dissimilarity space using *sv0* except for mean curvature histograms where the best results are obtained using *1nn* (using dissimilarities again).

10.7.3 Combining Different Modalities

As a further step to understand how information from multiple sources can be combined to reach better classification accuracy, we develop another experiment where we combine information from multiple modalities. We have 182 dissimilarity matrices from each of the four modalities. It is not possible to exhaustively search the whole solution space to find the best solution (optimum subset for combination), so instead, we choose the most accurate four ROI-dissimilarity pairs from each modality and do an exhaustive search on the combination of these matrices to get the best result. We can see the selected dissimilarity matrices and their base accuracies in Table 10.4. With a total of 16 dissimilarity matrices (modality-ROI-dissimilarity triples), we can get the best accuracy 86.84 % (last row in Table 10.4), which contains two dissimilarity matrices from intensities (*ldlpfc-pdf-kl-orig* and *ldlpfc-pdf-bc* both having 75.44 % accuracy) and one dissimilarity matrix from shape index (*rdlpfc-hist-chi* with 60.53 % accuracy). This accuracy is the best accuracy, which has been reached using dissimilarity combination and cannot be reached using only one modality. Applying the same methodology, we can reach only 76.32 % accuracy with *1nn* and 83.33 % accuracy with svm on the original feature space. This also shows us why it is important to combine useful information from different sources to come up with better accuracy. We see that the accuracy can be increased when complementary information using different modalities are combined.

In a medical application, besides increasing accuracy, the interpretability of the results is also important. We use this experiment to deduce information on the use of ROIs, their complementary information, and how each modality relates to the detection of schizophrenia. For this purpose, we select all the combinations of distance matrices with accuracies above 82 % (we have 69 different combinations) and count the occurrences of dissimilarity matrices for every combination. From Table 10.4, we can see that most of the combinations include *ldlpfc* of SMRI and the shape index of *rthal*. This shows us that these two modality-ROI pairs contribute and complement other dissimilarity matrices, and by using these two in combination, we increase accuracy. After these two dissimilarity matrices, we see that mean curvature of *rthal* and shape index of *rdlpfc* are used in combination the most. These are followed by *ldlpfc* of histogram intensities and the mean curvature of *rec*. With ADC, we see that most used ROI is *rstg*, which has been selected 38 times. This

Table 10.4 Most accurate four dissimilarity matrices from each modality, their single performances, and number of occurrences in the combination of most accurate results

Selected dissimilarity	Accuracy	Occurrences
SMRI-ldlpfc-pdf-js	76.32	60
SH-rthal-hist-l1	59.65	57
MCUR-rthal-pdf-bc	67.54	52
SH-rdlpfc-hist-chi*	60.53	50
SMRI-ldlpfc-pdf-bc*	75.44	48
SMRI-ldlpfc-pdf-kl-orig*	75.44	48
MCUR-rec-pdf-l1	63.16	47
SMRI-lamyg-pdf-bc	78.07	42
ADC-rstg-hist-l2	65.79	38
SH-lamyg-hist-emd	64.91	38
ADC-rstg-pdf-bc	70.18	20
MCUR-rec-pdf-l2	63.16	17
ADC-rdlpfc-pdf-emd	65.79	14
ADC-rstg-pdf-js	65.79	9
SH-lhg-hist-intersect	65.79	8
MCUR-lhippo-pdf-emd	62.28	1
Dissimilarities with * in the optimal combination are	86.84	

also shows us that the DWI information is the least complementary modality in this scenario, and one can design experiments without this modality, focusing on the other modalities. We can use this information to decrease the costs of the operation, that is, not performing DWI analysis. Also we see that the most accurate dissimilarity matrix (SMRI-*lamyg-pdf-bc*) is the eighth most used dissimilarity when we consider combination. This interesting fact shows us that when doing combination, the complementary information is more important than individual accuracies.

Another interesting fact is that some ROIs are more discriminative when the structural information is considered, and some are more discriminative when we consider DWI. The ROIs selected from the structural analysis in this experiment are those considered crucial for the impaired neural network in schizophrenia and comply with current studies in the literature [21], in contrast DWI is particularly keen in exploring the microstructural organization of white matter, therefore providing intriguing information on brain connectivity [13], but does not have complementary contribution in this context.

With this analysis, we can open a new perspective of how to use each of these modalities to get better accuracies. One can use this information to setup new experiments considering the contributions of these ROIs on these modalities.

10.7.4 Discussion

In this case study, a novel approach based on dissimilarity-based pattern recognition is proposed for the detection of schizophrenic brains. Several dissimilarity measures are proposed to deal with histograms of different types for different ROIs. ROI-based classification on the dissimilarity space shows improvements of the standard NN rule and the support vector classifier on the original space. Moreover, a Multi-ROI classification strategy is obtained by simply averaging the similarity matrices observed in each ROI. Such an approach improves the single-ROI one, by highlighting the complementary information enclosed in the several ROIs. This confirms the benefit of combining dissimilarity information and fusing information from various regions in the brain.

We investigate further to combine information from multiple modalities such as intensities, ADC values and geometric information. We can see that some ROIs are discriminative when we use intensities; some are useful when DWI data is considered. Geometric properties of some ROIs play a part in schizophrenia detection. We show that we get the best accuracy when we combine multiple modalities.

We can interpret the results of combining multiple modalities to set up further experiments in this context. Our results show that the least contributing modality is the DWI. With this information, one can skip using this modality and focus more on histograms of intensities and geometric information. Also, one can use this result to reduce the costs of this operation, by not performing DWI measurements and without the patient undergoing further medical operations.

We would like to emphasize that in building the (combined) dissimilarities no parameters are optimized w.r.t. performance. The proposed approach of combining dissimilarities on the dissimilarity space opens new perspectives in neuroanatomy classification by allowing the possibility to exploit dissimilarity measures where one does not have to deal with technical difficulties such as the metric requirements of distance based classification and kernel restrictions of support vector machines.

10.8 Case Study 2: Brain Classification by Generative Embeddings

In this case study, we use *Heat Kernel Signatures* to extract histogram-based features from SMRI and use the generative embedding score spaces mentioned in Sect. 10.5.3 and apply IT kernels [52]. We used average hold out methodology with 30 repetitions using stratification. For estimating the C value of the SVM and q value for the IT kernels, we used 5-fold cross-validation on the training set. To estimate the number of topics, we used the Bayesian Information Criterion (BIC) [67], which penalizes the likelihood with a penalty term on the number of free parameters in a way that larger models which do not increase the likelihood significantly are discouraged. In the pLSA model, the number of free parameters is calculated as

$(D-1)Z+(W-1)Z+(Z-1)$. Then the BIC becomes

$$\mathrm{BIC}=\frac{1}{2}\big((D-1)Z+(W-1)Z+(Z-1)\big)\log\sum_{j=1}^{W}\sum_{i=1}^{D}n(d_i,w_j).$$

10.8.1 Proposed Approach

Kernels on probability measures have been shown to be very effective in classification problems involving text, images, and other types of data [24, 40]. Given two probability measures p_1 and p_2, representing two objects, several information theoretic kernels (ITKs) can be defined [52]. In this work, we use the Jensen–Shannon kernel (JS), Jensen–Tsallis kernel (JT), and weighted JT kernel (since results were similar, we omit the weighted JT kernel (version B) [52]—we will refer to weighted JT kernel (version A) as JT-W). Once the generative model is estimated, the generative score spaces are calculated.

The approach herein proposed consists in defining a kernel between two observed objects x and x' as the composition of the score function with one of the JT kernels presented above. Formally,

$$k\big(x,x'\big)=k_q^i\big(\phi_\Theta(x),\phi_\Theta\big(x'\big)\big), \tag{10.15}$$

where $i\in\{\mathrm{JT},\mathrm{A},\mathrm{B}\}$ indexes one of the Jensen–Tsallis kernels, and ϕ_Θ is one of the generative embeddings defined in Sect. 10.5.3. Notice that this kernel is well defined because all the components of $\phi_\Theta^{\mathrm{FE}}$ are non-negative.

We consider two types of kernel-based classifiers: K-NN and SVM. Recall that positive definiteness is a key condition for the applicability of a kernel in SVM learning. It was shown in [52] that k_q^A is a positive definite kernel for $q\in[0,1]$, while k_q^B is a positive definite kernel for $q\in[0,2]$. Standard results from kernel theory [69, Proposition 3.22] guarantee that the kernel k defined in (10.15) inherits the positive definiteness of k_q^i, thus can be safely used in SVM learning algorithms.

10.8.2 Results

We compare the results of our proposed approach with the linear kernel as a reference (which is the most used solution in the hybrid generative discriminative approach case, e.g., the Fisher Kernel). As classifiers we used Support Vector Machines and K-Nearest Neighbor (with K set to 1, i.e., the nearest neighbor rule). When possible, the classifiers have been applied also in the original domain (namely without the application of the generative embedding step).

Results are displayed in Table 10.5. In the table, "NN" stands for nearest neighbor results, while "SVM" refers to SVM results. "Linear" is the linear kernel, whereas

Table 10.5 Results on the brain classification task. See text for details

Embedding	Linear		JS		JT		JT-W	
	NN	SVM	NN	SVM	NN	SVM	NN	SVM
TPM-1	51.56	50.00	54.22	59.56	50.33	62.67	58.44	64.33
TPM-2	61.00	68.56	58.89	68.89	54.33	65.78	63.11	70.22
FESS-1	56.11	50.00	56.89	50.00	50.00	36.89	58.44	50.00
FESS-2	62.89	50.00	62.67	60.00	50.00	67.44	60.11	72.00
LLR	57.33	50.00	58.78	61.56	50.00	63.78	61.44	63.56
FSH	61.78	50.00	58.44	70.22	55.33	67.33	61.89	69.89
TOP	51.89	50.00	51.89	50.00	50.00	50.00	50.00	50.00
PD	75.22	50.00	74.78	50.00	62.67	**80.56**	72.56	**80.78**
ORIG	61.00	77.00	60.22	74.33	50.33	70.56	50.00	73.78

"JS", "JT" and "JT-W" stand for Jensen–Shannon, Jensen–Tsallis, and Weighted Jensen–Tsallis kernels, respectively. The acronyms of the generative embeddings follow the notation described in Sect. 10.5.3: "TPM-1" is the posterior topic mixture for a single pLSA, "TPM-2" is the posterior topic mixture starting from one pLSA per class,"FESS-1" is the Free Energy Score Space for a single pLSA, "FESS-2" is the Free Energy Score Space obtained starting from one pLSA per class, "LLR" is the Log-Likelihood Ratio score space, "FSH" is the Fisher Score space, "TOP" is the TOP kernel score space and "PD" is the Posterior Divergence Score space. The standard errors of means, in all runs, were all less than 2.52.

From the table, different observations may be drawn:

- In almost all cases, the use of IT kernels over generative embeddings outperforms the linear kernel over the same embeddings, this being really evident in some cases.
- At the same time, the intermediate use of a generative embedding is almost always beneficial with respect to use the linear and the IT kernels on the original space.
- It is evident from the table that the best generative embedding is the very recently proposed Posterior Divergence Score Space. It seems that this generative embedding has a slight preference to be used with the IT kernels.
- There is no significant difference among the various IT kernels, even if it may be argued that the Weighted Jensen–Tsallis one is the most positive.
- Comparing the classifiers, there is no huge difference between the SVM and the Nearest Neighbor performances, thus confirming the goodness of the devised similarity measure.

10.9 Case Study 3: Scale Selection by MKL

In this case study, we use *Heat Kernel Signatures* to extract histogram based features (see also [16]) using different scales and using these as different sources for

Multiple Kernel Learning paradigm. The data is extracted from sMRI scans of the left thalamus of 30 schizophrenic patients and 30 healthy controls. Several kernels are computed (i.e., one kernel per scale), and a set of weights are estimated for the kernel combination. In this fashion, we can choose the most discriminative scales by selecting those associated to the highest weights, and vice versa. Moreover, kernel combination leads to a new similarity measure which increases the classification accuracy. It is important to note that in our approach we aim at selecting the best shape characteristics for classification purposes, hence, our selection is driven by the performance of a Support Vector Machine (SVM) classifier.

10.9.1 Methodology

The contribution of geometric features extracted at each scale are combined by employing the MKL strategy as described in Sect. 10.6.2. Each shape representation r_i is associated to a kernel k_m by leading to $n = P$ kernels. Indeed, both the weights $(\eta_1, \ldots, \eta_P)$ and the SVM parameters are estimated. In order to obtain the best classification accuracy, according to the *max-margin* paradigm an *alternating* approach is used between the optimization of kernel weights and the optimization of the SVM classifier. In each step, given the current solution of kernel weights, MKL solves a standard SVM optimization problem with the combined kernel. Then, a specific procedure is applied to update the kernel weights. Once the MKL procedure is completed, we obtain a two-fold advantage: (i) we can select the best scale contributions by keeping only the scales associated to the highest weighs, and (ii) we can compose a new kernel from the weighted contributions of the best scales, which can be evaluated for classification purposes.

10.9.2 Experimental Protocol

In our experiments, we apply leave-one-out (LOO) cross-validation to assess the performance of the technique. Since LOO is used as the cross-validation technique, we do not report standard deviations or variances. We compare our results using k-fold paired t-test at $p = 0.05$. We collect geometric features at 11 scales generating different shape representations $r_{01}, \ldots, r_{11}$. In practice, each representation r_i is a feature vector x_i which is plugged in the MKL framework. We employ the dot product as basic kernel function (i.e., linear kernel) since it avoids the estimation of free kernel parameters. Different strategies to combine the different shape representations have also been evaluated:

- Single Best Kernel (`Single-best`)—An SVM is trained separately per each representation. Therefore, the performances of the classification are evaluated separately at each scale. By doing so, we can evaluate the independent contributions coming from the different sources of information and select the best one.

Table 10.6 Single-kernel SVM accuracies

r01	r02	r03	r04	r05	r06	r07	r08	r09	r10	r11
75.00	**78.33**	76.67	76.67	73.33	*66.67	68.33	70.00	76.67	71.67	70.00

Table 10.7 MKL accuracies

SVM	SVM-con	RBMKL	SMKL	GLMKL
*78.3 (10, 11.7)	83.3 (8.3, 8.3)	*81.7 (10, 8.3)	**86.7 (6.7, 6.7)**	85.0 (8.3, 6.7)

- Feature concatenation (SVM-con)—The contributions coming from the different sources are concatenated into a single feature vector. Then, a single SVM is employed for classification [19].
- Rule-based MKL (RBMKL)—As baseline MKL approach, the so-called rule-based method is evaluated: the kernels computed at each scale are combined by simply taking their average (i.e., $\forall m, \eta_m = 1/P$).
- Simple MKL (SMKL)—A simple but effective MKL algorithm is employed [60] by addressing the MKL problem through a weighted 2-norm regularization formulation with additional constraint on the weights that encourages sparse kernel combination.
- Group Lasso MKL (GLMKL)—It denotes the group Lasso-based MKL algorithms proposed by [43, 86]. A closed form solution for optimizing the kernel weights based on the equivalence between group-lasso and MKL is proposed. In our implementation, we used l_1-norm on the kernel weights and learned a convex combination of the kernels.

10.9.3 Results

The first evaluation scores are shown in Table 10.6, which reports the single-best kernel accuracies for all feature representations. We can observe that the best performance is obtained at 78.33 % using r02 which is shown as bold face in the table. The entries marked with "*" show the accuracies which are statistically significantly less accurate than the best algorithm using k-fold paired t-test at $p = 0.05$.

Second, concatenating the features in a single vector leads to 83.33 % accuracy. Third, using the proposed three different MKL algorithms, we combined the eleven kernels by introducing the weights η_m. Table 10.7 reports the results of the best single-kernel SVM, the accuracy of the concatenated feature set, and the three MKL-based algorithms trained. The values in parentheses show the percentage of controls classified as schizophrenia and the percentage of patients classified as healthy, respectively. We achieve an accuracy of 86.67 %, reached by combining eleven kernels with the SMKL approach. This result is better than all other MKL settings and single-kernel SVMs. Further, GLMKL achieves 85 % accuracy which is

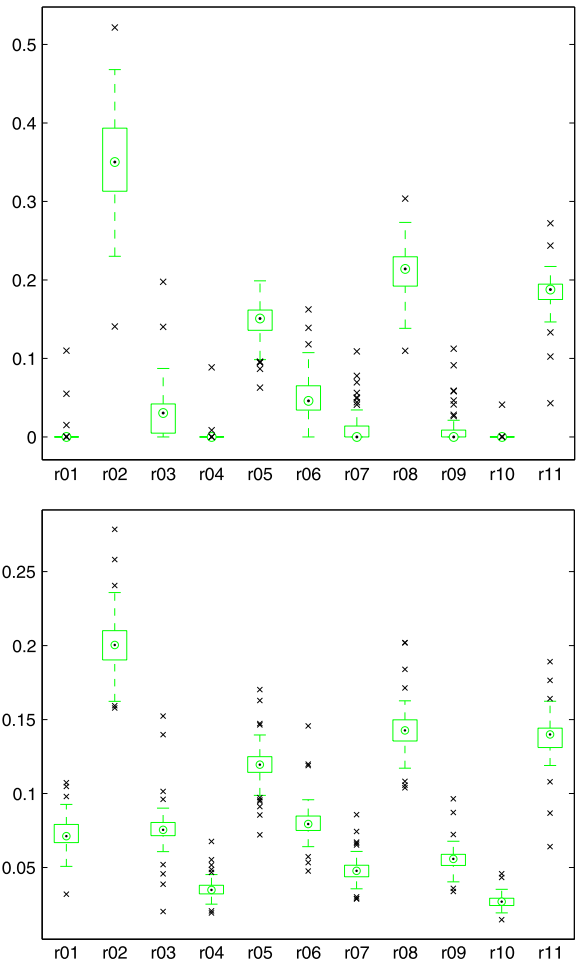

Fig. 10.7 Combination weights in MKL using the linear kernel: (*top*) using SMKL, (*bottom*) using GLMKL

still higher than that reached by the feature concatenation method. We can also note that we cannot overcome SVM-con when we use RBMKL, as the latter gives equal weight to each kernel. In fact, if there are inaccurate representations in the given set, the overall mean combination accuracy may be less of that reached using the single best. Conversely, when the weights are automatically estimated, such as in SMKL and GLMKL the selection of the most reliable information is carried out by the MKL procedure and the overall performance improves.

In Fig. 10.7, we plotted the weights of MKL for SMKL and GMKL algorithms to show the coherency of the weights. As expected, the best representation is r02, which has the highest weights. Although the other representations with high weights (r08, r11 and r05) do not provide accurate single-kernel SVMs results, their contributions to the overall accuracy in the combination is higher than those given by the other kernels. This demonstrates that when considering combinations, even a

Table 10.8 MKL accuracies on the selected subset of representations

SVM	SVM-con	RBMKL	SMKL	GLMKL
*78.3 (10, 11.7)	*83.3 (6.7, 10)	*83.3 (6.7, 10)	**88.3 (6.7, 5)**	85.0 (6.7, 8.3)

representation which does not lead to precise results may contribute to raise the overall combination accuracy. Moreover, we can also deduce that these four representations are the most useful in discriminating between healthy and schizophrenic subjects, and we may focus the attention on these properties only.

Using this information, we also performed the above pipeline using only these four representations, and we can observe the results in Table 10.8. Using this subset, we get the highest accuracy with SMKL, reaching 88.33 % of accuracy. We can also observe an increase in RBMKL.

10.9.4 Discussion

We have shown in general that MKL algorithms perform better than both single-best kernel SVMs and feature concatenation strategies. We have also observed that RBMKL (which does not compute weights while combining kernels) does not outperform the feature concatenation approach. Conversely, when the kernel combination is carried out by estimating proper weights, a substantial improvement is instead obtained. The kernel weights also allow us to extract useful information: it is interesting to observe that, for both MKL algorithms with the highest accuracy, four representations have the maximum effect (i.e., the highest weights), namely r02, r08, r11, and r05, with r02 being the best single-kernel. We use this information to select a smaller number of representations to reduce the costs of the feature extraction phase. Finally, we can also observe that by using such subset we can reach the best accuracy overall.

10.10 Conclusions

We have defined a set of new approaches to deal with schizophrenia detection from MRI images. We have proposed a working pipeline which takes into account different aspects of the disease. We have successfully applied the dissimilarity-based technique described in Chap. 2 to our medical application. In particular, we have shown that brain classification on dissimilarity space reaches a substantial improvement over standard feature-based approaches. Moreover, we have shown that combining dissimilarities represents a natural and effective approach to merge different sources of information. In this fashion, we were able to exploit complementary information about different parts of the brain, different acquisition modalities, and

different brain properties. Moreover, we have shown that our new paradigm to define data descriptors by generative embedding (see Chap. 4) is very effective and works well on our medical application. This research has opened new perspectives in the medical application which have been envisaged by our work. In particular, we have shown that a further improvement can be obtained by adapting random subspace method [81] to create the dissimilarity space.

Furthermore, we are working on employing advanced dissimilarity-based techniques to encode shape properties. Our preliminary results have shown an improvement by using Multiple Kernel Learning to improve the diffusion based shape description. Finally, we have shown in our experiments that DWI data was not important to improve the classification accuracy when multimodal approach was employed. This encourages us to exploit more advanced imaging techniques such as Diffusion Tensor MRI or Functional MRI to further improve schizophrenia detection.

References

1. Alpaydın, E.: Introduction to Machine Learning. MIT Press, Cambridge (2004)
2. Amaddeo, F., Tansella, M.: Information systems for mental health. Epidemiol. Psichiatr. Soc. **18**(1), 1–4 (2009)
3. Andreone, N., Tansella, M., Cerini, R., Versace, A., Rambaldelli, G., Perlini, C., Dusi, N., Pelizza, L., Balestrieri, M., Barbui, C., Nose, M., Gasparini, A., Brambilla, P.: Cortical white-matter microstructure in schizophrenia. diffusion imaging study. Br. J. Psychiatry **191**, 113–119 (2007)
4. Ashburner, J., Friston, K.J.: Voxel-based morphometry-the methods. NeuroImage **11**(6), 805–821 (2000)
5. Awate, S.P., Yushkevich, P., Song, Z., Licht, D., Gee, J.C.: Multivariate high-dimensional cortical folding analysis, combining complexity and shape, in neonates with congenital heart disease. In: Proceedings of the 21st International Conference on Information Processing in Medical Imaging, IPMI'09, pp. 552–563 (2009)
6. Bach, F.R., Lanckriet, G.R.G., Jordan, M.I.: Multiple kernel learning, conic duality, and the SMO algorithm. In: Proceedings of the 21st International Conference on Machine Learning, pp. 41–48 (2004)
7. Baiano, M., Perlini, C., Rambaldelli, G., Cerini, R., Dusi, N., Bellani, M., Spezzapria, G., Versace, A., Balestrieri, M., Mucelli, R.P., Tansella, M., Brambilla, P.: Decreased entorhinal cortex volumes in schizophrenia. Schizophr. Res. **102**(1–3), 171–180 (2008)
8. Bellani, M., Brambilla, P.: The use and meaning of the continuous performance test in schizophrenia. Epidemiol. Psichiatr. Soc. **17**(3), 188–191 (2008)
9. Bicego, M., Lovato, P., Ferrarini, A., Delledonne, M.: Biclustering of expression microarray data with topic models. In: Proceedings of the International Conference on Pattern Recognition, pp. 2728–2731 (2010)
10. Bicego, M., Lovato, P., Oliboni, B., Perina, A.: Expression microarray classification using topic models. In: Proceedings of the 2010 ACM Symposium on Applied Computing, SAC'10, New York, NY, USA, pp. 1516–1520 (2010)
11. Bicego, M., Pekalska, E., Tax, D.M.J., Duin, R.P.W.: Component-based discriminative classification for hidden Markov models. Pattern Recognit. **42**, 2637–2648 (2009)
12. Bosch, A., Zisserman, A., Munoz, X.: Scene classification via pLSA. In: Proceedings of the European Conference on Computer Vision, ECCV'06, pp. 517–530 (2006)

13. Brambilla, P., Tansella, M.: Can neuroimaging studies help us in understanding the biological causes of schizophrenia? Int. Rev. Psychiatry **19**(4), 313–314 (2007)
14. Bronstein, A.M., Bronstein, M.M.: Shape recognition with spectral distances. IEEE Trans. Pattern Anal. Mach. Intell. **33**(5), 1065–1071 (2011)
15. Browne, A., Jakary, A., Vinogradov, S., Fu, Y., Deicken, R.: Automatic relevance determination for identifying thalamic regions implicated in schizophrenia. IEEE Trans. Neural Netw. **19**(6), 1101–1107 (2008)
16. Castellani, U., Mirtuono, P., Murino, V., Bellani, M., Rambaldelli, G., Tansella, M., Brambilla, P.: A new shape diffusion descriptor for brain classification. In: Fichtinger, G., Martel, A., Peters, T. (eds.) Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, MICCAI'11. Lecture Notes in Computer Science, vol. 6892, pp. 426–433 (2011)
17. Castellani, U., Perina, A., Murino, V., Bellani, M., Rambaldelli, G., Tansella, M., Brambilla, P.: Brain morphometry by probabilistic latent semantic analysis. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, MICCAI'10, MICCAI, pp. 177–184 (2010)
18. Cha, S.H., Srihari, S.N.: On measuring the distance between histograms. Pattern Recognit. **35**(6), 1355–1370 (2002)
19. Chang, C.C., Lin, C.J.: In: LIBSVM: a Library for Support Vector Machines (2001). http://www.csie.ntu.edu.tw/~cjlin/libsvm
20. Cheng, D.S., Bicego, M., Castellani, U., Cerruti, S., Bellani, M., Rambaldelli, G., Atzori, M., Brambilla, P., Murino, V.: Schizophrenia classification using regions of interest in brain MRI. In: Proceedings of Intelligent Data Analysis in Biomedicine and Pharmacology, IDAMAP'09, pp. 47–52 (2009)
21. Corradi-Dell'Acqua, C., Tomelleri, L., Bellani, M., Rambaldelli, G., Cerini, R., Pozzi-Mucelli, R., Balestrieri, M., Tansella, M., Brambilla, P.: Thalamic-insular disconnectivity in schizophrenia: evidence from structural equation modeling. Hum. Brain Mapp. **33**, 740–752 (2012)
22. Cortes, C., Mohri, M., Rostamizadeh, A.: Learning non-linear combinations of kernels. In: Advances in Neural Information Processing Systems, vol. 22, pp. 396–404 (2010)
23. Cristani, M., Perina, A., Castellani, U., Murino, V.: Geo-located image analysis using latent representations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8 (2008)
24. Cuturi, M., Fukumizu, K., Vert, J.P.: Semigroup kernels on measures. J. Mach. Learn. Res. **6**, 1169–1198 (2005)
25. Davatzikos, C.: Why voxel-based morphometric analysis should be used with great caution when characterizing group differences. NeuroImage **23**(1), 17–20 (2004)
26. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. B **39**(1), 1–38 (1977)
27. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification, 2nd edn. Wiley-Interscience, New York (2000)
28. Duin, R.P.W.: Prtools, a Matlab toolbox for pattern recognition version 4.0.14 (2005). http://www.prtools.org/
29. Edelstein, W.A., Bottomley, P.A., Pfeifer, L.M.: A signal-to-noise calibration procedure for NMR imaging systems. Med. Phys. **11**, 180–185 (1984)
30. Fan, Y., Shen, D., Gur, R.C., Gur, R.E., Davatzikos, C.: COMPARE: classification of morphological patterns using adaptive regional elements. IEEE Trans. Med. Imaging **26**(1), 93–105 (2007)
31. Frey, B.J., Jojic, N.: A comparison of algorithms for inference and learning in probabilistic graphical models. IEEE Trans. Pattern Anal. Mach. Intell. **27**(9), 1392–1416 (2005)
32. Gerig, G., Styner, M., Shenton, M.E., Lieberman, J.A.: Shape versus size: improved understanding of the morphology of brain structures. In: Proceedings of the International Conference on Medical Image Computing, MICCAI'01, pp. 24–32 (2001)

33. Giuliani, N.R., Calhouna, V.D., Pearlson, G.D., Francis, A., Buchanan, R.W.: Voxel-based morphometry versus region of interest: a comparison of two methods for analyzing gray matter differences in schizophrenia. Schizophr. Res. **74**(2–3), 135–147 (2005)
34. Gönen, M., Alpaydın, E.: Localized multiple kernel learning. In: Proceedings of the 25th International Conference on Machine Learning, pp. 352–359 (2008)
35. Gönen, M., Alpaydın, E.: Multiple kernel learning algorithms. J. Mach. Learn. Res. **12**, 2181–2238 (2011)
36. Gönen, M., Ulaş, A., Schüffler, P.J., Castellani, U., Murino, V.: Combining data sources non-linearly for cell nucleus classification of renal cell carcinoma. In: Pelillo, M., Hancock, E.R. (eds.) Proceedings of the International Workshop on Similarity-Based Pattern Analysis, SIMBAD'11. Lecture Notes in Computer Science, vol. 7005, pp. 250–260. Springer, Berlin (2011)
37. Hofmann, T.: Learning the similarity of documents: an information-geometric approach to document retrieval and categorization. In: Proceedings of the Conference on Advances in Neural Information Processing Systems, NIPS'02, pp. 914–920 (2000)
38. Hofmann, T.: Unsupervised learning by probabilistic latent semantic analysis. Mach. Learn. **42**(1–2), 177–196 (2001)
39. Jaakkola, T.S., Haussler, D.: Exploiting generative models in discriminative classifiers. In: Proceedings of the Conference on Advances in Neural Information Processing Systems, NIPS'98, Cambridge, MA, USA, vol. 11, pp. 487–493 (1998)
40. Jebara, T., Kondor, R., Howard, A.: Probability product kernels. J. Mach. Learn. Res. **5**, 819–844 (2004)
41. Kawasaki, Y., Suzuki, M., Kherif, F., Takahashi, T., Zhou, S.Y., Nakamura, K., Matsui, M., Sumiyoshi, T., Seto, H., Kurachi, M.: Multivariate voxel-based morphometry successfully differentiates schizophrenia patients from healthy controls. NeuroImage **34**(1), 235–242 (2007)
42. Kittler, J., Hatef, M., Duin, R.P.W., Matas, J.: On combining classifiers. IEEE Trans. Pattern Anal. Mach. Intell. **20**(3), 226–239 (1998)
43. Kloft, M., Brefeld, U., Sonnenburg, S., Zien, A.: l_p-norm multiple kernel learning. J. Mach. Learn. Res. **12**, 953–997 (2011)
44. Koenderink, J.J., van Doorn, A.J.: Surface shape and curvature scales. Image Vis. Comput. **10**, 557–565 (1992)
45. Kuncheva, L.I.: Combining Pattern Classifiers: Methods and Algorithms. Wiley-Interscience, New York (2004)
46. Lanckriet, G.R.G., Cristianini, N., Bartlett, P., Ghaoui, L.E., Jordan, M.I.: Learning the kernel matrix with semidefinite programming. J. Mach. Learn. Res. **5**, 27–72 (2004)
47. Lee, W.J., Duin, R.P.W., Loog, M., Ibba, A.: An experimental study on combining Euclidean distances. In: 2nd International Workshop on Cognitive Information Processing (CIP), pp. 304–309 (2010)
48. Lewis, D.P., Jebara, T., Noble, W.S.: Nonstationary kernel combination. In: Proceedings of the 23rd International Conference on Machine Learning, pp. 553–560 (2006)
49. Li, X., Lee, T.S., Liu, Y.: Hybrid generative-discriminative classification using posterior divergence. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR'11, pp. 2713–2720 (2011)
50. Ling, H., Okada, K.: Diffusion distance for histogram comparison. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR'06, vol. 1, pp. 246–253 (2006)
51. Liu, Y., Teverovskiy, L., Carmichael, O., Kikinis, R., Shenton, M., Carter, C.S., Stenger, V.A., Davis, S., Aizenstein, H., Becker, J.T., Lopez, O.L., Meltzer, C.C.: Discriminative MR image feature analysis for automatic schizophrenia and Alzheimer's disease classification. In: Proceedings of the Medical Image Computing and Computer-Assisted Intervention, MICCAI'04, pp. 393–401 (2004)
52. Martins, A.F.T., Smith, N.A., Xing, E.P., Aguiar, P.M.Q., Figueiredo, M.A.T.: Nonextensive information theoretic kernels on measures. J. Mach. Learn. Res. **10**, 935–975 (2009)
53. Ng, A.Y., Jordan, M.I.: On discriminative vs. generative classifiers: a comparison of logistic regression and naive Bayes. In: Proceedings of the Conference on Advances in Neural Infor-

mation Processing Systems, NIPS'02, vol. 14, pp. 841–848 (2002)
54. Nyúl, L.G., Udupa, J.K., Zhang, X.: New variants of a method of MRI scale standardization. IEEE Trans. Med. Imaging **19**(2), 143–150 (2000)
55. Pekalska, E., Duin, R.P.W.: The Dissimilarity Representation for Pattern Recognition. Foundations and Applications. World Scientific, Singapore (2005)
56. Perina, A., Cristani, M., Castellani, U., Murino, V., Jojic, N.: Free energy score space. In: Proceedings of the Conference on Advances in Neural Information Processing Systems, NIPS'09, vol. 22, pp. 1428–1436 (2009)
57. Perina, A., Cristani, M., Castellani, U., Murino, V., Jojic, N.: A hybrid generative/discriminative classification framework based on free-energy terms. In: Proceedings of the IEEE International Conference on Computer Vision, ICCV'09, pp. 2058–2065 (2009)
58. Pohl, K.M., Sabuncu, M.R.: A unified framework for MR based disease classification. In: IPMI'09: Proceedings of the 21st International Conference on Information Processing in Medical Imaging, pp. 300–313 (2009)
59. Pruessner, J., Li, L., Serles, W., Pruessner, M., Collins, D., Kabani, N., Lupien, S., Evans, A.: Volumetry of hippocampus and amygdala with high-resolution MRI and three-dimensional analysis software: minimizing the discrepancies between laboratories. Cereb. Cortex **10**(4), 433–442 (2000)
60. Rakotomamonjy, A., Bach, F.R., Canu, S., Grandvalet, Y.: Simple MKL. J. Mach. Learn. Res. **9**, 2491–2521 (2008)
61. Raviv, D., Bronstein, A.M., Bronstein, M.M., Kimmel, R.: Volumetric heat kernel signatures. In: Workshop on 3D Object Retrieval, pp. 39–44 (2010)
62. Ray, K.M., Wang, H., Chu, Y., Chen, Y.F., Bert, A., Hasso, A.N., Su, M.Y.: Mild cognitive impairment: apparent diffusion coefficient in regional gray matter and white matter structures. Radiology **24**, 197–205 (2006)
63. Reuter, M., Wolter, F.E., Shenton, M., Niethammer, M.: Laplace–Beltrami eigenvalues and topological features on eigenfunctions for statistical shape analysis. Comput. Aided Des. **41**(10), 739–755 (2009)
64. Rovaris, M., Bozzali, M., Iannucci, G., Ghezzi, A., Caputo, D., Montanari, E., Bertolotto, A., Bergamaschi, R., Capra, R., Mancardi, G.L., Martinelli, V., Comi, G., Filippi, M.: Assessment of normal-appearing white and gray matter in patients with primary progressive multiple sclerosis—a diffusion-tensor magnetic resonance imaging study. Arch. Neurol. **59**, 1406–1412 (2002)
65. Rubner, Y., Tomasi, C., Guibas, L.J.: The Earth mover's distance as a metric for image retrieval. Int. J. Comput. Vis. **40**(2), 99–121 (2000)
66. Rujescu, D., Collier, D.A.: Dissecting the many genetic faces of schizophrenia. Epidemiol. Psichiatr. Soc. **18**(2), 91–95 (2009)
67. Schwarz, G.: Estimating the dimension of a model. Ann. Stat. **6**, 461–464 (1979)
68. Serratosa, F., Sanfeliu, A.: Signatures versus histograms: definitions, distances and algorithms. Pattern Recognit. **39**(5), 921–934 (2006)
69. Shawe-Taylor, J., Cristianini, N.: Kernel Methods for Pattern Analysis. Cambridge University Press, Cambridge (2004)
70. Shenton, M.E., Dickey, C.C., Frumin, M., McCarley, R.W.: A review of MRI findings in schizophrenia. Schizophr. Res. **49**(1–2), 1–52 (2001)
71. Smith, N., Gales, M.: Speech recognition using SVMs. In: Proceedings of the Conference on Advances in Neural Information Processing Systems, NIPS'02, vol. 14, pp. 1197–1204 (2002)
72. Smith, N.D., Gales, M.J.F.: Using SVMs to classify variable length speech patterns. Tech. Rep. CUED/F-INFENG/TR–412, Cambridge University Engineering Department (2002)
73. Sun, J., Ovsjanikov, M., Guibas, L.: A concise and provably informative multi-scale signature based on heat diffusion. In: Proceedings of the Symposium on Geometry Processing, SGP'09, pp. 1383–1392 (2009)
74. Swain, M.J., Ballard, D.H.: Color indexing. Int. J. Comput. Vis. **7**(1), 11–32 (1991)

75. Taylor, W.D., Hsu, E., Krishnan, K.R.R., MacFall, J.R.: Diffusion tensor imaging: background, potential, and utility in psychiatric research. Biol. Psychiatry **55**(3), 201–207 (2004)
76. Timoner, S.J., Golland, P., Kikinis, R., Shenton, M.E., Grimson, W.E.L., Wells III, W.M.: Performance issues in shape classification. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, MICCAI'02, pp. 355–362 (2002)
77. Toews, M., Wells III, W., Collins, D.L., Arbel, T.: Feature-based morphometry. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, MICCAI'09, pp. 109–116 (2009)
78. Tomasino, B., Bellani, M., Perlini, C., Rambaldelli, G., Cerini, R., Isola, M., Balestrieri, M., Caligrave, S., Versace, A., Mucelli, R.P., Gasparini, A., Tansella, M., Brambilla, P.: Altered microstructure integrity of the amygdala in schizophrenia: a bimodal MRI and DWI study. Psychol. Med. **41**(2), 301–311 (2010)
79. Tsuda, K., Kawanabe, M., Rätsch, G., Sonnenburg, S., Müller, K.R.: A new discriminative kernel from probabilistic models. Neural Comput. **14**, 2397–2414 (2002)
80. Ulaş, A., Castellani, U., Mirtuono, P., Bicego, M., Murino, V., Cerruti, S., Bellani, M., Atzori, M., Rambaldell, G., Tansella, M., Brambilla, P.: Multimodal schizophrenia detection by multiclassification analysis. In: Martín, C.S., Kim, S.W. (eds.) Proceedings of the Iberoamerican Congress on Pattern Recognition, CIARP'11. Lecture Notes in Computer Science, vol. 7042, pp. 491–498. Springer, Berlin (2011)
81. Ulaş, A., Castellani, U., Murino, V., Bellani, M., Tansella, M., Brambilla, P.: Heat diffusion based dissimilarity analysis for schizophrenia classification. In: M.L. et al. (ed.) IAPR International Conference on Pattern Recognition in Bioinformatics, PRIB'11. Lecture Notes in Bioinformatics, vol. 7036, pp. 306–317. Springer, Berlin (2011)
82. Ulaş, A., Duin, R.P.W., Castellani, U., Loog, M., Mirtuono, P., Bicego, M., Murino, V., Bellani, M., Cerruti, S., Tansella, M., Brambilla, P.: Dissimilarity-based detection of schizophrenia. Int. J. Imaging Syst. Technol. **21**(2), 179–192 (2011)
83. Ulaş, A., Yıldız, O.T., Alpaydın, E.: Eigenclassifiers for combining correlated classifiers. Inf. Sci. **187**, 109–120 (2012)
84. Vapnik, V.N.: Statistical Learning Theory. Wiley, New York (1998)
85. Voets, N.L., Hough, M.G., Douaud, G., Matthews, P.M., James, A., Winmill, L., Webster, P., Smith, S.: Evidence for abnormalities of cortical development in adolescent-onset schizophrenia. NeuroImage **43**(4), 665–675 (2008)
86. Xu, Z., Jin, R., Yang, H., King, I., Lyu, M.R.: Simple and efficient multiple kernel learning by group Lasso. In: Proceedings of the 27th International Conference on Machine Learning, ICML'10, pp. 1175–1182 (2010)
87. Yoon, U., Lee, J.M., Im, K., Shin, Y.W., Cho, B.H., Kim, I.Y., Kwon, J.S., Kim, S.I.: Pattern classification using principal components of cortical thickness and its discriminative pattern in schizophrenia. NeuroImage **34**(4), 1405–1415 (2007)