UNIVERSITA' DEGLI STUDI DI MILANO-BICOCCA

DIPARTIMENTO DI FISICA G. OCCHIALINI

CORSO DI DOTTORATO IN FISICA E ASTRONOMIA

CICLO XXVIII

DEVELOPMENT AND VALIDATION OF A DECISION SUPPORT SYSTEM FOR THE AUTOMATIC DIAGNOSIS OF MEDICAL IMAGES FROM BRAIN MRI STUDIES

CHRISTIAN SALVATORE

Supervisor: Prof. Marco Paganoni

External supervisor: Dott.ssa Isabella Castiglioni

Coordinator of PhD school: Prof. Giberto Chirico

Academic Year 2014-2015

A mia moglie Alice.

Ai miei genitori.

A mia sorella,
alla mia famiglia
e ai miei amici.

Prediction is very difficult,
especially if it's about the future.
- Niels Bohr -

Table of contents

Abstract

Decision support systems for the assisted diagnosis in medicine are computer-based information systems designed to assist physicians and clinicians with decision-making tasks by automatically determining diagnosis, or improving the diagnostic confidence. This could result in the possibility to perform early and differential diagnosis of neurodegenerative pathologies, such as Alzheimer's Disease (AD) and Parkinson's Disease (PD), for which definite diagnosis still remains a crucial issue.

Among decision support systems, multivariate Machine Learning (ML) methods are recently growing in popularity within the neuroimaging community. Among these, supervised ML methods are based on algorithms able to automatically extract multiple information from image sets without requiring a-priori hypotheses of where this information may be coded in the images. These methods have been proposed as a revolutionary approach for identifying sensitive biomarkers (or combinations of them) allowing for automatic classification of individual subjects.

The aim of this thesis was to implement, optimize and validate a ML method able to perform automatic diagnosis of medical images by means of structural Magnetic Resonance Imaging (MRI) data. This method consists of 3 phases: 1) image pre-processing, mainly devoted to the co-registration of data from different patients to the same reference system (i.e., a standard coordinate space); 2) feature extraction and selection, performed through Principal Components Analysis and Fisher's Discriminant Ratio, respectively, with the aim of extracting and selecting the most discriminative features for classification; 3) classification, performed by Support Vector Machine (SVM), with the aim of estimating the parameters that define the predictive model to be used for the classification of new (unseen) subjects.

Moreover, in order to allow the identification of new MRI-related biomarkers useful for the diagnosis of the considered pathology, I also developed and implemented a method for the generation of pattern distribution maps of brain structural differences, which reflect the importance of each image voxel for SVM classification. This point results to be particularly important because these maps could help to identify potential biomarkers for the diagnosis of neurological diseases.

In order to test the feasibility of the implemented ML method, I applied it to the diagnosis of 3 different pathologies: AD, PD and Eating Disorders (ED).

For the application to the diagnosis of PD, we acquired T1-weighted brain structural MR images of 56 patients and 28 healthy control (CN) subjects. The group of 56 patients consisted of 28 patients with clinically diagnosed PD and 28 with clinical diagnosis of probable or possible Progressive Supranuclear Palsy (PSP), that is a parkinsonian condition with similar symptoms to PD but different treatment and

prognosis. The classifier allowed individual differential diagnosis with the following accuracy (specificity/sensitivity): in Leave-One-Out validation, PD vs. CN 92.7 (92.3/93.4)%, PSP vs. CN 97.0 (98.2/95.9)%, PSP vs. PD 98.2 (98.8/97.8)%; in half-splitting validation, PD vs. CN 93.5 (90.6/97.4)%, PSP vs. CN 92.2 (92.5/92.4)%, PSP vs. PD 92.2 (91.3/94.4)%. The following MRI-related brain biomarkers were identified to be used for the differential diagnosis of PD and PSP: midbrain, pons, corpus callosum and thalamus, four critical regions which are highly consistent with typical neuropathological and imaging findings described in patients with PSP.

For the application to the diagnosis of AD, we obtained T1-weighted brain structural MR images of patients from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database. We enrolled 162 CN, 137 AD, 76 Mild Cognitive Impairment (MCI) patients who converted to AD within 18 months (MCIc) and 134 MCI who did not convert to AD within 18 months (MCInc). A total of 509 subjects from 41 different radiology centers were considered. Classification performances evaluated by nested Cross Validation were 0.76 ± 0.11 for AD vs. CN, 0.72 ±6 0.12 for MCIc vs. CN, 0.66 ± 0.16 for MCIc vs. MCInc. Voxels influencing the classification of AD with respect to CN resulted to be localized in the temporal pole, superior temporal cortex, medial temporal cortex including hippocampus and entorhinal cortex, amygdala, thalamus, putamen, caudate, insular cortex, gyrus rectus, lateral orbitofrontal cortex, inferior frontal cortex, superior frontal cortex, anterior cingulate cortex, precuneus and in the posterior cerebellar lobule. The major part of voxels influencing the classification of both MCIc vs. CN and MCIc vs. MCInc was similar to the one previously found in AD.

The application of the proposed method to ED was made with the aim of extracting the most important voxels for the classification. For this study, we enrolled 17 ED patients and 17 CN. T1-weighted brain structural MR images were acquired and submitted to the ML method. The classifier allowed individual ED vs. CN diagnosis with accuracy (specificity/sensitivity) of 0.85 (0.73/0.93). The voxel-based pattern distribution map of brain structural differences showed that voxels influencing ED vs. CN discrimination were localized in the occipital cortex, posterior cerebellar lobule, precuneus, sensorimotor and premotor cortices, anterior cingulated cortex and orbitofrontal cortex, all brain regions involved in the regulation of appetite and emotional processing.

The reported results encourage the application of ML methods as decision support systems for the assisted diagnosis in clinical practice by means of brain structural MRI studies. Moreover, these methods could allow the identification of possible MRI-related biomarkers useful for the diagnosis of the considered disease.

Chapter 1.

INTRODUCTION

The aim of this chapter is to introduce the reader to the themes on which the whole thesis is based. In particular, a description of the physical principles on which Nuclear Magnetic Resonance (NMR) is based will be provided, as it is the method we used for extracting information about the patient's status, as it will be deeply explained in Chapter 2. Moreover, Decision Support Systems (DSSs) will be outlined, with a particular remark on Support Vector Machine (SVM), that is a multivariate machine learning (ML) method able to perform binary classification and that was used during this thesis as method to predict the status of a patient starting from his NMR data (see Chapter 2).

1.1 Nuclear Magnetic Resonance

The Nuclear Magnetic Resonance (NMR), on which Magnetic Resonance Imaging (MRI) is based, was discovered in 1946 by physicists Felix Bloch and Edward Purcell. NMR is a technique that exploits the ability of atomic nuclei to absorb energy from a magnetic field and electromagnetic radiation release. The nucleus of an atom consists of neutrons (particles at zero charge) and protons (positively charged particles) and has its own characteristic angular momentum (spin). The consequent rotation of the protons generates a magnetic field known as magnetic moment $\vec{\mu}$, oriented as the spin vector $\vec{I}$

$$\vec{\mu} = \frac{\gamma h}{2\pi}\vec{I} \qquad (1.1.1)$$

with γ gyromagnetic ratio and h Planck's constant. Generally, the nuclear magnetic moment of nuclei in a body is casually oriented, with a resultant magnetization equal to zero.

When the body is placed in an external magnetic field, this condition causes the orientation of the magnetic moments of the paramagnetic nuclei, such as hydrogen, gadolinium, and manganese, according to the direction of the field. As a consequence, a not-null total magnetization is produced.

In presence of a static magnetic field which produces a spin polarization, a radiofrequency signal of a proper frequency can induce a transition between spin states. This proper frequency is known as the Larmor frequency and it depends on the gyromagnetic ratio of the nucleus and on the magnetic field, as follows

$$\upsilon_0 = \frac{\gamma}{2\pi} B_0 \qquad (1.1.2)$$

This is the frequency of the nuclei precession around the magnetic field. The radio frequency signals can induce a rotation of the spin of 90 or 180 degrees.

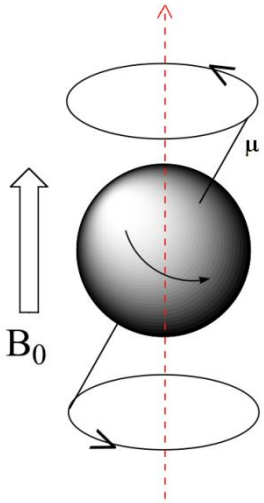


Figure 1.1.1 *Representation of the magnetic moment μ*

The *spin flip*, caused by the absorption of the energy of the signal, places some of the spins in their higher energy state. If the radio frequency signal is then switched off, the relaxation of the spins back to the lower state produces a measurable amount of RF signal at the resonant frequency associated with the spin flip.

The relaxation can only occur if the system is capable of exchanging energy. The nuclei can in fact change the energy level exchanging their magnetic energy with the thermal energy of the sample (spin-lattice interaction) or giving each other magnetic energy (spin-spin interactions, which do not alter the populations of the magnetic levels).

The first interaction consists for the longitudinal magnetization vector Mz into returning to equilibrium and it is characterized by a time constant T_1 (that is the time taken by the spins to recover 63% of the longitudinal magnetization).

The spin-lattice relaxation time (T_1) depends on the composition and the structure of the sample and on the intensity of the magnetic field. Materials with molecules having great freedom of movements (such as water) and a consequent difficulty for the molecules themselves to release energy, have very long T_1 constant, just like in the case of a strong magnetic field, which causes precession of nuclei in a faster way.

Figure 1.1.2 *Schematic representation of the longitudinal magnetization after the application of a radio frequency signal*

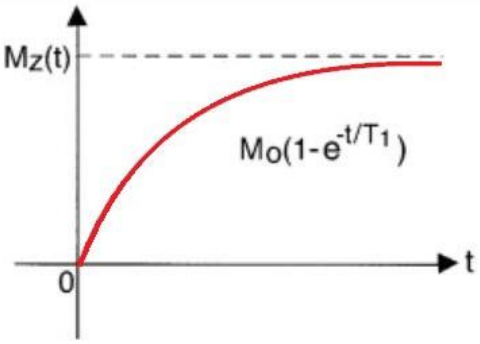


Figure 1.1.3 *Trend of the longitudinal magnetization after the application of a radio frequency signal*

The spin-spin interaction consists in the decay of the transverse magnetization vector Mxy due to the loss of phase coherence of the nuclear magnetic moments and it is characterized by a time constant T_2 (time required for the transverse magnetization to return to 37% of its original value).

The spin-spin relaxation time (T_2) expresses the interaction between the intrinsic magnetic moment of a nucleus and that of nearby nuclei, this parameter giving indications of magnetic homogeneity of the sample and about chemical nature of the environment around each single nucleus.

The effective magnetic field experienced by the individual spin is not equal for everyone. For example, the spins in itself represent a source of additive magnetic fields (micro-spatial different distribution of spin is reflected in a different manner on the single spin) and the motion of electrons produces magnetic fields too, with a consequent screen effect. T_2 is influenced by the inhomogeneity of external and local magnetic fields.

The two ways of relaxation give the possibility to weight the images according to one of the time-constants (T_1 or T_2). T_1-weighted images show tissues with the following gray scale: lipid tissues (brightest), muscles, cartilage, fluids, ligaments and bones (darkest). T_1-weighted sequences are often collected before and after infusion of MRI contrast agents. In the brain T_1-weighted scans provide appreciable contrast between gray and white matter.

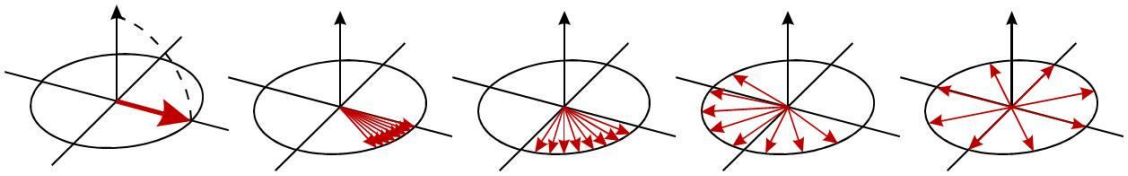


Figure 1.1.4 *Schematic representation of the transversal magnetization after the application of a radio frequency signal*

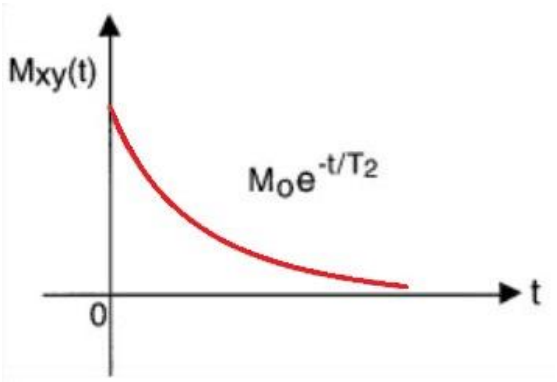


Figure 1.1.5 *Trend of the transversal magnetization after the application of a radio frequency signal*

Like the T_1-weighted scan, lipid's tissues are differentiated from water, but in this case fat appears to be darker and water appears to be brighter. In the case of cerebral and spinal study, the CSF (cerebrospinal fluid) appears brighter in T_2-weighted images. These scans are therefore particularly well suited to imaging edema.

There is a third possible weight for the images and it is the proton density (ρ). Proton density, also called Spin density, expresses the ratio between the number of resonant protons and the volume of the tissue under exam. Differently from T_1 or T_2 weighted scans, signal changes only come from differences in the amount of available hydrogen nuclei in the water molecules. This is a consequence of the fact that a significant contribution to the signal is given not by all protons present in the unit volume, but only by those contained in the hydrogen nuclei from liquid water. For example, ice gives no signal and blood gives a much higher one.

Finally, among the most used, there are two kind RF pulse sequences: saturation recovery and spin-echo. The first one is the simplest sequence and it is constituted by two pulses at 90 degrees, which completely overturn the longitudinal magnetization Mz in the xy plane. During the time interval between one pulse and the successive, the magnetization along the z axis increases progressively up to the original value.

The second pulse sequence alternates a 90 degrees pulse with a 180 degrees pulse. After the first impulse, due to local magnetic field inhomogeneities (variations in the magnetic field in different regions of the sample that are constant in time), some spins slow down due to lower local field strength, while some speed up due to higher field strength, this causing the Mxy signal decay. During this decay, the second pulse is applied so that the slower spins lead ahead of the main moment and the fast ones trail behind. Progressively, the fast moments reach the main moment and the slow moments drift back toward the main moment, causing a new increase of Mxy and the fall of the magnetization along the z axis.

In clinical applications in which $T_E/2 \gg T_R$, where T_R is the recovery time (time between two 90° pulses) and T_E is the echo time (time between the 90° pulse and the maximum amplitude of the echo), the measured signal (S) is equal to

$$S \propto \rho(1 - e^{-\frac{T_R}{T_1}})e^{-\frac{T_E}{T_2}} \qquad (1.1.3)$$

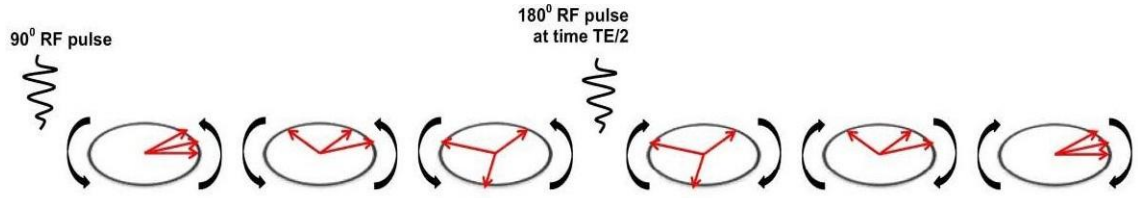


Figure 1.1.6 *Schematic representation of the transversal magnetization with two RF pulses*

The choice of T_R and T_E allows one to choose between T_1, T_2 or ρ weighted images. T_1-weighted images can be obtained with long T_R and T_E ($T_E \ll T_2$), T_2-weighted images need both T_R and T_E to be short ($T_R \gg T_1$) and proton density images need long $T_R$and short T_E ($T_E \ll T_2$ and $T_R \gg T_1$).

1.2 Decision Support Systems for assisted diagnosis

A (clinical) decision support system (DSS) is an interactive computer-based information system which is designed to assist physicians and other health professionals with decision-making tasks, such as determining diagnosis of patient data. In other words, DSSs link health observations with clinical knowledge in order to influence choices with the aim of improving the diagnostic accuracy (Miller et al., 1994). Such an active-knowledge system is usually based on a database of data or images, that is used to compare an individual data (coming from the patient to be studied) with reference images (the database), in order to generate case-specific clinical advices. The main feature offered by these methods is the one of giving objective assessments (i.e., free from arbitrary clinical reasoning) which are based on mathematical models for data analysis; resulting information has to be available in a reasonably short interval of time (if compared with typically diagnostic times). This approach is less biased than qualitative visual inspection of images and it provides diagnostic signs that are otherwise easily missed. The use of DSSs has been proven to improve diagnostic accuracy: a systematic review (Garg et al., 2005) of 100 studies concluded that DSSs improved practitioner performance in 64% of the studies; another systematic review (quantitative analysis) of 70 studies (Kawamoto et al., 2005) found that DSSs significantly improved clinical practice in 68% of trials. Thanks to the large number of evaluations proving their usefulness and, particularly, to the increased performance of imaging systems (e.g., spatial resolution, acquisition protocols, reconstruction algorithms), DSSs are ready to become clinically available. Indeed, all these emerging approaches are possible solutions to address the clinical need of making accurate and early diagnosis and to enhance the diagnostic confidence about a specific disease (also during monitoring of diseases progression). Nevertheless, their availability in current clinical practice is still limited by the need of a large number of control data (even healthy subjects) and the huge amount of computational cost needed for algorithm processing.

Until the last few years, basing on the method they make use of, the most promising DSSs included Region-of-interest measures, image segmentation and mass univariate statistical approaches, e.g. Voxel-Based Morphometry (Ashburner and Friston, 2000), that compares an individual image (for example, a MR image) with reference images of control subjects.

However, recent advances in statistical learning have attracted strong interest toward multivariate pattern analysis (MVPA). This approach aims at analyzing a distributed pattern of activity in order to extract the information of interest, which can be represented, for example, by a single variable of interest (such as the diagnostic label of a patient).

In general, the main advantage of MVPA-based methods naturally descends from their multivariate nature. Specifically, they are able to detect spatially distributed activations and cerebral patterns over a set of pixels. In other words, MVPA-based methods are able to automatically extract multiple information from image sets without requiring a priori hypotheses of where this information may be coded in the images. Because of this, MVPA-based methods show a relatively higher sensitivity than methods based on conventional univariate analysis (Mahmoudi et al., 2012; Pereira et al., 2009; Schrouff et al., 2013).

Interestingly, these methods showed promising results in medicine when coupled with a ML approach (e.g. Salvatore et al., 2015b). The main characteristic of ML methods is that they are able to design a predictive model from a set of training data, and this model can subsequently be used to evaluate information of interest about new (unseen) data.

Among MVPA ML methods, Support Vector Machine (SVM) attracted strong interest within the neuroimaging community for its ability in the classification medical images. In the following section, a description of SVM will be given, that is, a multivariate and supervised classifier able to separate binary groups of data and that was used as classifier for the implementation of the ML method during this thesis work, as described in Chapter 2.

1.3 Support Vector Machine

Supervised ML is an attractive approach to data modelling which has been successfully used in many applications for image classification and, particularly, for functional brain image studies to support diagnosis by means of pattern recognition techniques (Scholkopf et al., 2003; Gunn et al., 1998).

Among supervised ML methods, SVM algorithm was first proposed by Vladimir N. Vapnik and Alexey Ya. Chervonenkis in 1963 as a (linear) binary classifier able to generate a predictive model for the discrimination of new samples.

Let us suppose that we have to study a particular kind of disease involving two different classes, the former containing images representing patients characterized by a particular feature (or pathology), the latter containing images representing patients who are not characterized by this feature (i.e., normal subjects). The aim of supervised predictive models, in general, and of SVM, in particular, is to find a mapping that, given some information about these subjects (for example, structural information encoded in MRI data), is able to predict the class to which new subjects belong.

Imagine we are given N observations, each one consisting of a pair: an input vector $x_n \in R^N$, n = 1, ...N and the corresponding target value $t_n \in \{\pm 1\}$, given to us

by a trusted source. Data are assumed to be *iid*, i.e., independently drawn from an unknown probability distribution P(x; t) and identically distributed. In our example, $x_n \in R^N$ might be a vector of pixel values (representing images of patients) and $t_n \in \{\pm 1\}$ would be +1 for images belonging to the feature class, -1 for images belonging to the normal class. The aim is to estimate a function that will correctly classify unseen examples (x; t), i.e., we want f(x) = t for samples (x; t) that were also generated from P(x; t). This problem reduces to the goal of separating the two classes by a function which is induced from available examples, in order to produce a classifier, based on input-output data training, which will work well on unseen samples. Let us now make the further assumption that the training data set is linearly separable in feature space and consider the example in Figure 1.3.1: here there are many possible linear classifiers that can separate the data, but there is only one that maximizes the margin (i.e., the distance between it and the nearest data point of each class). This classifier is termed the optimal separating hyper-plane. Intuitively, we would expect this boundary to generalize well as opposed to the other possible boundaries (Figure 1.3.1).

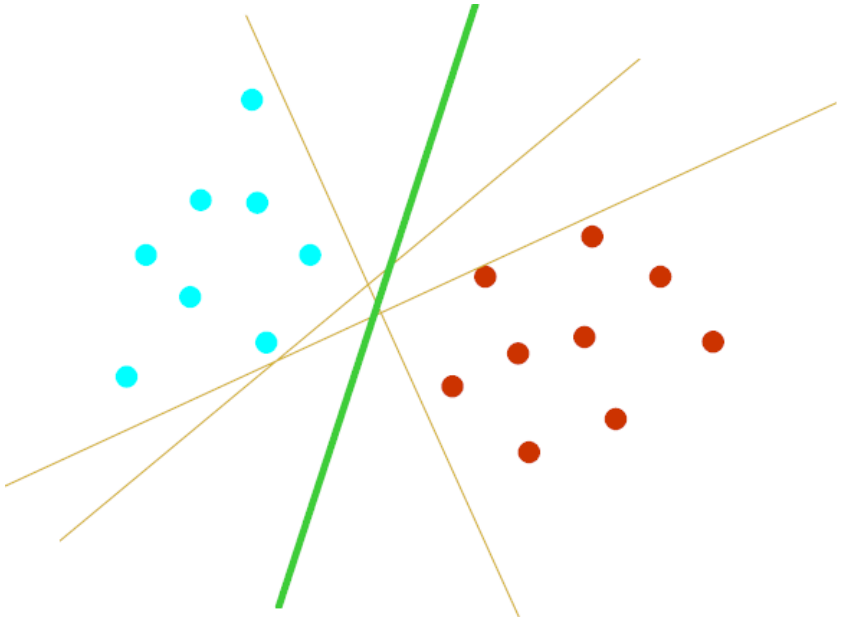


Figure 1.3.1 *Optimal separating hyper-plane*

If we consider the following class of hyper-planes (linear model)

$$y(x) = \omega^T \varphi(x) + b \qquad (1.3.1)$$

Where $\varphi(x)$ denotes a fixed feature-space transformation and the explicit parameter b is called bias parameter, then new data points x can be classified according to the sign of $y(x)$. In fact, as we assumed the training data set to be linearly

separable in feature space, by definition there exists at least one choice of the parameters w and b such that a function of the form 1.3.1 satisfies $y(x_n) > 0$ for points having $t_n = +1$ and $y(x_n) < 0$ for points having $t_n = -1$. Furthermore, $t_n \cdot y(x_n) > 0$ for all training data points. As we said before, there may exist many such solutions that separate the classes exactly. This problem can be solved by introducing the concept of the margin, which is defined to be the smallest distance between the decision boundary and any of the samples, as illustrated in Figure 1.3.2.

In SVM, the decision boundary is chosen to be the one for which the margin is maximized. The maximum margin solution can be motivated using computational learning theory, also known as statistical learning theory or VC (Vapnik-Chervonenkis) theory.

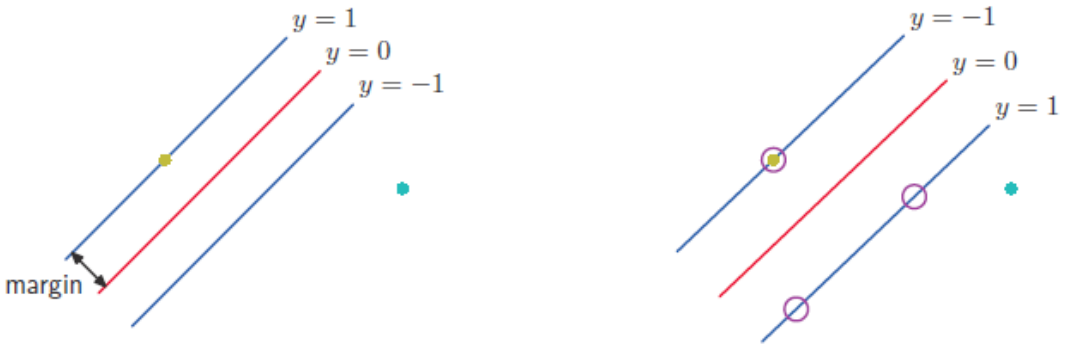


Figure 1.3.2 *Margin and support vectors*

It must be noted that the margin is defined as the perpendicular distance between the decision boundary and the closest of the data points, as shown on the left part of Figure 1.3.2. Maximizing the margin leads to a particular choice of decision boundary, as shown on the right part of Figure 1.3.2. The location of this boundary is determined by a subset of the data points, known as support vectors, which are indicated by the circles. The margin is defined as the perpendicular distance between the decision boundary and the closest of the data points, as shown on the left part of Figure 1.3.2. Maximizing the margin leads to a particular choice of decision boundary, as shown on the right of Figure 1.3.2. The location of this boundary is determined by a subset of the data points, known as support vectors, which are indicated by the circles.

In general, the perpendicular distance of a point x from a hyper-plane defined by y(x) = 0, where y(x) takes the form specified in equation 1.3.1, is given by $|y(x)|/\|w\|$. Furthermore, in this case we are only interested in solutions for which all data points are correctly classified, so that $t_n \cdot y(x_n) > 0$ for all n. Thus the distance of a point x_n to the decision surface is given by

$$\frac{t_n \cdot y(x_n)}{\|w\|} = \frac{t_n \cdot (\omega^T \varphi(x) + b)}{\|w\|} \qquad (1.3.2)$$

The margin is given by the perpendicular distance to the closest point x_n from the data set, and we wish to optimize the parameters w and b in order to maximize this distance. Thus the maximum margin solution is found by solving

$$argmax_{W,b}\left\{\frac{1}{\|w\|}min_n[t_n\cdot(\omega^T\varphi(x)+b)]\right\} \quad (1.3.3)$$

where we have taken the factor $\frac{1}{\|w\|}$ outside the optimization over n because w does not depend on n. Solving this optimization problem simply reduces to the requirement of minimizing $\|w\|^2$

$$argmax_{W,b}\left\{\frac{1}{2}\|w\|^2\right\} \quad (1.3.4)$$

(the factor of $\frac{1}{2}$ is included for later convenience), subject to the constraints

$$t_n\cdot(\omega^T\varphi(x)+b\ \geq 1, n=1,\dots N \quad (1.3.5)$$

It appears that the bias parameter b has disappeared from the optimization. However, it is determined implicitly via the constraints, because these require that changes to $\|w\|$ be compensated by changes to b. This constrained optimization problem is dealt with by introducing Lagrange multipliers $\mathrm{a_n} > 0$, with one multiplier $\mathrm{a_n}$ for each of the constraints in equation 1.3.5, giving the Lagrangian function

$$L(\omega,b,a)=\frac{1}{2}\|w\|^2-\sum_{n=1}^{N}a_n\{t_n\cdot(\omega^T\varphi(x)+b)-1\} \quad (1.3.6)$$

where $\mathrm{a}=(\mathrm{a_1},\dots\mathrm{a_N})^{\mathrm{T}}$. The negative sign in front of the Lagrange multiplier term is a consequence of the fact that minimization is performed with respect to w and b, and maximization is performed with respect to a. By setting the derivatives of $L(\omega,b,a)$ with respect to w and b equal to zero, the two following conditions can be obtained:

$$\omega=\sum_{n=1}^{N}a_n\cdot t_n\cdot\varphi(x) \quad (1.3.7)$$

and

$$\sum_{n=1}^{N}a_n\cdot t_n=0 \quad (1.3.8)$$

Eliminating w and b from $\mathrm{L}(\omega,\mathrm{b},\mathrm{a})$ using these conditions, then, gives the dual representation of the maximum margin problem in which we maximize

$$\tilde{L}(a)=\sum_{n=1}^{N}a_n-\frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N}a_m a_n t_m t_n k(x_n,x_m) \quad (1.3.9)$$

with respect to the constraints

$$a_n \geq 0, n = 1, \dots N \tag{1.3.10}$$

$$\sum_{n=1}^{N} a_n \cdot t_n = 0 \tag{1.3.11}$$

Here, the kernel function is defined by $k(x, x') = \varphi(x)^T \cdot \varphi(x)'$: this kernel formulation makes clear the role of the constraint that the kernel function $\mathrm{k(x, x')}$ be positive definite, because this ensures that the Lagrangian function $\tilde{L}(a)$ is bounded below, giving rise to a well-defined optimization problem. In order to classify new data points using the trained model, we evaluate the sign of y(x) defined in equation 1.3.1. This can be expressed in terms of the parameters an and the kernel function by substituting for ω using 1.3.7 to give the hyper-plane decision function

$$y(x) = \sum_{n=1}^{N} a_n \cdot t_n \cdot k(x, x_n) + b \tag{1.3.12}$$

It can be shown that a constrained optimization of this form satisfies the Karush-Kuhn-Tucker (KKT) conditions, which, in this case, require that the following three properties hold:

$$a_n \geq 0 \tag{1.3.13}$$

$$t_n \cdot y(x_n) - 1 \geq 0 \tag{1.3.14}$$

$$\mathrm{a_n\{t_n \cdot y(x_n) - 1\} = 0} \tag{1.3.15}$$

Thus for every data point, either $\mathrm{a_n = 0}$ or $\mathrm{t_n \cdot y(x_n) = 1}$. Any data point for which $\mathrm{a_n = 0}$ will not appear in the sum in 1.3.12 and, hence, they will play no role in making predictions for new data points. The remaining data points are called support vectors, and because they satisfy $\mathrm{t_n \cdot y(x_n) = 1}$, they correspond to points that lie on the maximum margin hyper-planes in feature space, as illustrated in Figure 1.3.2. This property is central to the practical applicability of SVM. Once the model is trained, a significant proportion of the data points can be discarded and only the support vectors retained. Having solved the quadratic programming problem and found a value for a, we can then determine the value of the threshold parameter b by noting that any support vector $\mathrm{x_n}$ satisfies $\mathrm{t_n \cdot y(x_n) = 1}$. Using 1.3.12, this gives

$$t_n\left(\sum_{m \in S} a_m t_m k(x_n, x_m) + b\right) = 1 \tag{1.3.16}$$

where S denotes the set of indices of the support vectors. Although we can solve this equation for b using an arbitrarily chosen support vector x_n, a numerically more stable solution is obtained by first multiplying through by $\mathrm{t_n}$, making use of $\mathrm{(t_n)^2 = 1}$, and, then, averaging these equations over all support vectors and solving for b to give

$$b = \frac{1}{N_S}\sum_{n\in S}(t_n - \sum_{m\in S} a_m t_m k(x_n, x_m)) \quad (1.3.17)$$

where $\mathrm{N_S}$ is the total number of support vectors.

In practice, a separating hyper-plane may not exist, e.g., if a high noise level causes a large overlap of the class-conditional distribution. In order to overcome this limitation, Cortes and Vapkin (1995) proposed a modified version of SVM introducing the idea of soft margin, which is useful when training classes cannot be sharply discriminated. Specifically, the soft margin approach allows to misclassify a fraction of training samples, while preserving the ability of the hyper-plane to maximizing its distance from the nearest samples of the two classes.

In order to allow for the possibility of violating 1.3.14, the general approach has to be modified so that data points are allowed to be on the *wrong side* of the margin boundary, but with a penalty that increases with the distance from that boundary. For the subsequent optimization problem, it is convenient to make this penalty a linear function of this distance; to do this, we introduce slack variables

$$\xi_n \geq 0, n = 1, \dots N \quad (1.3.18)$$

with one slack variable for each training data point. These are defined by $\xi_n = 0$ for data points that are on or inside the correct margin boundary and $\xi_n = |\mathrm{t_n} - \mathrm{y(x_n)}|$ for other points. Thus, a data point that is on the decision boundary $\mathrm{y(x_n)} = 0$ will have $\mathrm{\xi_n} = 1$ and points with $\mathrm{\xi_n} > 1$ will be misclassified. In this case, classification constraints relax to

$$t_n \cdot y(x_n) \geq 1 - \xi_n, n = 1, \dots N \quad (1.3.19)$$

in which the slack variables are constrained to satisfy $\xi_n \geq 0$. Data points for which $\xi_n = 0$ are correctly classified and are either on the margin or on the correct side of the margin. Points for which $0 < \xi_n \leq 1$ lie inside the margin, but on the correct side of the decision boundary, and those data points for which $\xi_n \geq 1$ lie on the wrong side of the decision boundary and are misclassified, as illustrated in Figure 1.3.3. This is sometimes described as relaxing the hard margin constraint to give a soft margin and it allows some of the training set data points to be misclassified. Note that while slack variables allow for overlapping class distributions, this framework is still sensitive to outliers, because the penalty for misclassification increases linearly with ξ. In order to realize a soft margin classifier, we now have to minimize the

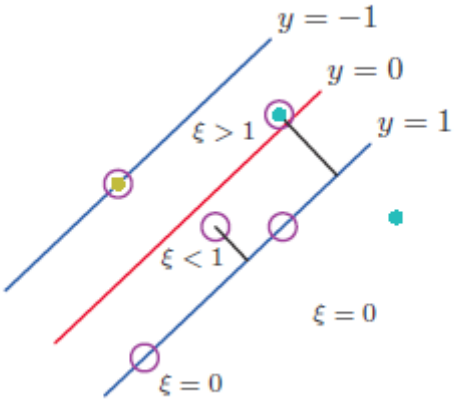


Figure 1.3.3 *Slack variables in support vector classifiers objective function. Data points with circles around them are support vectors.*

$$\frac{1}{2}\|w\|^2 + C\sum_{n=1}^{N}\xi_n \qquad (1.3.20)$$

subject to the constraints 1.3.18 and 1.3.19; here, the parameter C > 0 controls the trade-off between the slack variable penalty and the margin. Incorporating kernels and rewriting it in terms of Lagrange multipliers, this again leads to the problem of maximizing 1.3.9, subject to the constraints

$$0 \leq a_n \leq C \qquad (1.3.21)$$

$$\sum_{n=1}^{N} a_n t_n = 0 \qquad (1.3.22)$$

for $n = 1, \ldots N$ (constraints in 1.3.21 are known as box constraints).

The only difference from the separable case is the upper bound C on the Lagrange multipliers a_n. In this way, the influence of the individual patterns (which could be outliers) gets limited. As above, the solution shows that predictions for new data points are again made by using 1.3.12. The threshold b can be computed by exploiting the fact that for all support vectors x_n, with $0 \leq a_n \leq C$, the slack variable $\xi_n = 0$ and, hence, will satisfy 1.3.16. As before, a subset of the data points may have $a_n = 0$, in which case it does not contribute to the predictive model 1.3.12. The remaining data points constitute the support vectors.

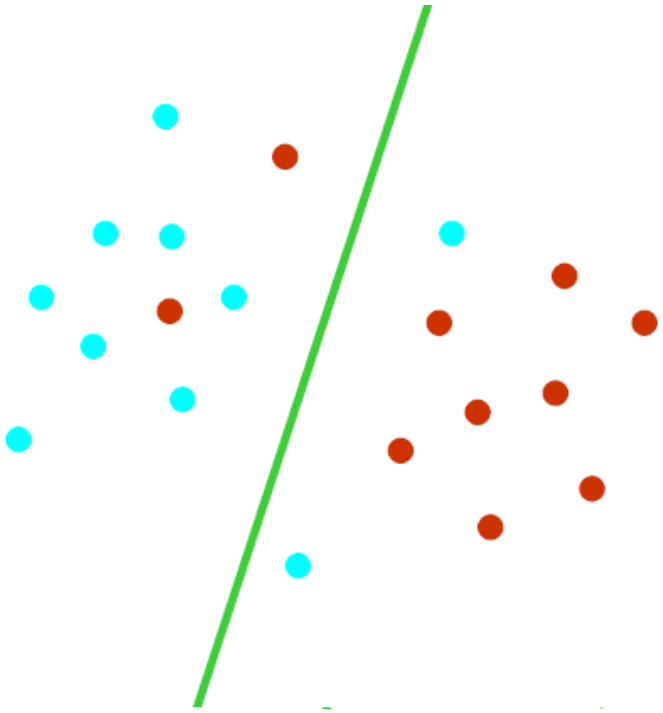


Figure 1.3.4 *Generalized optimal separating hyper-plane*

In practice, when dealing with a SVM-based classifier, the main parameter to be set for a study is the kernel function. The most popular kernels in literature are the linear, the polynomial and the Gaussian Radial Basis Function (RBF) kernels:

$$k(x_i, x_j) = (x_i \cdot x_j) \tag{1.3.23}$$

$$k(x_i, x_j) = (x_i \cdot x_j)^d \tag{1.3.24}$$

$$k(x_i, x_j) = (x_i \cdot x_j + 1)^d \tag{1.3.25}$$

$$k(x_i, x_j) = exp\left(-\gamma \|x_i - x_j\|^2\right) \qquad \gamma > 0 \tag{1.3.26}$$

where relation 1.3.23 is associated to linear kernels, relations 1.3.24-25 are associated to (homogeneous and inhomogeneous) polynomial kernels and relation 1.3.26 is associated to Gaussian RBF kernels, respectively.

As a general remark, it must be considered that non-linear kernels have been observed to be more flexible than linear ones in solving discrimination problems (Orrù 2012). This feature can bring to better classification performance in the training set, but usually lower generalization power (i.e., worse classification performance) in the testing set. This is referred to as an overfitting problem, that is, an overfitting of the generated predictive model to the dataset used for training the classifier. This drawback is particularly enhanced when the sample size is small with respect to the number of features. Because of this, in classification of neuroimaging data, in which

the sample size (number of patients) is usually smaller than the number of features (number of voxels), linear kernels should be preferred.

Both kernel function and related parameters can be (a priori) set to default values or they can be chosen through parameter optimization, which is performed using a grid-search approach. In the grid-search approach, a subset of the parameter space is defined, and classification performances are evaluated for each parameter configuration. When the whole subset of the parameter space has been explored, the configuration corresponding to the best classification performance is chosen as the optimal one. It is important to note that data used for parameter optimization and for training/testing the classifier should not be the same, because this could bring to a biased estimate of the generalization error.

SVM classification method has been used in imaging studies with good results: for example, Osuna et al. (1997) applied SVM classifiers to the problem of face recognition; Bonneville et al. (1998) studied SVM for improving the classification of human brain PET images with a resulting estimate error rate of 7.1%; Magnin et al. (2008) used this method for classification of Alzheimer's Disease (AD) using whole-brain anatomical MRI images reaching an overall mean accuracy of 94.5%. For a review of the use of SVM in structural neuroimaging studies for the diagnosis of AD, see the paper by Salvatore et al. (2015b).

Chapter 2.
MATERIALS AND METHODS

The aim of this chapter is to present and describe the ML method implemented and developed during my doctoral thesis. This method is based on Principal Components Analysis and Fisher's Discriminant Ratio for feature extraction and selection, respectively, while SVM is employed as classification algorithm.

The implementation of the algorithm is described in Section 2.1, while in Sections 2.2-2.5 the application of this method to four clinical problems is reported. In particular, these applications are useful as validation settings for the proposed method.

2.1 The Machine Learning Method

In this work, we aimed at performing automatic binary classification of MR images of patients by means of a ML approach. Specifically, in order to classify different groups of subjects by means of their T1-weighted structural MRIs, we implemented a ML classifier. The flow of the whole process (after the phase involving patient recruitment and data acquisition) consists of 3 steps: 1) image pre-processing, mainly devoted to the co-registration of data from different patients to the same reference system (that is, to a standard coordinate space); 2) feature extraction and feature selection, with the aim of extracting and choosing the most discriminative features for classification; 3) classification, i.e., estimation of the parameters defining the predictive model and prediction of the label of new (unseen) subjects. It is worth noting that all these steps are applied during both the training phase and the testing phase of the classifier separately. The training of the classifier is performed using a training set of subjects with known labels and it ends with the estimation of the parameters defining the predictive model. The testing phase of the classifier is performed using a testing set without known labels (unlabeled subjects) and it ends with the prediction of the label by means of the predictive model previously computed. The processing outlined above is schematically represented as a flowchart in Figure 2.1.1.

Among these three main phases of the proposed ML method, in the following Subsections 2.1.1 and 2.2.2 the two phases that I personally implemented, developed

and tested during my doctoral thesis will only be described: 1) feature extraction and selection; 2) classification (training and prediction).

Moreover, in Subsection 2.2.3 the implementation of the extraction of biomarkers is described, intended as the detection of the most important voxels for group discrimination and classification.

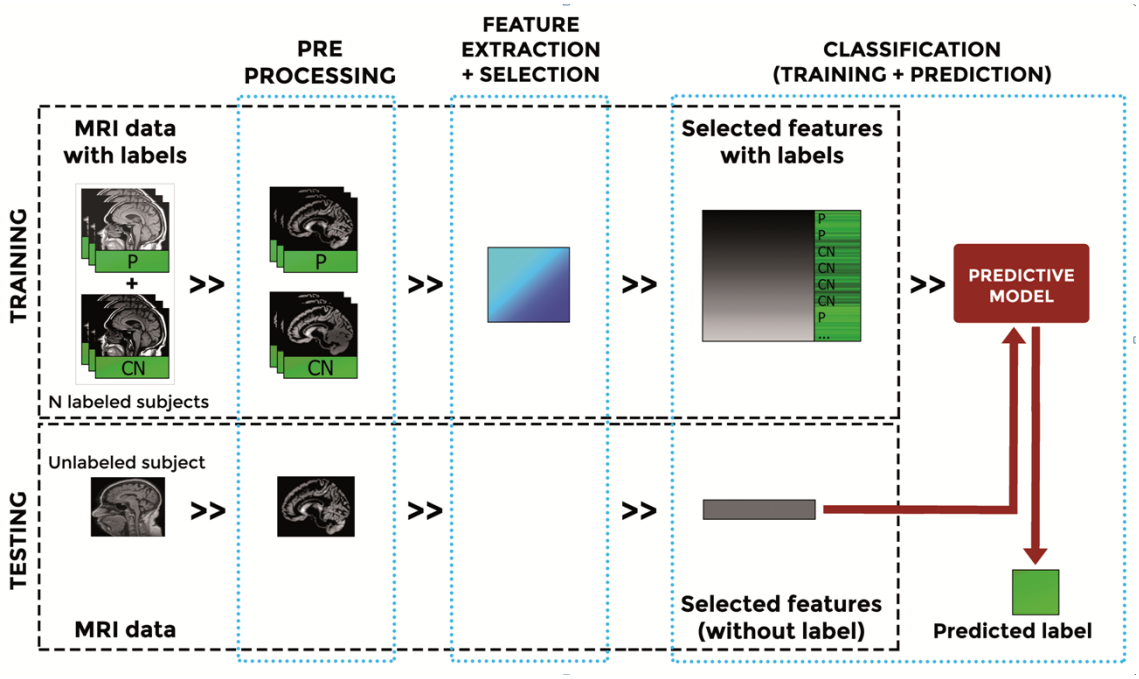


Fig 2.1.1 *Flowchart representing the main phases of the ML method, i.e., image pre-processing, feature extraction and selection, classification (for both training and testing/prediction)*

2.1.1 Feature extraction and selection

In order to identify the most discriminative features among groups, I implemented an automatic feature extraction and selection technique to be applied to MR images before training and classification with SVM.

Feature extraction is the operation of generating a new set of features as a function of the original input data. The new set of features should have the following characteristics with respect to the original input data: 1) reduced redundancy, by removing those features carrying no more information than the selected subset; 2) increased relevance (or reduced irrelevance), by removing those features that don't provide useful information for class discrimination (independently from the selected subset); 3) reduced dimensions, as a consequence of points 1) and 2).

This operation results in two main advantages. First, the reduction of the number of features to be handled by the classification algorithm usually enhances classification performance; this is only possible if the feature extraction technique is able to retain significant information for discrimination while discarding irrelevant and redundant information. Second, the reduction of the dimension of the feature set may result in a decrease of computational costs for the learning process. This point is

particularly important if we think that we are dealing with medical MR images, which are composed of a number of features (i.e., voxels) that are typically in the order of 10^6 or even more.

On the other side, the final purposes of feature selection are the same than those of feature extraction, e.g. dimensionality reduction or removal of redundant and irrelevant data. This is the reason why feature extraction and feature selection are often confused. However, there is one main difference between feature extraction and feature selection: the first generates a new set of features as a function of the original input data, while the second selects a subset of features from the original input data. As a consequence of this, feature selection doesn't suffer from the problem of making interpretation of the results difficult, which is typical of some feature extraction techniques (because in this last case we have to deal with a newly generated set of features that could show completely different characteristics with respect to the original one). On the contrary, feature selection could even aid the interpretation of the predictive model, because it restricts the dataset to a subset of discriminative features without operating any transformation on them.

Besides reducing computational costs and improving model interpretability, selection of a subset of relevant features from the original dataset before the generation of the predictive model also helps reducing overfitting problems, which brings to a better evaluation of the generalization ability of the classifier.

The technique of feature extraction and feature selection implemented in this work consists of two subsequent phases: 1) application of Principal Components Analysis (PCA), for feature extraction; 2) ranking of the extracted features by means of a Fisher's Discriminant Ratio (FDR) criterion, for feature selection. These two phases are described in detail below.

Principal Components Analysis

PCA (Habeck et al., 2008; Alvarez et al., 2009; Salas-Gonzalez et al., 2010) is a feature extraction technique that is based on two successive steps.

The first step consists in the application of an orthogonal transformation to a given dataset of (possibly) correlated variables, which results in a set of orthogonal (uncorrelated) variables. These variables are called principal components of the original dataset and they define the PCA subspace. Principal components of a given dataset are the eigenvectors of the covariance matrix of that dataset. The main characteristic of principal components is that they are sorted in descending order according to the proportion of explained variance of the input dataset, maintaining the constraint to be orthogonal with one another.

The second step consists in the projection of each variable of the original dataset onto the PCA subspace; this operation results in the reduction of the original set of observed variables into a smaller set of features called PCA coefficients (low dimension representation of the original samples). The total number of PCA coefficients is equal to the number of Principal Components extracted from the original dataset. The

number of Principal Components, in turn, is at most equal to the value of the smaller dimension of the original dataset – 1.

Mathematically, let us suppose X to be a dataset of 3D brain MR images X_i, with *i* ranging from 1 to *N* and *N* being the number of MR images (samples) in the dataset. If each image X_i is considered in the form of a vector of dimension *V* (in this case V could represent the total number of voxels in each image), than the dimension of *X* is equal to *NxV*. Under the condition that the dataset *X* has zero mean, PCA-space can be defined as the space spanned by the eigenvectors of the covariance matrix *C* of the dataset *X*:

$$C = X \cdot X^T \tag{2.1.1}$$

(if the dataset X had non-zero mean, then the average X_M could simply be subtracted from each image X_i to satisfy this condition).

As outlined above, application of PCA to a given dataset results in a number of principal components (i.e., eigenvectors) with non-null eigenvalues which is at most equal to the value of the lower dimension of the data matrix - 1. If N is the number of subjects in the dataset, and if N is smaller than the number of considered features, there will only be N-1 eigenvectors (principal components) with non-null eigenvalues. The other eigenvectors will have a zero eigenvalue associated, so that they will not be considered.

Fisher's Discriminant Ratio

While PCA returns coefficients containing information about the proportion of explained variance of the whole input dataset, FDR gives information about the class discriminatory power, i.e., it considers both input dataset and class labels.

In order to obtain FDR of each component, the following formula was applied to the PCA coefficients obtained from the feature extraction procedure:

$$FDR_k = \frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2} \tag{2.1.2}$$

where 1 and 2 represent the two classes involved (for example, pathological and normal condition), while μ_i and σ_i^2 are the mean and the variance of the i^{th} class, respectively. As it can be seen in formula 2.1.2, FDR is a measure of the separation of the two considered classes: for each PCA coefficient k, the higher is the value of FDR, the greater is the discriminatory power of that coefficient with respect to classes 1 and 2. In this way, FDR criterion was implemented in this work in order to rank PCA coefficients in a descending order according to their class discriminatory power.

2.1.2 Classification

The classification algorithm of the proposed ML method was based on SVM. A detailed description of SVM was given in Section 1.3. However, here we report the decision function used by SVM to predict the belonging class of a new (unseen) sample:

$$y(x) = \sum_{n=1}^{N} w_n \cdot t_n \cdot k(x, x_n) + b \qquad (2.1.3)$$

where y is the predicted class for sample x; *N* is the total number of samples included in the training set (both classes); w_n is a weight assigned during the training phase by SVM to each sample *n* of the training set; t_n is the label of the sample *n* of the training set; $k(x,x_n)$ is a kernel function; b is a threshold parameter.

In order to perform classification of features that were previously extracted and selected from MRI data through PCA and FDR, an SVM classifier was implemented. In the whole work, we used the Matlab platform to both implement and optimize the SVM classifier. Our code also included algorithms of the biolearning toolbox of Matlab. In all applications of the implemented ML algorithm described in this thesis, SVM classification was always performed using a linear kernel. The reasons for this are that 1) it is able to improve generalization ability and 2) it is the only kernel function that allows the computation of weights and, thus, the generation of voxel-based pattern distribution maps of brain structural differences (see subsection 2.1.3 below). Moreover, linear kernels allows to improve computational efficiency with respect to non-linear ones and this is a useful feature when dealing with big datasets (e.g. Application II in Section 2.3).

2.1.3 Computation of activation patterns and extraction of biomarkers

Extraction of biomarkers is one of the most important points in the proposed work, because it allows the computation of the most significant features (i.e., voxels) for group discrimination and the generation of image maps of voxel-based pattern distribution of brain structural differences between two given groups or conditions (e.g., pathological vs. healthy controls). This technique, coupled with a correct interpretation of these maps, could help to identify potential biomarkers for the diagnosis of a given pathology.

The most significant features for group discrimination can be computed during the training phase of the SVM classifier. As it can be seen from equation 2.1.3, during the training step, the SVM classifier assigns a specific weight w_n to each data-sample *n* of the training set. This weight reflects the importance of each given sample for the definition of the separating hyper-plane, i.e., the importance of each given sample for binary class separation. In our case, a weight can be associated to each MR image used to train the classifier. It is worth noting that this weight is non-null only for support vectors, while its sign is positive or negative whether the associated support vector belongs to the first or the second binary labeled class.

In order to extract the most important voxels for classification, the following procedure can be used: after training the SVM classifier, each sample of the training set is multiplied with the corresponding weight w_n assigned by SVM. The resulting weighted samples are then summed voxel-by-voxel, so that a vector of weights representing the importance of each voxel for SVM classification is generated (Kloppel et al., 2008; Focke et al., 2011). This vector can then be mapped onto a standard reference system (e.g. a standard stereotactic brain) to show the importance of each MR image voxel for SVM binary group discrimination.

However, as stressed by Haufe and colleagues (Haufe et al., 2014), this procedure is not sufficient in the case of backward models and, on the contrary, it could bring to misinterpretations in terms of the studied brain processes.

Backward models and activation patterns

Let us consider to have a source of information to be measured, and to have a number M of measurement channels. Data coming from channel m will be referred to as x_m. Let us make the further assumption that N data samples $x(n) = 1, \dots N$ are available (i.e., number of measurements), and that K latent factors $s(n)$ are *hidden* in the data. Finally, each latent factor $s_k(n)$ is linked to the corresponding target variable $y_k(n)$.

Backward models are models able to *extract* latent factors $\hat{\boldsymbol{s}}(n)$ as a function of the observed data. Indeed, the general purpose of backward models is the mapping of the data into a more informative representation, with the aim of showing the signals of interest as low-dimensional isolated components. In this sense, they are the opposite of forward models, which respect the functional dependency between factors and data (that is reversed in the case of backward models). Backward models are roughly used when there is the need to transform observations in order to find a representation (possibly at lower a dimension) in which they exhibit certain desired characteristics. In this sense, backward models roughly correspond to discriminative models in ML.

When dealing with the linear case, the transformation matrix $\boldsymbol{W} \in R^{M \times K}$, i.e., the matrix of filters, summarizes the mapping from observations to factors, so that the backward model can be written as follows:

$$W^T x(n) = \hat{s}(n) \qquad (2.1.4)$$

In supervised backward models, $\boldsymbol{W}$ can be chosen so that $\hat{\boldsymbol{s}}(n)$ approximates a target variable, where typically $K < M$. In this case, we can speak of *decoding*, a term used in analogy to the term of *encoding* used for supervised forward modeling (Naselaris et al., 2011).

Generally, one factor $\hat{\boldsymbol{s}}_{\boldsymbol{k}}(n)$ is extracted for each column $\boldsymbol{W}_{\boldsymbol{k}} \in R^M$ of $\boldsymbol{W}$, and that column of $\boldsymbol{W}$ is referred to as the extraction filter for the corresponding factor. As a consequence, an M-dimensional weight vector is associated to each factor, that is what happens also in forward modeling. However, in contrast to forward modeling, where the contribution of a factor to the measured data $\boldsymbol{x}(n)$ is obtained by

multiplying it with a latent factor $\hat{\boldsymbol{s}}_{\boldsymbol{k}}(n)$, here it is multiplied to the measured data $\boldsymbol{x}(n)$ with the aim of obtaining the latent factor $\hat{\boldsymbol{s}}_{\boldsymbol{k}}(n)$. As it can be understood, the filter vector $\boldsymbol{w}_{\boldsymbol{k}}$ does not coincide with the activation pattern $\boldsymbol{a}_{\boldsymbol{k}}$ of the same factor $\hat{\boldsymbol{s}}_{\boldsymbol{k}}(n)$, and there is no reason to think that these two entities should be similar or should have the same significance. Indeed, the interpretation of these entities is rather different:

> *When projecting observed data onto an extraction filter w_k, the result will be a latent component exhibiting certain desired properties (e.g., allow good classification or maximize the similarity to a target variable) (Haufe et al., 2014)*

In general, a filter is designed with the aim of amplifying the signal of interest and reducing (or suppression) all signals of no interest, which include noise, artifacts, redundant signals and signals coming from the source but regarding processes not under study (that is typically one of the main problems when dealing with brain processes). In this sense, extraction filters are functions of signal and noise components in the data, and it is the second task (i.e., reduction of signals of no interest) to be the responsible of the difference between filters and activation patterns. This is why the filter weights (e.g., the weights assigned by SVM as described above) do not allow one to draw conclusions about the features (e.g., brain voxels) in which the corresponding factor is expressed.

As written above, the general purpose of backward models is the mapping of the data into a more informative representation, with the aim of showing the signals of interest as low-dimensional isolated components.

However, the filters which compose the matrix W do not represent the expression of factors (from data) in the measured channels, as they only show how to combine information coming from different channels in order to extract factors from data. This is a problem when we are dealing with neurophysiological data to be interpreted: in this case, there is the need to compute activation patterns from extraction filters.

Let us now consider the case in which the number K of extracted linearly independent factors is exactly equal to M, i.e., the number of measurement channels. In this case, the matrix $\boldsymbol{W}$ of the extraction filters is square and invertible, so that multiplying Equation 2.1.4 with $\boldsymbol{W}^{-\boldsymbol{T}}$ (inverse and transpose), the following relation can be obtained:

$$x(n) = W^{-T}\hat{s}(n) \qquad (2.1.5)$$

in which the activation pattern $\boldsymbol{A}$ corresponds to $\boldsymbol{W}^{-\boldsymbol{T}}$. Under this simplifying condition, the backward model can be interpreted as a forward model, which means that the activation patterns $\boldsymbol{A}$ are simply given by the extraction filters $\boldsymbol{W}$.

On the other side, when dealing with the general case in which $K \leq M$, the extraction filters do not form an invertible matrix anymore. As a consequence, solving the problem of finding a pattern matrix $\boldsymbol{A}$ that indicates in which channels the

extracted factors are reflected is not trivial. We are looking for a linear forward model (corresponding to our backward model) which can be described by the following equation:

$$x(n) = A\hat{s}(n) + \epsilon(n) \tag{2.1.6}$$

where $\epsilon(n)$ is an M-dimensional noise vector. We assume (without loss of generality) that $E[\boldsymbol{x}(n)]_n = E[\hat{\boldsymbol{s}}(n)]_n = E[\boldsymbol{\epsilon}(n)]_n$, where $E[\cdot]_n$ denotes expectation over samples. Because of this, the associated covariance matrices are given by $\boldsymbol{\Sigma_x} = E[\boldsymbol{x}(n)\boldsymbol{x}(n)^T]_n$, $\boldsymbol{\Sigma_{\hat{s}}} = E[\hat{\boldsymbol{s}}(n)\hat{\boldsymbol{s}}(n)^T]_n$ and $\boldsymbol{\Sigma_\epsilon} = E[\boldsymbol{\epsilon}(n)\boldsymbol{\epsilon}(n)^T]_n$. We then make the further assumptions that the latent factors $\hat{\boldsymbol{s}}(n)$ are linearly independent, which implies that $rank\ (\boldsymbol{W}) = K$ (i.e., W must have full rank), and that the noise $\boldsymbol{\epsilon}(n)$ is uncorrelated with the latent factors $\hat{s}$, i.e. if $E[\epsilon(n)\hat{s}(n)^T]_n = 0$. As a consequence of these assumptions, the model described in Equation 2.1.6 can be called the *corresponding forward model* to the discriminative model of Equation 2.1.4.

This approach leads to the consequence *(from the paper by Haufe and colleagues)* that for any backward model

$$W^T x(n) = \hat{s}(n) \tag{2.1.7}$$

the corresponding forward model is unique, and its parameters are obtained by

$$A = \Sigma_x W {\Sigma_{\hat{s}}}^{-1} \tag{2.1.8}$$

The proof of this theorem can be found in the Appendix A of the paper by Haufe et al. (2014). What is important to note here is that A is the matrix whose columns are the activation patterns. Differently from associated filters which form the matrix W, activation patterns allow the correct interpretation of results, i.e., they actually represent the effect directions and strengths of the extracted latent factors in the measurement channels.

It is worth noting that, in practice, population covariances Σ_x, $\Sigma_{\hat{s}}$ and Σ_ϵ that appear in the equations above and on which these results depend can be substituted by their sample empirical values. This simplification allows the derivation of activation patterns in practice.

Finally, if the estimated factors $\hat{s}(n)$ are uncorrelated, the computation of activation patterns in practice is simplified. Specifically, activation patterns can be obtained multiplying the covariance Σ_x with the matrix W, i.e., as covariance between data and latent factors:

$$\boldsymbol{A} \propto \Sigma_x \boldsymbol{W} = Cov[\boldsymbol{x}(n), \hat{\boldsymbol{s}}(n)] \tag{2.1.9}$$

where

$$Cov[\boldsymbol{x}(n), \hat{\boldsymbol{s}}(n)] = \begin{bmatrix} Cov(x_1(n), \hat{s}_1(n)) & \cdots & Cov(x_1(n), \hat{s}_K(n)) \\ \vdots & \ddots & \vdots \\ Cov(x_M(n), \hat{s}_1(n)) & \cdots & Cov(x_M(n), \hat{s}_K(n)) \end{bmatrix} \quad (2.1.10)$$

This simplifying condition is verified for many backward models, including the method used during this thesis.

In this work, in order to ensure the correct interpretation of weights assigned by SVM, I adapted this method proposed for the computation of activation patterns in backward models by Haufe and colleagues to the extraction of the most important voxels for SVM classification. By mapping back the computed pattern from the PCA space to the MRI images space and by superimposing it onto a standard stereotactic brain for visualization purposes, a map of voxel-based pattern distribution of MR image differences among considered groups could be obtained, this map providing information about the localization of the most important areas used by SVM during the classification process.

2.2 Application I: Parkinson's Disease

Parkinson's disease (PD) is the second most common neurodegenerative disease affecting millions of people worldwide. Achieving an individual differential diagnosis still remains the primary goal in the clinical practice of PD, given the need of tailoring the best individual treatment. Diagnosis of PD is particularly prone to errors (Tolosa et al., 2006) because motor symptoms reported in PD can also be found in parkinsonian conditions such as Progressive Supranuclear Palsy (PSP), that is a parkinsonism with clinically similar symptoms to PD, but different treatment response (PSP patients are less responsive to treatment) and different prognosis (PSP has a more rapid disease progression). In particular, PSP results to be one of the most difficult parkinsonisms to be discriminated from PD, especially in early stages of diseases, when the typical clinical signs are not clearly evident yet (Litvan et al., 1996; Gelb et al., 1999).

The diagnosis of PSP is still based on the clinical history of the patient, while brain MRI is only used to exclude concomitant diseases. Moreover, given that images only undergo visual inspection, the resulting diagnostic accuracy (as well as sensitivity and specificity) is poor (Tolosa et al. 2006). In the last years, advanced processing techniques for neuroimaging have been developed. One of the objectives of these techniques is the identification of potential neuroimaging biomarkers able to improve the diagnostic confidence in the clinical practice. These efforts produced significant results (e.g. Shi et al., 2013), even if most studies only reported results at a group level, which is a limitation for a possible translation to an individual diagnosis in clinical settings.

In this first application of the proposed ML method to PD, we aimed to perform individual differential diagnosis of PD and PSP. Classification was made by means of structural T1-weighted MRI data collected from a cohort of PD and PSP patients and

healthy controls (CN). Moreover, in order to identify potential MR-related biomarkers for the early and differential diagnosis of PD, we also generated pattern distribution maps of brain structural differences reflecting the importance of each image voxel for the SVM classification, as explained in section 2.1.3.

2.2.1 Participants

In this study, 56 patients and 28 healthy control subjects were enrolled. The group of 56 patients consisted of 28 patients with a clinical diagnosis of PD (Gelb et al., 1999) and 28 patients with clinical diagnosis of probable or possible PSP (Litvan et al., 1996). Neurologists working in the field of movement disorders for more than 10 years examined all 84 subjects (patients and healthy controls). Age at onset of the disease, duration and severity of symptoms as assessed by the Unified Parkinson's Disease Rating Scale (UPDRS), and the Hoehn–Yahr (H&Y) stage were obtained and recorded for all patients. Moreover, the Mini Mental State Examination (MMSE) was used to assess the general cognitive status of each patient. The healthy control subjects had no history of neurologic or psychiatric diseases, with normal neurological examinations. The 28 healthy control subjects were matched for age to both patient groups.

2.2.2 MR images

MR images were acquired at the Institute of Neurology, University "Magna Graecia", Catanzaro. For each subject in both patients and healthy controls groups, we performed one brain structural MRI study. MRI scans were obtained by means of a 1.5-T Signa NV/I unit (GE Medical Systems, USA) and data were acquired using a 3D T1-weighted spoiled gradient echo sequence with the following parameters: TR = 15.2ms; TE = 6.7ms; flip angle = 15°; FOV = 24cm. Slice thickness was of 1.2mm and each slice had a resolution of 256 x 256 pixels. This process resulted in a T1-weighted 3D dataset for each subject. Motion artifacts were negligible for all scans by visual inspection.

Original datasets were converted from DICOM format to 3D NIfTI format using the dcm2nii tool included in the MRICron software (http://www.mccauslandcenter.sc.edu/mricro/mricron/). After this, the pre-processing procedure mainly consisted of 4 steps: 1) cropping and 2) re-orientation of converted images; 3) skull stripping, which was achieved using the BET tool of the FSL 4.1 software (Smith et al., 2004; Jenkinson et a., 2012); 4) spatial normalization to MNI space, which was performed by co-registration to the MNI template (MNI152_T1_1mm_brain) (Grabner et al., 2006) included in the FSL 4.1 software.

At the end of this pre-processing phase, images were imported into the Matlab platform using the 'Tools For NifTI And ANALYZE Image' toolbox (http://www.mathworks.com/matlabcentral/fileexchange/8797) and limited within a bounding box. Final whole-brain volumes consisted of 145x178x133 voxels. It is worth noting that in this first application of the proposed ML method to PD, no smoothing or segmentation were applied to MRI data.

2.2.3 The classifier

Classification was performed through the ML method described in Section 2.1 and visually outlined in Figure 2.1.1. The whole procedure is also describe in the paper by Salvatore et al. (2014). Specifically, feature extraction and feature selection were applied to pre-processed whole-brain MR images by means of PCA and FDR. After this step, extracted and ranked features were used as input to the SVM classifier.

SVM was used for the classification of PD vs. CN, PSP vs. CN and PSP vs. PD. Classification was performed using a linear kernel and for a number k of PCA coefficients ranging from 1 to the total number of extracted PCA coefficients.

2.2.4 Performance evaluation

In order to test the feasibility of this method for the automatic classification of PD and PSP, we used two different validation approaches: Leave-Out-Out (LOO) and half-splitting.

LOO is a particular case of cross validation in which the testing set only consists of one sample, while the training set is made up of all the remaining (N-1) samples of the whole dataset. By performing a number of rounds of LOO validation equal to the number of sample in the whole dataset, it is possible to test all samples in turn. The main strength of this validation technique is that it can be applied also when the dimension of the dataset is relatively small. Indeed, in LOO the largest portion of the dataset is reserved for the training of the classifier, which is useful when the number of samples is small in order to adequately compute the predictive model. On the other side, by reserving the largest part of the datasample for the training phase, the testing of the predictive value of the classifier is made only using a very small portion of the dataset, which can more easily bring to overoptimistic estimates of the predictive power (overfitting) with respect to other validation techniques. This is due to the fact that the classifier could be excessively tuned to the large amount of data used for the training.

Half-splitting, on the other side, is a validation approach in which half of the whole datasample is used for training the classifier and the remaining half is used for testing. One of the advantages of this procedure is the use of the same number of subjects for training and classification, which has been supposed to affect classification performance in terms of sensitivity and specificity of the classifier (see below). However, when half-splitting is performed using only one data partition, and especially when dealing with small heterogeneous datasets, half-splitting validation technique could be sensible to variability, bringing to significant variations in performance of classification.

For each validation method and for each of the three binary comparisons described above (PD vs. CN, PSP vs. CN, PSP vs. PD), accuracy, specificity and sensitivity of classification were computed as follows:

$$Accuracy_i = \frac{T_{CC}}{T} \qquad (2.2.1)$$

$$Specificity_i = \frac{X_{CC}}{X_{CC} + Y_{MC}} \quad (2.2.2)$$

$$Sensitivity_i = \frac{Y_{CC}}{Y_{CC} + X_{MC}} \quad (2.2.3)$$

where *I* indicates the number of employed principal components; T is the number of classified images; T_{CC} is the number of images that were correctly classified; X_{CC} (X_{MC}) is the number of images belonging to the first class that were correctly classified (misclassified); Y_{CC} (Y_{MC}) is the number of samples belonging to the second class that were correctly classified (misclassified). In particular, as it can be seen from equations 2.2.1-2.2.3, specificity represents the ability of the algorithm to correctly classify samples belonging to the first class (usually, the CN class), while sensitivity is the ability of the algorithm to correctly classify samples belonging to the second class (usually, the pathological class).

Once accuracy, specificity and sensitivity were computed, we calculated overall mean accuracy, specificity and sensitivity as mean values over a number of principal components ranging from 1 to PC, where PC is the whole number of extracted principal components. Furthermore, $accuracy_{>80}$, $specificity_{>80}$ and $sensitivity_{>80}$ were calculated as mean values over a range of principal components for which accuracy, specificity and sensitivity fell above 80%. The dependency of accuracy, specificity and sensitivity on the number of principal components was studied.

2.2.5 Diagnostic MR-related biomarkers

The most important voxels for SVM classification were computed as reported in section 2.1.3 using the whole number of subjects in the dataset (28 PD, 28 PSP and 28 Controls) and for each of the three binary comparisons described above (PD vs. CN, PSP vs. CN, PSP vs. PD).

In particular, for this first application of the implemented ML method to the diagnosis of PD, importance of voxels was computed without applying the correction needed for obtaining activation patterns in backward models.

2.3 Application II: Alzheimer's Disease

In this second application of the proposed ML method, I studied the early diagnosis of AD. AD is the first most common neurodegenerative disease, which affects millions of people worldwide (Martin et al., 2012). To date, individual diagnosis of AD in clinical practice is mainly based on neuropsychological assessment and clinical examinations, but definite diagnosis can only be achieved through post-mortem analysis (Blennow et al., 2006; Knopman et al., 2001). In order to aid clinicians to develop new treatments as well as to monitor their effectiveness, the identification of sensitive and specific markers of early AD are needed.

The study of normal and pathological ageing is gaining more and more interest given the increase in life expectancy and the prevalence of age-related cognitive disorders. Some of the main objectives of research in this field is represented by the possibility of detecting predictors of degenerative disorders for early and differential diagnosis, as well as improve efficacy of cognitive and pharmacological approaches in the treatment of these conditions. Indeed, considering the high costs on national healthcare systems given by exams and therapies of degenerative diseases, research with the aim of improving early and differential diagnosis is needed.

Clinical diagnostic criteria for AD, developed by the National Institute of Neurologic and Communicative Disorders and Stroke and the Alzheimer's Disease and Related Disorders Association (NINCDS-ADRDA), define cognitive impairment as necessary for the diagnosis of definite, probable or possible AD (McKhann et al., 1984). Moreover, the presence senile plaques and neurofibrillary tangles were afterwards introduced as neuropathological conditions for AD diagnosis (Hyman and Trojanowski, 1997).

In 2011, the National Institute on Aging-Alzheimer's Association developed revised diagnostic criteria for AD, which proposed additional features for the diagnostic process, such as measurement of cerebrospinal fluid (CSF), amyloid and tau, neurogenetic testing and neuronal injury biomarkers as measured by neuroimaging instruments, such as PET and MRI. In particular, PET provides measurements of metabolism/amyloid markers (Jagust et al., 2006; Fox and Schott, 2004), while MRI provides measurements of atrophic regions associated to AD, even before dementia is apparent (Sperling et al., 2011).

Among other methods, MRI has the advantage that it is a non-invasive technique. Because of this, the development of advanced MR image processing has been increasing in the past years in order to identify MR-related biomarkers useful for improving the accuracy of clinical diagnosis of AD. Among those studies focused on the identification of structural brain differences among AD patients and CN, the major part was based on a priori-defined regions of interest (ROIs) or on mass univariate image analysis methods, such as Voxel Based Morphometry (e.g. Busatto et al., 2003). These approaches, differently from multivariate techniques, are not able to detect spatially distributed patterns.

In order to overcome these limitations, in the last few years there has been a growing interest within the neuroimaging community in multivariate pattern analysis, including ML algorithms. Several studies have assessed the feasibility of these techniques in the automatic diagnosis of AD by means of brain structural MRI data (e.g. Davatzikos et al., 2008; Klöppel et al., 2008; Cuingnet et al., 2011; Hidalgo-Muñoz et al., 2014). Some studies showed good results also for the prediction of conversion to AD in the early phases of the disease (e.g. Tufail et al., 2012; Moradi et al., 2015). Among these, Kloppel et al. (2008) performed automatic ML classification by means of brain structural MR images, and they were also able to identify MR-related spatially-distributed biomarkers useful for the differential diagnosis of AD versus Fronto-Temporal Lobar Degeneration (FTD) and versus CN.

To date, definite early and differential diagnosis of AD by structural MRI data is one of the major challenges about neurodegenerative disorders, due to the difficulty of quantifying patterns of structural change during early stages of AD or during clinically normal stages (Davatzikos et al., 2008). Patients suffering from AD at a prodromal stage are often clinically classified as Mild Cognitive Impairment (MCI), but it is still difficult to predict MCI patients who will (MCIc) or will not (MCInc) convert to AD. The rate of conversion of MCI to AD has recently been estimated to be around 5-10% per year (Mitchell and Shiri-Feshki, 2009).

Given all these reasons, the identification of multivariate MR-related biomarkers for the diagnosis of AD and for the prediction of conversion of MCI to AD (MCIc vs. MCInc) is needed. Therefore, in this second application of the proposed ML method to AD, we aimed to perform early differential diagnosis of AD and to extract potential structural MR-related biomarkers for the prediction of conversion of MCI into AD.

2.3.1 Participants

Subjects included in this study were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). A total of 162 cognitively normal elderly controls (CN), 137 patients with diagnosis of AD, 76 patients with diagnosis of MCI who converted to AD within 18 months of follow-up (MCIc) and 134 patients with diagnosis of MCI who did not convert to AD within 18 months of follow-up (MCInc) were included in the study. MCI patients who had been followed less than 18 months were not considered. Demographic and clinical data (sex, age and mini-mental score) for each group are shown in Table 2.3.1 (see http://adni.loni.usc.edu/study-design/background-rationale/ for further description of groups). A total of 509 subjects from 41 different radiology centers were considered. Identification Numbers (IDs) of each subject involved in this study can be found in the supplementary material of the paper by Salvatore et al. (2015a). The ADNI was launched in 2003 by the National Institute on Aging (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the Food and Drug Administration (FDA), private pharmaceutical companies and non-profit organizations, as a $60 million, 5-year public private partnership. The primary goal of ADNI has been to test whether serial MR, PET, other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of MCI and early AD.

According to the ADNI inclusion criteria, enrolled subjects were all between 55 and 90 years of age and spoke either English or Spanish. Each subject was willing, able to perform all test procedures described in the protocol and had a study partner able to provide an independent evaluation of functioning. Inclusion criteria for CN were: Mini Mental State Examination (MMSE) scores between 24 and 30; Clinical Dementia Rating (CDR) (Mitchell and Shiri-Feshki, 2009) of zero; absence of depression, MCI and dementia. Inclusion criteria for MCI were: MMSE scores between 24 and 30; CDR of 0.5; objective memory loss, measured by education adjusted scores on Wechsler Memory Scale Logical Memory II (Wechsler, 1987), absence of significant levels of impairment in other cognitive domains; absence of dementia. Inclusion criteria for AD were: MMSE scores between 20 and 26; CDR of 0.5 or 1.0; NINCDS/ADRDA criteria for

probable AD (McKhann et al., 1984; Dubois et al., 2007). Detailed description of inclusion/exclusion criteria can be found in the ADNI protocol (http://www.adni-info.org/Scientists/ADNIStudyProcedures.aspx).

Table 2.3.1 Demographic and clinical data for the considered groups of participants

Group	#subjects	Age mean ± std [range]	Gender #Males / #Females	MMSE score mean ± std [range]	#centers
CN	162	76.3 ± 5.4 [60-90]	76 M / 86 F	29.2 ± 1.0 [25-30]	40
MCInc	134	74.5 ± 7.2 [58-88]	84 M / 50 F	27.2 ± 1.7 [24-30]	36
MCIc	76	74.8 ± 7.4 [55-88]	43 M / 33 F	26.5 ± 1.9 [23-30]	30
AD	137	76.0 ±7.3 [55-91]	67 M / 70 F	23.2 ± 2.0 [18-27]	39

2.3.2 MR images

For each of the 509 patients and healthy controls described above, we obtained one structural MR image from the ADNI database. We only considered T1-weighted structural MR images performed at 1.5 T. Moreover, in order to allow standardization of images from different sites and platforms, we only used images which had undergone: 1) geometry correction for gradient nonlinearity, by 3D gradwarp correction (Jovicich et al., 2006); and 2) intensity correction for non-uniformity, by B1 non-uniformity correction (Narayana et al., 1988). T1-weighted structural MR images of each subject were acquired according to the ADNI acquisition protocol (Jack et al., 2008). We used scans from the baseline visit (when available) or from the screening visit. According to the ADNI protocol, MR imaging examination was performed twice per visit. Scans were then rated by the ADNI investigators of the ADNI MR imaging quality control center at the Mayo Clinic on the basis of blurring/ghosting, flow artifact, intensity and homogeneity, signal-to-noise ratio (SNR), susceptibility artifacts, and gray-white/cerebrospinal fluid contrast (Jack et al., 2008). In this work, we used the image which was rated as the *best quality scan* for each subject. 3D MR images were downloaded from the ADNI dataset in 3D NIfTI format.

As for Application I, MR images underwent a pre-processing phase, which consisted of 4 steps: 1) cropping, 2) re-orientation, 3) skull-stripping and 4) spatial normalization to MNI standard space, performed by co-registration to the MNI template (MNI152_T1_1mm_brain) (Grabner et al., 2006). These procedures were performed through the VBM8 software package (Ashburner and Friston, 2000). At the end of this step, image size was of 121x145x121.

After this phase of spatial pre-processing, images underwent 1) segmentation into Gray Matter (GM) and White Matter (WM) tissue probability maps and 2) smoothing using an isotropic Gaussian kernel with Full Width at Half Maximum (FWHM) ranging from 2 to 12 mm^3 (step: 2mm^3).

The final set consisted of 21 images for each subject, depending on the considered tissue (whole-brain, GM, WM) and on the kernel used for smoothing (from 2 to 12 mm^3 FWHM or no smoothing).

2.3.3 The classifier

In order to perform classification, we applied the ML method described in Section 2.1 and depicted in Figure 2.1.1. In particular, feature extraction and feature selection were applied to pre-processed MR images (whole-brain, GM and WM) by means of PCA and FDR. After this step, extracted and ranked features were used as input to the SVM classifier.

SVM was used for the classification of AD vs. CN, MCIc vs. CN and MCIc vs. MCInc. Classification was performed using a linear kernel. The following parameters were varied: brain tissue (whole-brain, GM or WM images); smoothing (2-12 mm^3 FWHM or no smoothing); number of PCA coefficients (ranging from 1 to the total number of extracted PCA coefficients).

In order to show the impact of FDR-analysis on PCA coefficients, the explained variance was studied as a function of the number of considered principal components before and after sorting them in accordance to their FDR.

2.3.4 Optimization of classification and performance evaluation

In order to find the best parameter configuration for the classification of AD vs. CN, MCIc vs. CN and MCIc vs. MCInc, I performed an optimization of the considered parameters (i.e., brain tissue, kernel for smoothing and number of PCA coefficients to be used for classification). This optimization was performed through nested k-fold CV, which consists in splitting the original dataset into k subsets of (possibly) equal size and then performing an inner training-and-validation loop for parameter estimation and optimization and an outer test loop for performance evaluation. The inner training-and-validation loop is performed using k-1 subsets, while the outer test loop is performed using the k^{th} held-out subset, and by repeating the whole process k times it is possible to use all k subsets once for performance evaluation in the outer loop.

In particular, in this work, I used nested 20-fold CV. The inner training-and-validation loop was performed using 19/20 of the original dataset, and for each loop these 19/20 subsets were randomly split in half to perform training and validation on two independent datasets. The outer test loop was performed using the held-out 1/20 of the original dataset, for which a label (AD, MCIc, MCInc, CN) was predicted using the trained and optimized model and parameter set computed during the corresponding inner loop.

For each round of the inner loop, in order to chose the best parameter configuration, i.e., the optimal set of parameters (which brain tissue, which level of filtering, how many PCA coefficients), the algorithm aimed at minimizing the classification error (E) defined as follows:

$$E = 1 - Balanced\ Accuracy \quad (2.3.1)$$

$$Balanced\ Accuracy = \frac{1}{2}(Specificity + Sensitivity) \quad (2.3.2)$$

where specificity and sensitivity were computed as in 2.2.2-2.2.3. Optimization was performed as a function of the following parameters: 1) tissue map (whole-brain, GM and WM); 2) smoothing (FWHM = 2, 4, 6, 8, 10, 12 mm^3 or no smoothing); 3) number of PCA coefficients (from 1 to PC, where PC is the total number of extracted coefficients).

Balanced accuracy was computed as in 2.3.1 for each of the 20 separate rounds of the outer test loop. These results were then averaged across all 20 rounds in order to obtain the overall balanced accuracy.

This procedure (parameter optimization and accuracy evaluation) was performed for the binary classification of AD vs. CN, MCIc vs. CN and MCIc vs. MCInc.

2.3.5 Diagnostic MR-related biomarkers

In order to identify potential MR-related biomarkers for the early diagnosis of AD, extraction of voxel based pattern distribution of MR image differences between AD and CN, MCIc and CN, MCIc and MCInc was performed. Specifically, extraction was carried out according to the procedure described in section 2.1.3 for the computation of activation pattern in backward classification models. For each round of the inner training-and-validation loop and for the optimal set of parameters (minimum E) computed as described in subsection 2.3.4, a map of voxel-based pattern distribution of MR image differences among groups of subjects was generated. The resulting 20 maps (one for each round) were then averaged in order to obtain one final map, which was represented by a proper color scale and superimposed on a standard stereotactic brain for spatial localization. By performing this process for each of the three considered binary classifications, I finally obtained voxel-based pattern distribution maps of brain MR differences between AD and CN (AD diagnosis), MCIc and CN (early AD diagnosis), MCIc and MCInc (prognosis).

2.4 Application III: Eating Disorders

Eating disorders (ED) are a psychiatric condition with typical adolescent-onset that cause serious disturbances to everyday diet, such as eating extremely small amounts of food or severely overeating. It was demonstrated that the female gender is a potent risk factor for ED (Hoek and van Hoeken, 2003). However, it is still unknown how much this association can be attributed to biological rather than social factors (Treasure et al., 2010). ED presents various clinical phenotypes, the mot investigated among them being Anorexia Nervosa (AN) and Bulimia Nervosa (BN). AN is a mental disorder that leads to death in approximately 10% of cases (Nielsen, 2001). In accordance with the DSM-V criteria, diagnosis of AN is made following the criteria below: (a) Persistent restriction of energy intake relative to requirements leading to a

significantly low body; (b) Intense fear of gaining weight or becoming fat, even though underweight; (c) Disturbance in the way in which one's body weight or shape is experienced, undue influence of body weight or shape on self-evaluation, or denial of the seriousness of the current low body weight. On the other side, BN is characterized by frequent episodes of binge eating followed by inappropriate behaviors such as self-induced vomiting to avoid weight gain. In accordance with the DSM-V criteria, people with BN must exhibit binge eating and compensatory behaviors with a frequency of at least once a week.

To date, individual clinical diagnosis of ED is only based on a clinical interview complemented by physical, psychopathological and behavioural examinations aimed at assessing the existence of physical, emotional, behavioural and cognitive disturbances. However, a particular feature of ED diagnosis is that it is extremely unstable, with clinical features changing over time (i.e., weight normalization, Kong et al., 2014) and often switching from AN to BN (Fairburn and Harrison, 2003). For this reason, the identification of biomarkers useful for helping and improving not only early diagnosis, but especially treatment planning and monitoring of disease progression is strongly needed. This has lead to a considerable effort in developing advanced neuroimaging methods in the last few years (Titova et al., 2013; Gaudio and Quattrocchi, 2012; Van den Eynde and Treasure, 2009; Kaye et al., 2009; van Kuyck et al., 2009). Results from research in this field regarding AN reported global reductions of total gray and white matter (Seitz et al., 2014) and cortical thickness (King et al., 2014). As it was proposed in recent literature (Van den Eynde et al., 2012), AN patients are characterized by widespread brain abnormalities involving: (a) the mesolimbic regions (striatum, amygdale, hippocampus and cerebellum), (b) the dorsolateral prefrontal cortex, (c) the visual cortex and (d) the cerebellum. On the other side, BN are characterized the presence of a specific involvement of the reward neural system (ventral striatum, anterior cingulate cortex (ACC), orbitofrontal cortex (OFC), nucleus caudate), which could bring to the hypothesis that, as for addiction, during binge eating the feeling of satisfaction can be reached only after a greater consume of food with respect to the normal condition (Amianto et al., 2013; Marsh et al., 2015; Brooks et al., 2012).

Although significant results have been achieved, all these studies reported neurobiological abnormalities at a group level, comparing patients and controls, which has a consequently limited clinical translation power at the individual level. For this reason, the attention of the neuroimaging community has recently turned toward alternative kinds of analyses of neuroimaging data, such as multivariate ML techniques. However, to date there are no studies investigating the potential role of these methods in ED.

In this third application of the proposed ML method to ED, we aimed define reliable neuroimaging biomarkers useful to distinguish individual with diagnosis of ED patients from CN by means of structural T1-weighted MR images.

2.4.1 Participants

In this study, we considered a total of 103 patients with a first diagnosis of ED made by two specialized psychiatrists from 2011 to 2012. Clinical diagnosis was made through the Structured Clinical Interview for Diagnosis (SCID) for DSM-IV-TR. Inclusion criteria of patients to this research project were the following: (1) age range from 18 to 40 years; (2) female gender and (3) right-handedness. Exclusion criteria were: (1) neurological illness (such as Epilepsy or mental retardation); (2) Axis II disorders (using the SCID-II for DSM-IV-TR) to exclude comorbidity with personality disorders; (3) presence of brain lesions and history of cerebro-vascular disease, head trauma or hypertension; (4) psychotropic medication; (5) drug or alcohol abuse; (6) claustrophobia; (7) past recovery for ED symptoms or psychiatric disorders. After the evaluation of inclusion/exclusion criteria, 17 females with ED resulted to be eligible and were enrolled in this study. In particular, the group of 17 ED patients was composed of 11 patients with BN and 6 patients with AN (according to DSM-IV criteria for AN/BN restrictive-type). For all patients, duration of illness was rather short (mean duration: 16 ± 5 months).

The considered CN group to be compared with ED was composed of 81 healthy volunteers who were recruited by local advertisements. Inclusion criteria of CN to this research project were the following: 1) no previous histories of neurological or psychiatric diseases or abnormal brain MRIs; 2) being inside the normal range of the Italian version of Minnesota Multiphasic Personality Inventory-2 (MMPI-2) (Hathaway and McKinley, 1943). After the evaluation of inclusion/exclusion criteria, 17 CN females resulted to be eligible and were enrolled in this study. This group had similar demographical characteristics with respect to the enrolled ED patients. Potential confounding factors were also considered, including BMI, which was previously demonstrated to influence brain anatomy (Taki et al., 2008). In order to taking into account these factors, CN and ED patients were individually pair-matched according to their age, educational level and BMI (± 2). For further information, see the supplementary materials of the paper by Cerasa et al. (Cerasa et al., 2015).

In order to participate in this study, all participants gave written informed consent. This study was approved by the Local Ethical Committee according to the declaration of Helsinki.

2.4.2 Psychiatric assessment

All enrolled participants completed a battery of self-evaluation questionnaires. This battery included the following tests:

Eating Disorders Inventory-2 (EDI-2): a worldwide validated questionnaire that provides a multidimensional evaluation of the psychological characteristics of AN and BN (Garner, 1991);

Traumatic Experiences Checklist (TEC): a self-report measure addressing potentially traumatizing events (Nijenhuis et al., 2002). Different scores can be calculated including a cumulative score and scores for emotional neglect, emotional abuse, physical abuse, sexual harassment, sexual abuse and bodily threat from a person;

Dissociative Experiences Scale v. II (DES-II): a lifetime 28-item, self-rating questionnaire developed specifically as a screening instrument to identify subjects that are likely to have dissociative symptoms (Bernstein and Putnam, 1986);

Somatoform Dissociation Questionnaire-20 (SDQ-20): a self-rating scale developed to the investigated somatic component of dissociation. The SDQ-20 discriminates between dissociative and affective disorders (mood and anxiety disorders) and psychotic symptoms, but a cut-off score is not available (Nijenhuis et al., 1996);

Parental bonding instrument (PBI): PBI is a self-reporting scale with 25 items to rate paternal or maternal attitude during the first 16 years and has four items comprising care and overprotection factors (Parker et al., 1979). The Italian version of PBI was used to assess perceived parental rearing styles;

Eating attitude test-26 (EAT-26): a 26-item self-rated questionnaire for evaluating ED-related symptoms (Garner and Garfinkel, 1979). The results are presented as a total score (range, 0–78);

Body Image Dimensional Assessment (BIDA): a silhouette-based scale that starts from neutral figural stimuli and attributes a direct quantitative value to the subject's own current and ideal body image, the most sexually attractive figure and the most common figure of same-gender-and-age fellows (Segura-García et al., 2012);

Hamilton rating scale for anxiety (HAM-A): assessment of anxiety symptoms;

Beck Depression Inventory (BDI): definition of the depression status.

Statistical analysis was performed with STATISTICA Version 6.0 (www.statsoft.com). Assumptions for normality were tested for all continuous variables by using the Kolmogorov–Smirnov test. All variables were normally distributed, except for educational level. Then, Unpaired t-test and Mann–Whitney U-test were applied appropriately to assess potential differences between groups for all demographic and psychological variables. All statistical analyses had a 2-tailed alpha level of < 0.05 for defining significance.

2.4.2 MR images

Brain MRI scans were performed according to the routine protocol of the Institute of Neurology, University "Magna Graecia", Catanzaro. Structural MRI images were acquired by a 3T scanner with an 8-channel head coil (Discovery MR-750, GE, Milwaukee, WI, USA), using a 3D T1-weighted spoiled gradient echo sequence with the following parameters: TR: 9.2 ms, TE: 3.7 ms, flip angle 12°, voxel-size 1×1×1 mm^3. Subjects were positioned to lie comfortably in the scanner with a forehead-restraining strap and various foam pads to ensure head fixation. All acquired images were visually inspected by expert physicians and neuroradiologists to ensure that no signal artifacts were shown.

Original images were then imported into the Matlab platform (Matlab version R2011b, The MathWorks, Natick, MA) through the 'Tools For NIfTI And ANALYZE Image' toolbox (http://www.mathworks.com/matlabcentral/fileexchange/8797).

As for Application I, MR images underwent a pre-processing phase, which was performed by means of the VBM8 toolbox (Gaser et al., 1999) implemented in the

SPM8 software package (Ashburner and Friston, 2000). This pre-processing consisted of 4 steps: 1) cropping, 2) re-orientation, 3) skull-stripping and 4) spatial non-linear normalization to MNI standard space, performed by co-registration to the MNI template (MNI152_T1_1mm_brain) (Grabner et al., 2006). At the end of this step, image size was of 121x145x121.

After this phase of spatial pre-processing, images underwent smoothing using an isotropic Gaussian kernel with FWHM of 8 mm^3. Resulting non-modulated whole-brain images were used as input to the classifier.

We also automatically calculated the total gray matter (GM) and white matter (WM), as well as cerebrospinal fluid (CSF) volumes, by means of VBM8.

2.4.3 The classifier

Classification was performed through the ML method described in Section 2.1 and visually outlined in Figure 2.1.1. Specifically, pre-processed whole-brain MR images underwent feature extraction and selection by means of PCA and FDR. After this step, extracted and ranked features were used as input to the SVM classifier.

The classification of ED vs. CN was performed using a number k of PCA coefficients, where k runs from 1 to the total number of extracted PCA coefficients. As for the other applications, a linear kernel was used.

2.4.4 Performance evaluation

In order to evaluate the performance of the supervised machine-learning method, k-fold CV was performed with k = {10, 20}.

Accuracy, Specificity and Sensitivity were computed over the first k PCA coefficients, where k runs from 1 to the total number of extracted PCA coefficients, as in 2.2.1-2.2.3. It is worth noting that, for each round of CV, image pre-processing and feature extraction were performed separately on the training and the testing sets.

2.4.5 Diagnostic MR-related biomarkers

In order to identify potential MR-related biomarkers for the diagnosis of ED, extraction of voxel based pattern distribution of MR image differences between ED and CN was performed according to the procedure described in section 2.1.3 for the computation of activation pattern in backward classification models. The resulting map was normalized to a range between 0 and 1, expressed by a proper color scale and superimposed on a standard stereotactic brain for spatial localization.

2.5 Other applications

2.5.1 Autism Spectrum Disorder

Autism spectrum disorder (ASD) is a highly heterogeneous neurodevelopmental disorder with multiple causes, courses, and a wide range in symptom severity (Amaral et al., 2008).

Persistent deficits in social communication and interaction and presence of restricted, repetitive patterns of behavior, interests, or activities (DSM V, American Psychiatric Association, 2013) are the main features of ASD. However, motor impairments associated with ASD must not be forgotten, as they show high prevalence (79%) and can have a significant impact on quality of life and social development (Lai et al., 2014).

Motor abnormalities in ASD may have a very early onset in development (Teitelbaum et al., 1998, Brian et al., 2008) and they are often apparent over time (Fournier et al., 2010; Van Waelvelde et al., 2010) being a pervasive feature of the disorder. Recent studies have additionally provided proof for the specificity of motor impairments identified in high-functioning children with ASD compared to children with attention deficit/hyperactivity (ADHD) (Izawa et al., 2012; Ament et al., 2014) and to Typically Developing (TD) children matched by nonverbal IQ and receptive language (Whyatt & Craig, 2013). Overall, these findings suggest that motor abnormalities could be a consistent marker of ASD (Dowd et al., 2010). Different motor deficits have been reported in ASD, including hand movements such as reaching (e.g., Mari et al. 2003; Glazebrook et al. 2006; Forti et al. 2011), eye-hand coordination (e.g. Glazebrook et al., 2009; Crippa et al., 2013) and also anomalies in walking patterns (e.g., Rinehart et al. 2006; Nobile et al. 2011). The severity of motor deficits seems to be correlated with the degree of social withdrawal and the severity of symptoms (Freitag et al. 2007). Motor control has even been supposed to play a central role for social interaction and communication (Leary & Hill, 1996). Indeed, as highlighted by Minshew et al. (2004) studies on motor function could help elucidating the neurobiological basis and even improving the diagnostic definition of ASD.

Currently, clinical diagnosis of ASD is based on the clinical judgment of symptoms and on semistructured, play-based behavioral observations (Lord et al., 2000) including standardized interviews or questionnaires (e.g., Lord et al., 1994). Recently, there has been growing interest in the predictive value of neurobiological and behavioral measures in ASD in order to identify a well-defined phenotype of individuals and, possibly, to enable the perspective of computer-aided diagnosis of ASD.

In this Section, the application of the implemented ML method to ASD using kinematic data registered during a movement task is shown. The aim of this work, whose results were published in the paper by Crippa et al. (2015), is to show that the proposed classifier can be adapted to manage data from different modalities from MR images.

On the other side, from a clinical point of view, this study represents a proof-of-concept study to determine whether a simple upper-limb movement can be useful to

accurately classify low-functioning children with ASD (age ranging from 2 to 4 years). In particular, the ML method was applied for the classification of preschool children with ASD vs. typically developing (TD) children through kinematic data collected during a reach-grasp-drop task (described in subsection 2.5.1.2).

The choice of analyzing a simple motor task instead of more complex cognitive functioning-related tasks was made because the motor system can be more easily probed in low-functioning ASD children with respect to systems underlying complex cognitive functions.

Moreover, in addition to studying the predictive value of the ML method in classifying ASD vs. TD through this data, I applied feature selection coupled with *a posteriori* classification results to identify a limited set of kinematic features able to perform ASD vs. TD classification alone. These findings could suggest the hypothesis of a motor signature of autism.

2.5.1.1 Participants

In this study, we enrolled a total of 15 preschool-aged children with ASD and 15 TD children. The two considered groups were matched by mental age. IQ and mental age were assessed at the Child Psychopathology Unit of the Scientific Institute IRCCS Eugenio Medea by means of the Griffiths Mental Development Scales (Griffiths, 1970), as routinely made in the clinical practice with low-functioning children. In particular, this test is used since a poor score on the Griffiths scales at 1 and/or 2 years has been demonstrated to be a good predictor of impairment at school age (Barnett et al., 2004). All participants had normal or corrected-to-normal vision and were drug-naïve.

The participants in the ASD group were recruited over a period of 18 months. All participants in the clinical group had been previously diagnosed by a medical doctor specialized in child neuropsychiatry with expertise in autism. Diagnosis was made according with the criteria described in the Diagnostic and Statistical Manual of Mental Disorders-IV TR (American Psychiatric Association, 2000). The diagnoses were then confirmed independently by a child psychologist through direct observation and discussion with each child's parents. Seven children had been administered the Autism Diagnostic Observation Schedule (ADOS; Lord et al., 2000).

The participants in the control group were recruited by local pediatricians and from kindergartens to be mentally age-matched to the clinical sample from the normally developing population. The choice to match ASD and TD children by mental age was made following the assumption that it usually predicts the ability to understand task instructions, to use appropriate strategies and to inhibit inappropriate responses (Jarrold and Brock, 2004). TD children had no previous history of social/communicative disorders, developmental abnormalities, or medical disorders with central nervous system implications. All of the participants' legal guardians gave their informed written consent prior to the children's participation. The research was approved by the ethics board of our institute in accordance with the Declaration of Helsinki.

2.5.1.2 Procedure, apparatus and kinematic data acquisition

The procedure for the acquisition of movement data is describe below. The participants sat in front of a table of variable height, which was adjusted basing on the dimensions of the trunk of each child. The experimenter sat at the opposite side of the table, with one parent present in the room. At the beginning of each trial, the children's hands were resting at a set position 20 cm away from the ball support. The experimental task, that is visually described in Figure 2.5.1, roughly consisted of reaching a ball, grasping it and dropping it in a hole. Specifically, the participant grasped a rubber ball (6 cm diameter) that was placed over a support. This movement can be described as a reach-to-grasp movement, after which the participant drops the ball in a hole (7 cm diameter). The hole was located inside a see-through square box (21 cm high, 20 cm wide) and was large enough not to require fine movements.

For each participant, 10 trials were achieved: 5 consecutive trials on the left side (left hand) and 5 consecutive trials on the right side (right hand). The order of trial blocks was counterbalanced between participants. In order to visually illustrate the task (i.e., reach for the ball, grasp it and drop it in the hole) without any verbal cue to the participant, the experimenter performed the task. Practice trials, the number of which varied individually, were given to participants before recording in order to verify the children's understanding of the task. The participants were allowed to interrupt the experiment at will in order to rest. The experimental task was simple and interesting enough to ensure the full motivation and compliance of all participants across groups.

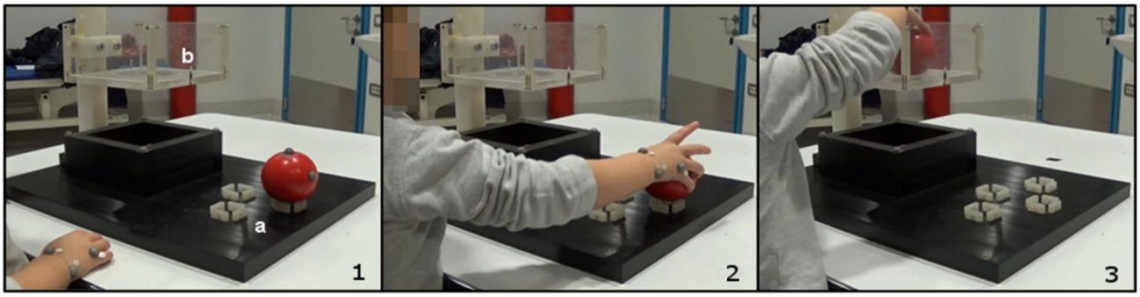


Fig 2.5.1 *The experimental task consisted of grasping a rubber ball (2) that was placed over a support (see 1, a); that is, a reach-to-grasp movement before they dropped it in a hole (3). The hole (1, c) was located inside a see-through square box (21 cm high, 20 cm wide) and was large enough not to require fine movements. The goal area is transparent to allow seeing through. 4 markers are placed on the basket under the goal area, 2 on the ball and 3 on each hand (attached to the ulnar and radial surfaces of the participant's wrist and to the hand dorsum on the 4th and 5th metacarpals)*

Kinematic data of the movements of each participant were registered and acquired through an optoelectronic system (The SMART D from BTS Bioengineering® Garbagnate Milanese, Italy). In order to track the movement of the participants during the task, passive markers (1 cm) were attached to the ulnar and radial surfaces of the participants' wrists and to the hand dorsum on the fourth and fifth metacarpals (as depicted in Figure 2.5.1). Moreover, 2 markers were placed on the ball and 4 markers were placed on the box edges under the goal area. Three-dimensional kinematic data

were then collected by 8 infrared-motion analysis cameras (4 per side at 2.5 m from participants) at 60 Hz, with a spatial accuracy less than 0.2 mm. All raw data were first preprocessed through Matlab (Mathworks® Natick, MA, USA); a fifth-order Butterworth, 8-Hz low-pass filter was applied, and movement segmentation and parameters estimation were computed.

After acquisition and preprocessing of kinematic data, the overall movement was divided into 2 sub-movements: *Sub-movement 1*—the movement necessary to reach the ball and place it on its support; *Sub-movement 2*— the movement to transport the ball from its support to the target box hole. For each of these sub-movements, statistics pertaining to a set of dependent measures was collected: (a) total movement duration (TD), (b) number of movement units[1] (MU), (c) peak velocity (PV), (d) time of PV from sub-movement onset (tPV), (e) peak acceleration (PA), (f) time of PA (tPA), (g) peak deceleration (PD), and (h) time of peak deceleration (tPD). Moreover, final movement accuracy was evaluated by the wrist inclination at the time of the ball drop (delta_WA), which was calculated as the angle between the palm and the vertical axis of the coordinate system (more precisely, the difference between WA at the end of the transport phase and at the time of peak deceleration). In sum, 17 kinematic measures were collected and these measures were used as input features to the ML classifier.

2.5.1.3 Data analysis

In order to compare the two groups of children on all kinematic measures with Group (ASD vs. TD) as a between-participant factor, an analysis of covariance (ANCOVA) was first performed (after checking that assumptions were not violated). IQ and chronological age were considered as between-participant covariates. The alpha level was set to .05 for all data analyses. Effect sizes for ANCOVA are reported using partial eta squared (η_p^2).

2.5.1.4 The classifier

In order to classify ASD vs. TD, we used a modified version of the ML method described in Section 2.1. In particular, in this case feature extraction (PCA) was not applied, as the number of features collected in this study (17) was already small enough to allow classification without the need of feature reduction. On the other side, feature selection was applied in order to select a subset of relevant features to be used for classification.

Because of this, the method used in this work involved 2 steps: 1) feature selection and 2) classification. Feature selection was performed by means of the FDR criterion in order to understand which of the collected kinematic features were the most important for the discrimination between ASD and TD. Ranked features were then used as input to SVM for the classification of ASD vs. TD.

[1] A movement unit is defined as an acceleration phase followed by a deceleration phase higher than 10mm/s, starting from the moment at which the increase or decrease in cumulative velocity is over 20mm/s (Von Hofsten, 1991; Thelen, Corbetta, & Spencer, 1996).

2.5.1.5 Performance evaluation and extraction of the most discriminant features

Performance of the classification algorithm was assessed for the classification of ASD vs. TD by using a LOO strategy, which was applied in this case as explained in 2.2.4 for Application I. A schematic description of the whole procedure is shown in Figure 2.5.2. It is worth noting that in this case, differently from applications I-III, the step of feature extraction is not performed.

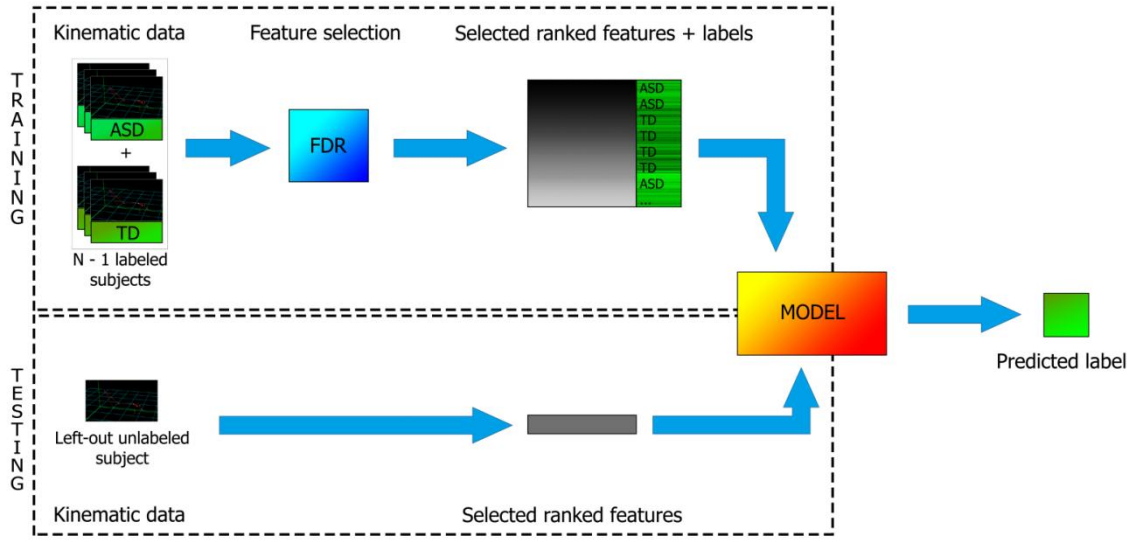


Fig 2.5.2 *Flowchart representing the main steps of the ML method applied to ASD, i.e., feature selection and classification (for both training and testing/prediction phases in the LOO process)*

In order to quantify the performance of the proposed classification algorithm, the accuracy, specificity, and sensitivity rates of the classifier when the first *i* selected features are used, were computed as in 2.2.1-2.2.3. As explained in subsection 2.2.4, accuracy of classification measures the rate of correctly classified samples in both positive (ASD) and negative (TD) classes. Specificity and sensitivity measure the rate of correctly classified samples in the positive (ASD) and in the negative (TD) class, respectively.

We then studied the dependency of accuracy, specificity, and sensitivity on the number *i* of selected features. The maximum values of accuracy, specificity, and sensitivity obtained in this way and referred to as maximum accuracy, specificity, and sensitivity, allowed the definition of the most discriminative features for the classification of ASD vs. TD.

Overall mean accuracy, specificity, and sensitivity rates were calculated as mean values of accuracy, specificity, and sensitivity over a number of features ranging from 1 to F, where F is the whole number of features (17).

2.6 Scalability, computational efficiency and use of cloud computing

In order to test the computational efficiency of the implemented ML algorithm for the classification of structural MR images, I used MRI data acquired in Application II to AD.

The operational time required by 1) the whole pre-processing and training of the classifier (including feature extraction and feature selection) and 2) the testing phase (including preprocessing and classification of the new dataset) was measured using the *tic and toc* functions implemented in Matlab. The operational time for the ML algorithm was measured while running on a system with 32 CPUs at 2.00 GhZ. The whole dataset for the 3 considered classification groups was used: AD vs. CN dataset (299 subjects), MCIc vs. CN dataset (238 subjects) and MCIc vs. MCInc dataset (210 subjects).

Chapter 3.
RESULTS AND DISCUSSION

3.1 Application I: Parkinson's Disease

3.1.1 Participants

In Table 3.1.1, demographic and clinical features about PD, PSP and CN groups included in this study are reported. As it can be seen, no significant differences were detected for demographical data among groups. However, as assessed by higher scores of UPDRS and H&Y scales, the group of patients with diagnosis of PSP showed a more rapid disease progression (with a fatal prognosis after few years) and a more critical clinical status in terms of motor disability with respect to PD patients.

Table 3.1.1 Demographic and clinical data of enrolled subjects.

Variables	CN	PD	PSP	*p* values
N°	28	28	28	
Gender (% males)	54%	54%	64%	
Age (years)	67.5 ± 7.1	68.2 ± 5	69.4 ± 5.7	n.s.
Disease Duration (years)	-	8.0 ± 4.8	3.0 ± 1.6	<0.001[II]
Age at Onset (years)	-	60.8 ± 5.6	67.2 ± 3.0	<0.001[II]
UPDRS-ME	-	24 (10 - 45)	34 (24 - 47)	<0.001[§]
H&Y	-	3 (1 - 4)	4 (3 - 5)	<0.001[§]
MMSE	26.5 ± 2.1	25.7± 1.9	24 ± 4.8	<0.001[‡]

Note: Data are given as mean values (SD) or median values (range) when appropriate. [‡]= One-way ANOVA. [II]= Unpaired t test. [§]= Mann-Whitney test. UPDRS-ME: Unified Parkinson Disease Rating Scale-Motor Examination in "off" phase (off medications overnight). H&Y: Hoehn-Yahr. MMSE: Mini Mental State Examination.

3.1.2 MR images

Considering MRI studies, all subjects had no evidence of vascular lesions as evaluated in Fluid Attenuated Inversion Recovery (FLAIR) and by T2-weighted MRI. Both PD and PSP patients showed no evident structural abnormalities, and CN subjects had normal MRI scanning.

3.1.3 The classifier

In Figure 3.1.1, the 1-st, 2-nd and 3-rd extracted PCA coefficients are shown, as a representative example, for the PSP versus PD (28 vs. 28) classification (in this case, the total number of extracted PCA coefficients was equal to 55).

As a representative example, in Figure 3.1.2 the optimal separating hyper-plane (i.e., the decision function) designed by SVM for PSP (28) versus PD (28) group separation is shown.

3.1.4 Performance evaluation

In Table 3.1.2, accuracy, specificity and sensitivity of the implemented ML algorithm are reported for the classification of PD versus CN, PSP versus CN and PSP versus PD, when using LOO validation approach. Overall Mean Accuracy, Specificity and Sensitivity rates were calculated over a number of principal components ranging from 1 to 55. Overall Mean Accuracy (Specificity/Sensitivity) were 85.8 (86.0/86.0), 89.1 (89.1/89.5) and 88.9 (88.5/89.5)% for the classification of PD versus CN, PSP versus CN and PSP versus PD, respectively.

Figure 3.1.3 shows the considered metrics (accuracy, specificity and sensitivity) as a function of the number of principal components when using LOO validation approach. Data are shown, as a representative example, for a number of principal components ranging from 1 to 55 and for the classification of PSP versus PD. As expected, accuracy, specificity and sensitivity rates increase with the number of considered principal components, reaching a plateau. This can be explained by the use of the FDR criterion, which allows noisy information to be contained only in a small fraction of eigenvectors (the last eigenvectors). For instance, without applying the FDR criterion, the relevant information would be contained in the first few eigenvectors, while the remaining eigenvectors would only contribute with noisy information (Alvarez et al., 2009).

The range of principal components for which accuracy, specificity and sensitivity fell above 80% (Accuracy>80, Specificity>80 and Sensitivity>80) was found to be from 30 to 52 components, when using LOO validation approach. In this range, for each of the three classifications, Accuracy>80 was > 83.9%, with mean Accuracy>80 = {92.7; 97.0; 98.2}% for the classification of PD versus CN, PSP versus CN and PSP versus PD, respectively. Specificity>80 was > 81.3%, with mean Specificity>80 = {92.3; 98.2; 98.8}% for the classification of PD versus CN, PSP versus CN and PSP versus PD, respectively. Sensitivity>80 was > 80.6%, with mean Sensitivity>80 = {93.4; 95.9; 97.8}% for the classification of PD versus CN, PSP versus CN and PSP versus PD, respectively.

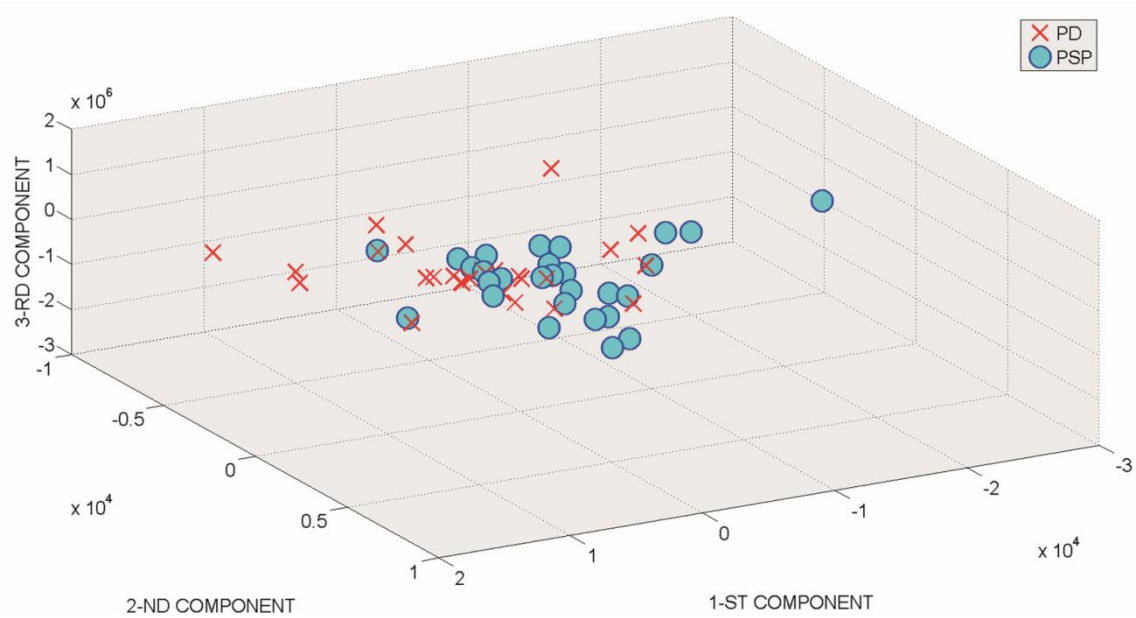


Figure 3.1.1 *PCA coefficients for the PSP versus PD binary labeled group (1st, 2nd and 3rd components).*

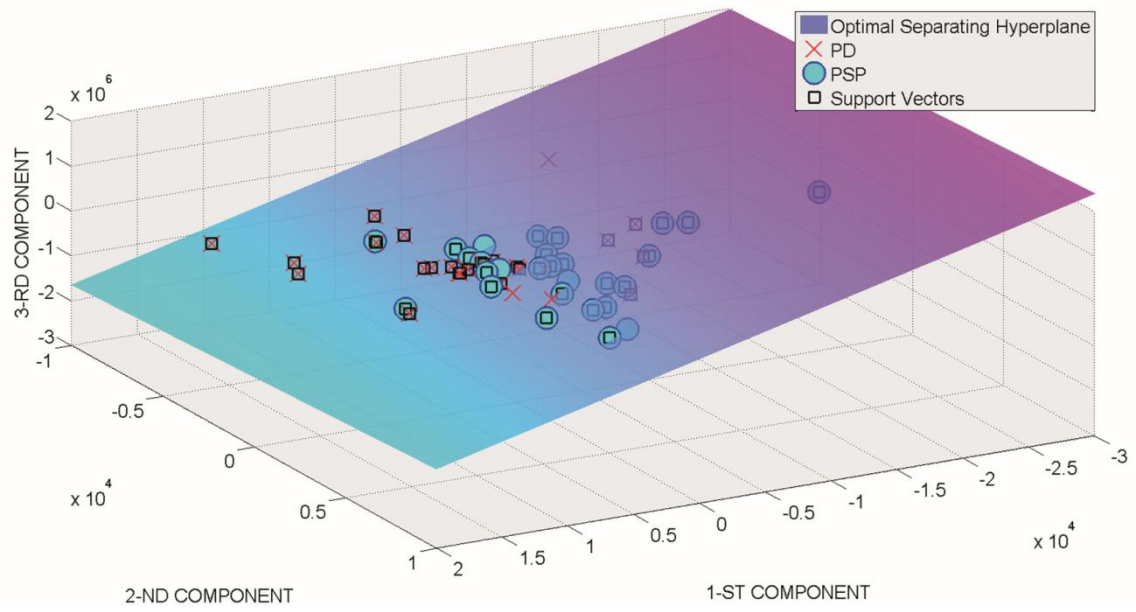


Figure 3.1.2 *Optimal separating hyper-plane for the PSP versus PD binary labeled group (1st, 2nd and 3rd components).*

Table 3.1.2 Accuracy, Specificity and Sensitivity rates of SVM using LOO validation.

	Overall Mean Accuracy (%)	Overall Mean Specificity (%)	Overall Mean Sensitivity (%)
	$Accuracy_{>80}$ (%) Mean (Min/Max)	$Specificity_{>80}$ (%) Mean (Min/Max)	$Sensitivity_{>80}$ (%) Mean (Min/Max)
PD vs. Controls	85.8	86.0	86.0
	92.7 (83.9/100)	92.3 (81.3/100)	93.4 (80.6/100)
PSP vs. Controls	89.1	89.1	89.5
	97.0 (92.9/100)	98.2 (92.9/100)	95.9 (90.0/100)
PSP vs. PD	88.9	88.5	89.5
	98.2 (94.6/100)	98.8 (96.3/100)	97.8 (93.1/100)

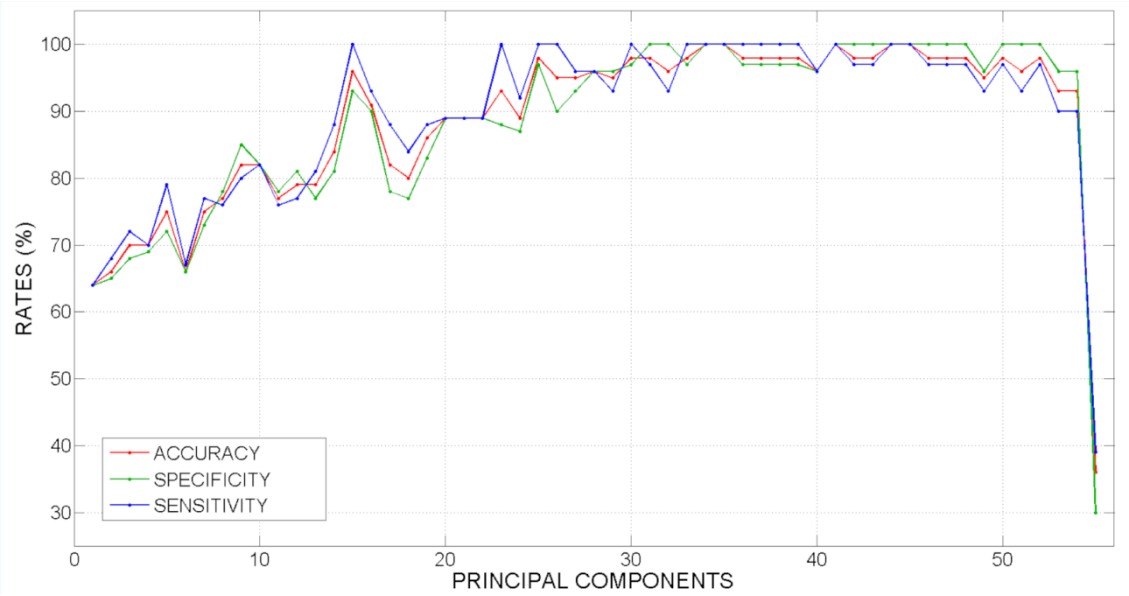


Figure 3.1.3 *Accuracy, Specificity and Sensitivity rates (%) of SVM versus Number of PCA components in LOO validation.*

In Table 3.1.3, accuracy, specificity and sensitivity of the implemented ML algorithm are reported for the classification of PD versus CN, PSP versus CN and PSP versus PD, when using half-splitting validation approach. Overall Mean Accuracy, Specificity and Sensitivity rates were calculated over a number of principal components ranging from 1 to 28. Overall Mean Accuracy (Specificity/Sensitivity) were 83.2 (81.9/85.4), 86.2 (92.1/82.9) and 84.7 (87.5/83.8)% for the classification of PD versus CN, PSP versus CN and PSP versus PD, respectively.

The range of principal components for which accuracy, specificity and sensitivity fell above 80% (Accuracy>80, Specificity>80 and Sensitivity>80) was found to be from 16 to 26 components when using half-splitting validation approach. In this range, for each of the three classifications, Accuracy>80 was > 85.7%, with mean Accuracy>80 = {93.5; 92.2; 92.2}% for the classification of PD versus CN, PSP versus CN and PSP versus PD, respectively. Specificity>80 was > 82.4%, with mean Specificity>80 = {90.6; 92.5;

91.3}% for the classification of PD versus CN, PSP versus CN and PSP versus PD, respectively. Sensitivity>80 was > 85.7%, with mean Sensitivity>80 = {97.4; 92.4; 94.4}% for the classification of PD versus CN, PSP versus CN and PSP versus PD, respectively.

Table 3.1.3 Accuracy, Specificity and Sensitivity rates of SVM using half-splitting validation.

	Overall Mean Accuracy (%)	Overall Mean Specificity (%)	Overall Mean Sensitivity (%)
	$Accuracy_{>80}$ (%) Mean (Min/Max)	$Specificity_{>80}$ (%) Mean (Min/Max)	$Sensitivity_{>80}$ (%) Mean (Min/Max)
PD vs. Controls	83.2	81.9	85.4
	93.5 (89.3/100)	90.6 (82.4/100)	97.4 (92.3/100)
PSP vs. Controls	86.2	92.1	82.9
	92.2 (85.7/96.4)	92.5 (85.7/100)	92.4 (85.7/100)
PSP vs. PD	84.7	87.5	83.8
	92.2 (89.3/96.4)	91.3 (82.4/100)	94.4 (86.7/100)

Classification performance reported above was consistent with previous studies, such as studies using manual morphological metrics (e.g., Massey et al., 2013; Quattrone et al., 2008; Oba et al., 2005; Schulz et al., 1999) or studies applying SVM to MRI data (Focke et al., 2011; Haller et al., 2012).

Interestingly, the most difficult task in literature among those explored in this section seems to be the discrimination between PD and CN. For example, in the study by Focke and colleagues (2011), the classification performance for the diagnosis of PD (PD vs. CN) was reported to be only marginally better than chance. Haller et al. (2012; 2013) applied SVM to Diffusion Tensor Imaging (DTI) and susceptibility-weighted images, obtaining classification performances comparable to those reported in this work for the diagnosis of PD. However, in these later works the clinical classification was made using a cohort of parkinsonisms that was heterogeneous in the group of patients and without considering CN subjects.

3.1.5 Diagnostic MR-related biomarkers

Figure 3.1.4 shows maps of voxel-based pattern distribution displaying the importance of each voxel for the SVM classification of PD versus CN (28 vs. 28), PSP versus CN (28 vs. 28) and PSP versus PD (28 vs. 28). The pattern is expressed according to the color scales reported in the figure.

When considering the direct classification of PD versus CN and PSP versus CN by SVM, the pattern distribution of the most important voxels for group separation was similar for both PD and PSP patients, as shown in the upper part of Figure 3.1.4. However, it is worth noting that, only for the classification of PD versus CN, SVM

classification revealed significant voxels within the medial part of the midbrain (encompassing the substantia nigra) and the caudal part of the pons.

For the differential classification of PSP versus PD by SVM, the most important voxels were found to be localized in the midbrain, pons, corpus callosum and thalamus, as shown in the bottom part of Figure 3.1.4.

Overall, brain regions detected as the most important to perform classification of PD versus PSP (i.e., midbrain, pons, corpus callosum and thalamus) are highly consistent with typical neuropathological (Steele et al., 1964) and imaging findings described in patients with PSP (Shi et al., 2013; Messina et al., 2012). Indeed, the volumetric atrophy of the brainstem plays a pivotal role in PSP, representing a hallmark of this pathology (Oba et al., 2005). Moreover, recent studies aiming to quantify white matter pathology in PSP by means of Diffusion Tensor Imaging (DTI), highlighted the involvement of the corpus callosum. The relevance of this finding is given by the fact that corpus callosum is the largest white matter tract in the brain, enabling interhemispheric communication, particularly with respect to motor coordination, and it is one of the damaged tracts in PSP (Knake et al., 2010; Canu et al., 2011).

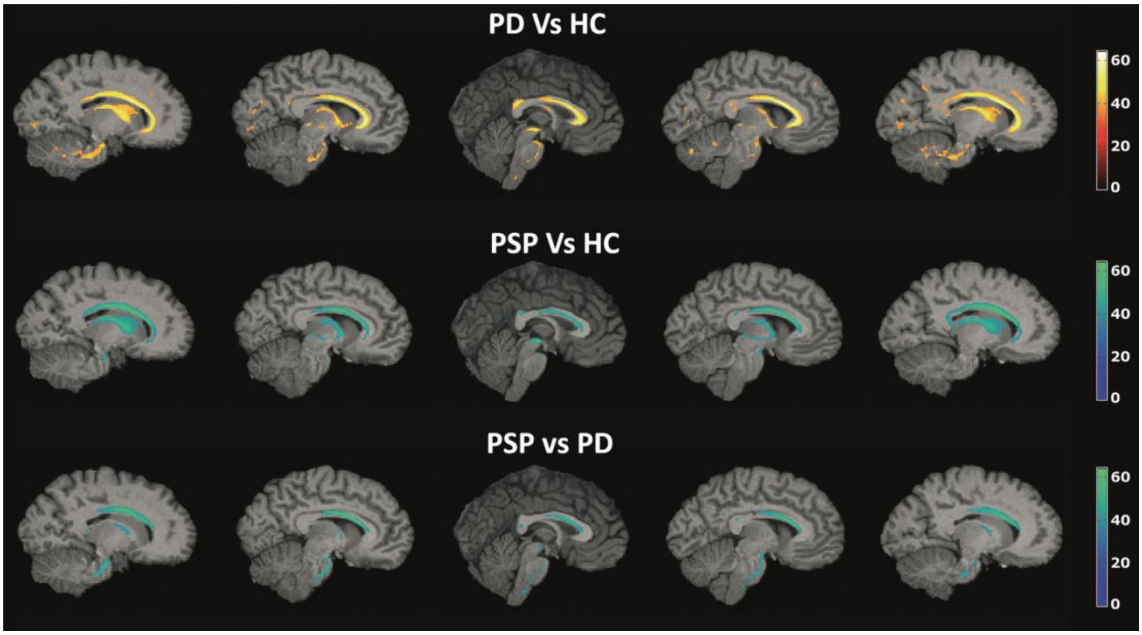


Figure 3.1.4 *Maps of voxel-based pattern distribution of brain structural differences (sagittal view, threshold = 60%). The importance of each voxel in the SVM classification is expressed according to the color scale. A, PD versus CN; B, PSP versus CN; C, PSP versus PD; PD: Parkinson's Disease; PSP: Progressive Supranuclear Palsy; CN: healthy controls.*

For the classification of PD versus CN by SVM, a more widespread pattern was highlighted by our analysis, involving several cerebral regions. This result seems to confirm the hypothesis that PD is a more heterogeneous clinical phenotype, characterized by several neural pathologies which may be topographically separated. Among the most important areas for the classification of PD versus CN highlighted by

our analysis, the medial part of the midbrain (encompassing the substantia nigra) and the caudal part of the pons are consistent with findings reported in the Braak's neuroanatomical model of the PD (Braak et al. 2003). Post-mortem studies by Braak and colleagues (see Del Tredici et al., 2002), based on the analysis of Lewy neuritis and Lewy bodies accumulation (that is, a proteic hallmark of PD), showed that various cerebral structures are damaged in a consistent and repeated pattern before substantia nigra. In a six-stage model of disease proposed by Braak et al. (2004), PD would begin (stage 1) in the medulla oblongata and in the olfactory bulb, and it would progress (stage 2) in a caudo-rostral pattern. Substantia nigra would be affected only during the onset of the motor symptoms (stage 3), which are often detected in the first neurological visit of the patient. Nevertheless, to date only one study (Jubault et al., 2009) about structural neuroimaging investigating the neural basis of PD did describe the occurrence of anatomical changes in this region. It is worth noting that most of the studies that have investigated structural abnormalities in PD were performed using standard mass-univariate analytical methods. The main advantage of SVM in this point is that it is able to take into account inter-regional correlations, being sensitive to subtle and spatially distributed differences in this way. This represents the optimal framework for investigating neurological diseases, in which a distributed network of regions is affected.

3.2 Application II: Alzheimer's Disease

3.2.1 Participants

No significant differences were found for age (Student's t-test with significance level at 0.05) and gender (Pearson's chi-square test with significance level at 0.05) among the groups of participants. On the other side, for MMSE scores, significant differences were found between patients (AD, MCIc) and CN using a Student's t-test with $p<0.0001$. This is consistent with previous studies that considered the same groups of ADNI subjects (Cuingnet et al., 2012).

3.2.2 MR images

In Figure 3.2.1, the results of co-registration to the MNI template and segmentation into GM and WM tissue probability maps are shown for a representative MR image of a patient with MCIc. Sagittal view of the original MR volume (A), same slice co-registered to the MNI space (B), and same slice segmented into GM (C) and WM (D) tissue probability maps are shown.

Overall, co-registration and segmentation were correctly performed for all MR images involved in this study, with no artifacts at visual inspection.

3.2.3 The classifier

In Figure 3.2.2, PCA coefficients resulting from feature extraction and feature selection are shown as a representative example for the classification of AD vs. CN. 1st,

2nd and 3rd components are shown when using GM tissue probability map and an isotropic Gaussian kernel with 10 mm^3 FWHM for smoothing. In this case, the number of the extracted PC was 141.

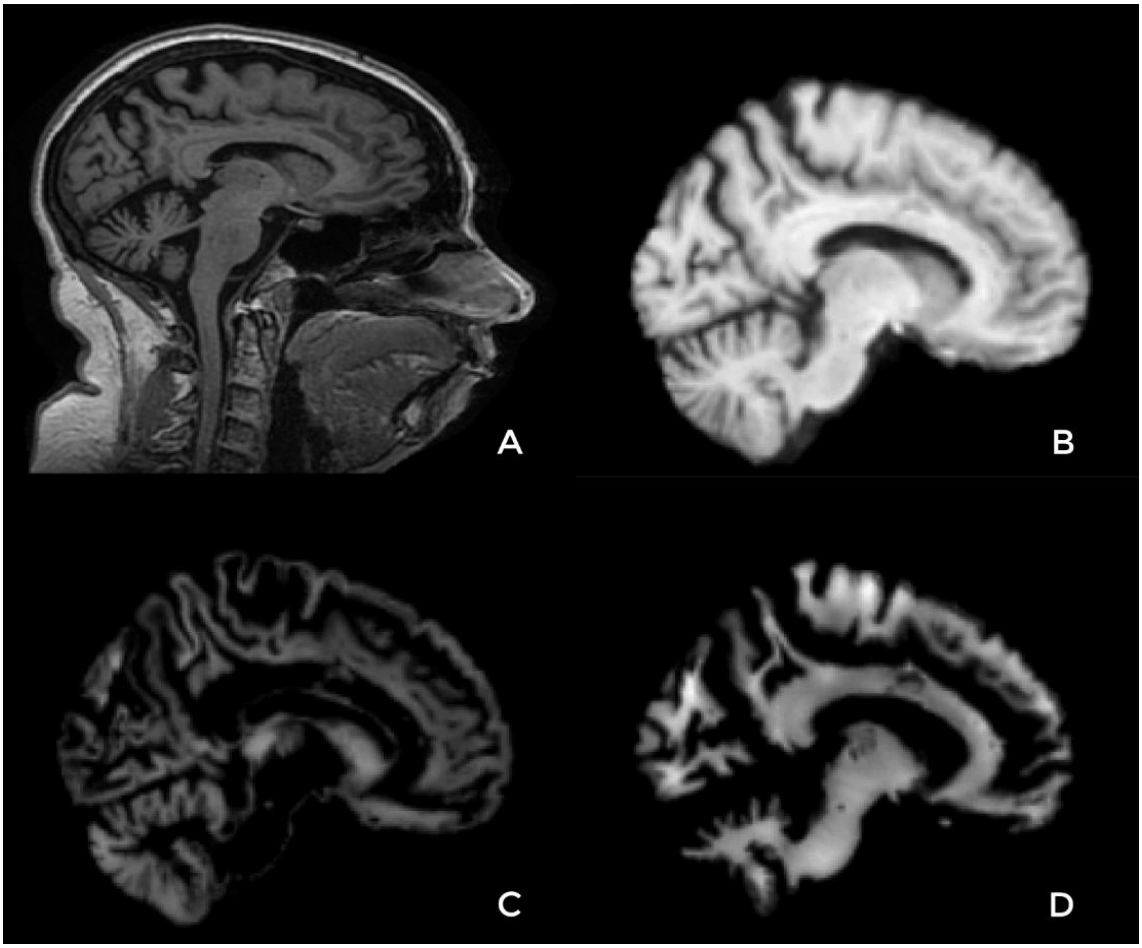


Figure 3.2.1 *Sagittal image of a MR scan from a MCIc patient: (A) original image; (B) same slice, deskulled and co-registered to the MNI space; same slice, segmented into Gray Matter (GM) (C) and into White Matter (WM) (D).*

The explained variance as a function of the number of considered principal components is shown in Figures 3.2.3-3.2.4, as representative example. Figure 3.3.3 shows the explained variance before sorting principal components according to their FDR, while Figure 3.3.4 shows the explained variance after sorting principal components according to their FDR. Plots are shown for the classification of AD versus CN, MCIc versus CN and MCIc versus MCInc when using GM tissue probability maps and no smoothing. As it can be seen looking at these two figures, the trend of explained variance as a function of the number of considered PCs was modified by the application of FDR-analysis. In particular, FDR ranking allowed the most discriminative information for class separation to be contained in the first few principal components. This is shown, for example, by the step in the explained variance in correspondence with a low number of components for the classification of both AD versus CN and MCIc versus CN in Figure 3.2.4.

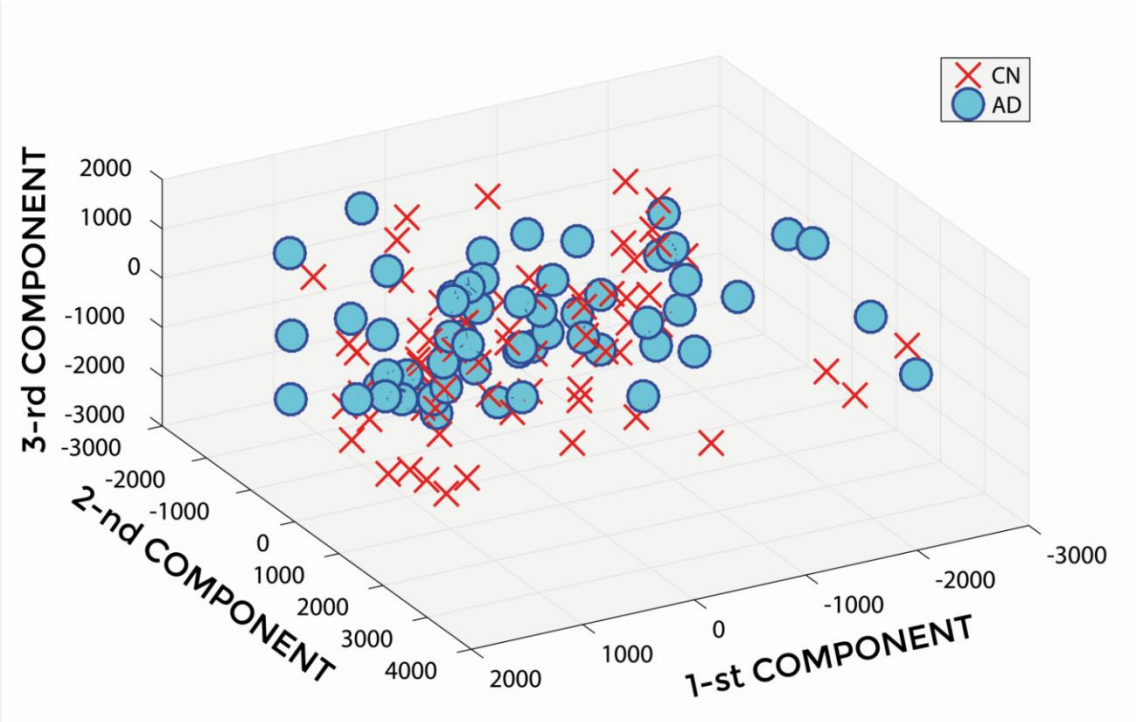


Figure 3.2.2 *PCA coefficients for the comparison between AD (o symbol) and CN (x symbol) when using GM tissue probability map and an isotropic Gaussian kernel with 10 mm3 FWHM for smoothing. 1st, 2nd and 3rd components are shown.*

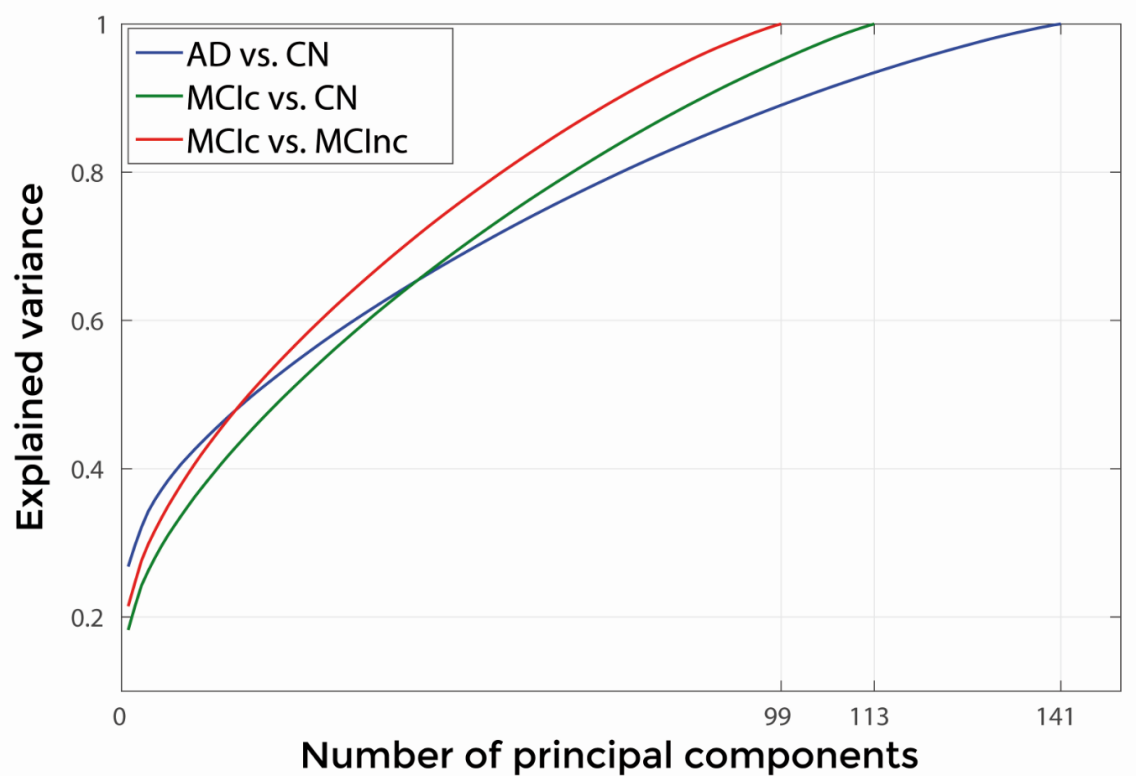


Figure 3.2.3 *Explained Variance as a function of the number of considered Principal Components, when using GM tissue probability map and no smoothing, for the following comparisons: AD vs. CN, MCIc vs. CN, MCIc vs. MCInc.*

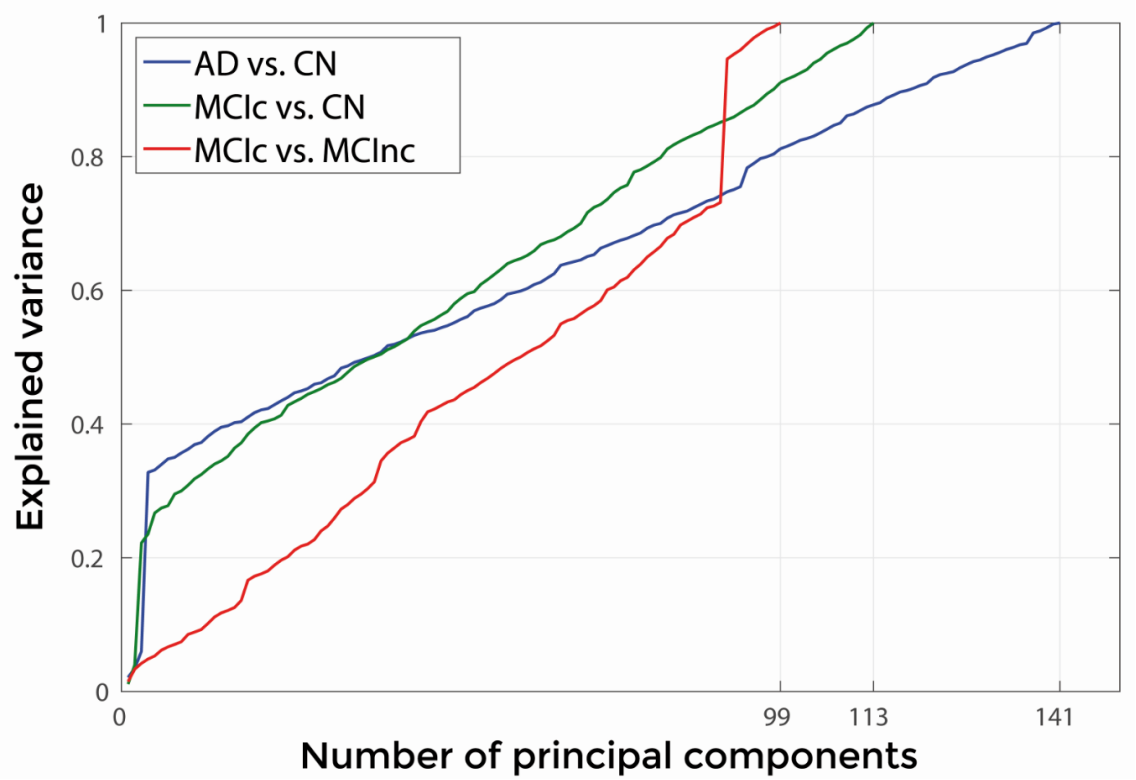


Figure 3.2.4 *Explained Variance as a function of the number of considered Principal Components sorted in accordance to their FDR, when using GM tissue probability map and no smoothing, for the following comparisons: AD vs. CN, MCIc vs. CN, MCIc vs. MCInc.*

In Figure 3.2.5, a representative example of the hyper-plane separating AD from CN subjects is shown when using the following parameters: 3 PCA coefficients, GM tissue probability map and an isotropic Gaussian kernel with 10 mm^3 FWHM for smoothing. The number of subjects involved in this case was 142, including 67 AD and 75 CN, while the total number of extracted PCA coefficients was 141.

3.2.4 Optimization of classification and performance evaluation

Figures 3.2.6-3.2.8 show E (i.e., 1 – Balanced Accuracy) as a function of the applied smoothing (FWHM – mm^3) and of the number of PCA coefficients. Plots are shown for the classification of AD versus CN, MCIc versus CN and MCIc versus MCInc when using GM tissue probability maps.

In Table 3.2.1, optimal parameters resulting from classifier optimization are reported. For all the classifications (AD versus CN, MCIc versus CN, MCIc versus MCInc), minimum values of E (for the 20 rounds of the nested 20-fold CV) were obtained –mostly– when using GM tissue probability maps (with a frequency of 100% for AD versus CN, 85% for MCIc versus CN and 80% for MCIc versus MCInc). On the other side, the number of PCA coefficients and the FWHM value of the isotropic Gaussian kernel for smoothing resulted to be different among the 20 rounds.

For the classification of AD versus CN, the best set of optimal parameters among the 20 rounds was: GM tissue probability map; 10 mm^3 FWHM of the isotropic Gaussian kernel for smoothing; 127 PCA coefficients. When using these parameters, E

reached its minimum value of 0.08. For the classification of MCIc versus CN, the best set of optimal parameters among the 20 rounds was: GM tissue probability map; 6 mm^3 FWHM of the isotropic Gaussian kernel for smoothing; 67 PCA coefficients. When using these parameters, E reached its minimum value of 0.14. For the classification of MCIc versus MCInc, the best set of optimal parameters among the 20 rounds was: GM tissue probability map; 2 mm^3 FWHM of the isotropic Gaussian kernel for smoothing; 34 PCA coefficients. When using these parameters, E reached its minimum value of 0.27.

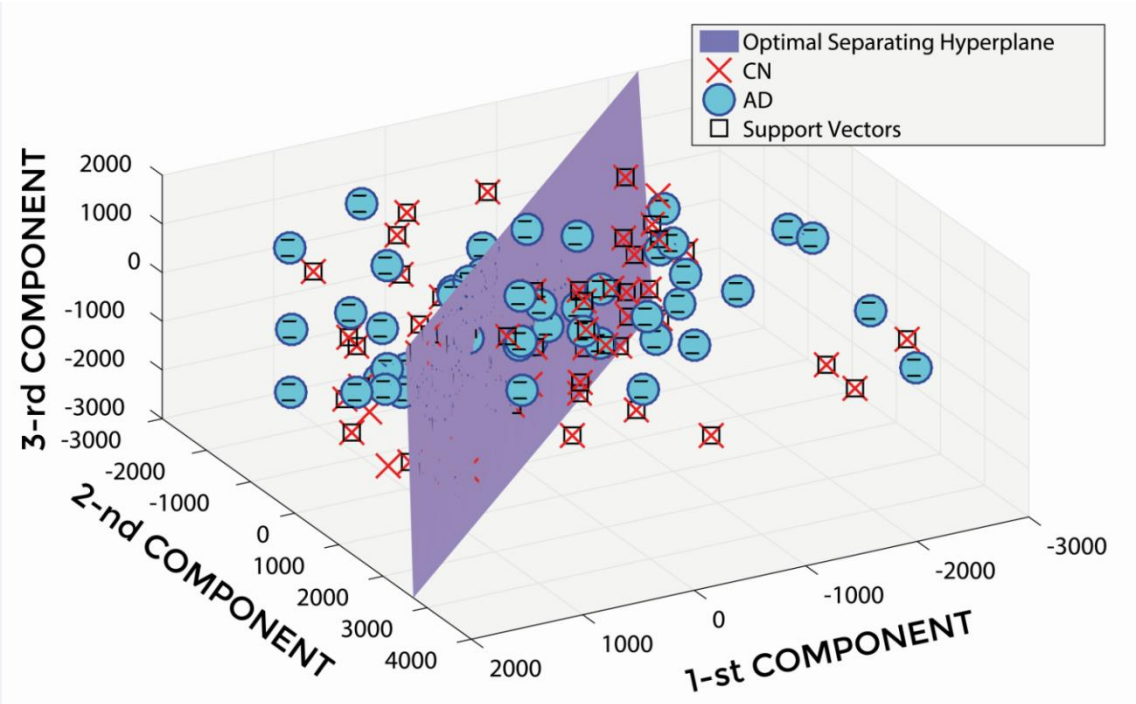


Figure 3.2.5 *Hyper-plane plane separating AD (o symbol) from CN (x symbol) PCA coefficients (3 PCA coefficients), and defined Support Vectors (□ symbol), when using GM tissue probability map and an isotropic Gaussian kernel with 10 mm3 FWHM for smoothing. 1st, 2nd and 3rd components are shown.*

The Overall Balanced Accuracy (averaged across all the 20 rounds of the nested 20-fold CV) was 0.76 ± 0.11 for the classification of AD vs. CN, 0.72 ± 0.12 for the classification of MCIc vs. CN, 0.66 ± 0.16 for the classification of MCIc vs. MCInc, respectively.

Since MMSE resulted significantly different between CN and patients (i.e., AD and MCIc), we tested our ML method after including MMSE as additional feature (besides MR data) to be given as input to the implemented classification algorithm. Balanced Accuracy resulted to be affected by the inclusion of MMSE among the input features (from 0.76 ± 0.11 to 0.99 ± 0.03 for AD vs. CN, from 0.72 ± 0.12 to 0.78 ± 0.16 for MCIc vs. CN, from 0.66 ± 0.16 to 0.60 ± 0.17 for MCIc vs. MCInc). This result is not surprising, given the distribution of the MMSE scores among groups that is shown in Figure 3.2.9.

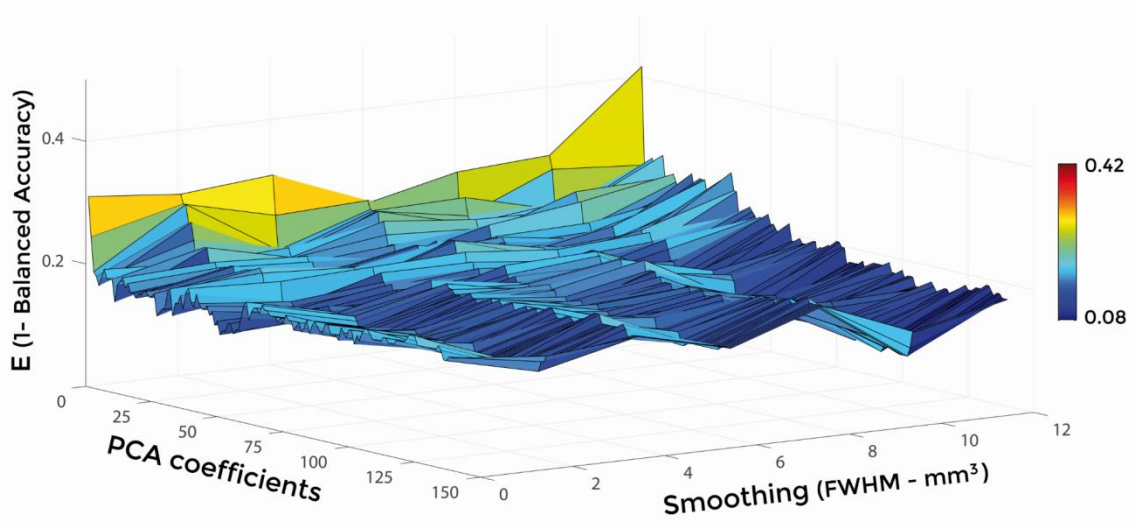


Figure 3.2.6 *E (1 – Balanced Accuracy) as a function of smoothing (FWHM – mm3) and number of PCA coefficients for the comparison between AD and CN when using GM.*

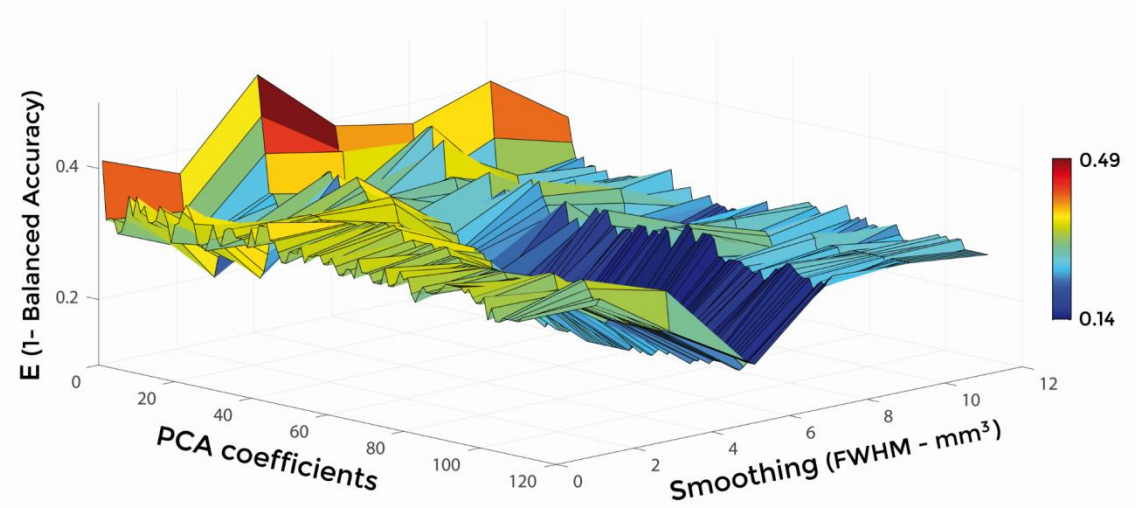


Figure 3.2.7 *E (1 – Balanced Accuracy) as a function of smoothing (FWHM – mm3) and number of PCA coefficients for the comparison between MCIc and CN when using GM.*

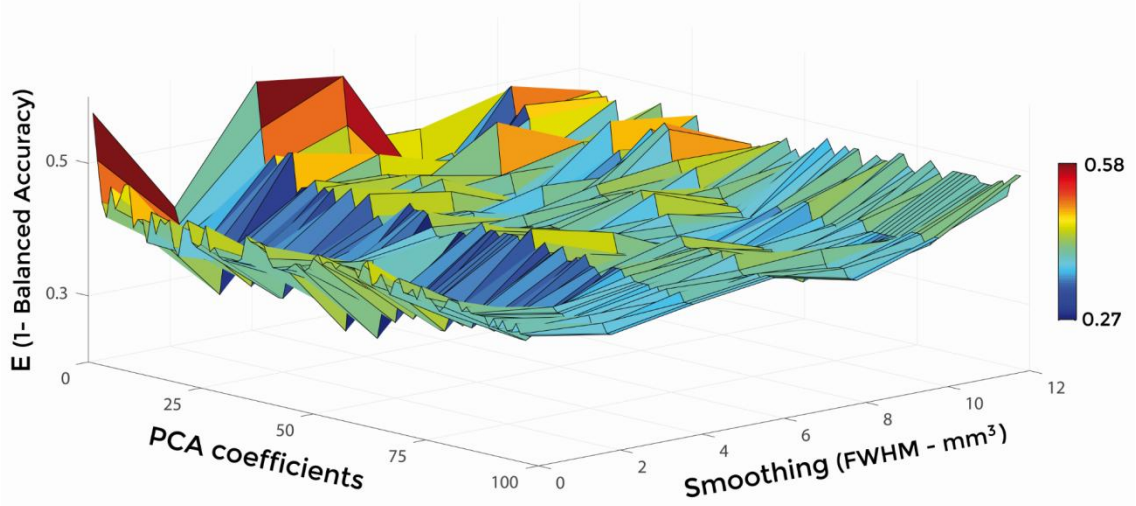


Figure 3.2.8 *E (1 – Balanced Accuracy) as a function of smoothing (FWHM – mm3) and number of PCA coefficients for the comparison between MCIc and MCInc when using GM.*

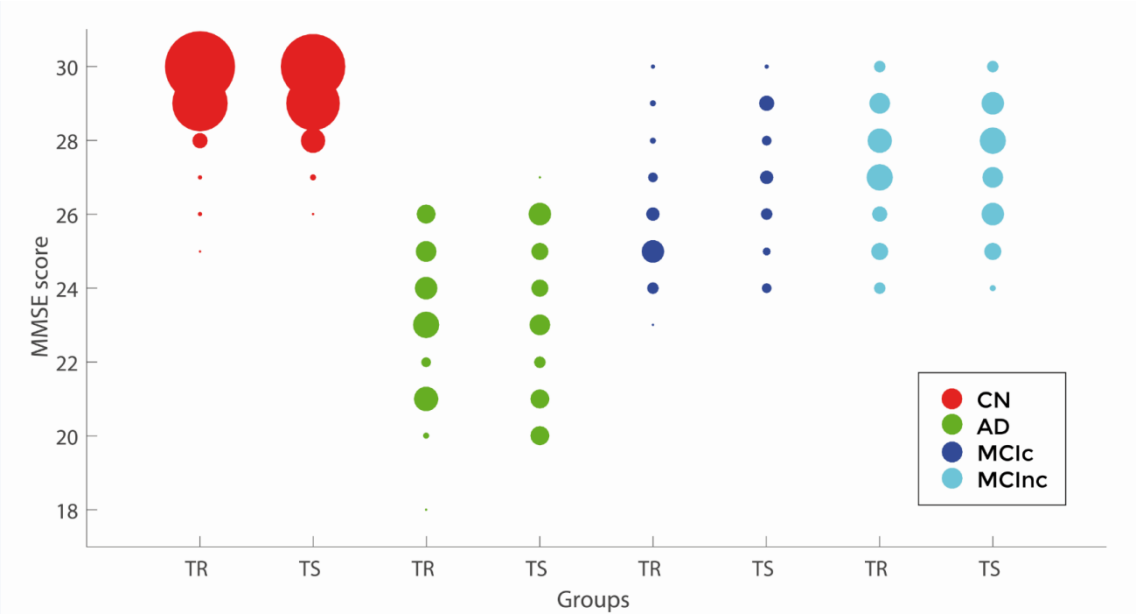


Figure 3.2.9 *Scatter plot reporting MMSE scores for the training (TR) and testing (TS) subsets of CN (red), AD (green), MCIc (blue) and MCInc (light blue) groups of subjects. For each MMSE score, the dimension of the circles in the plot is proportional to the number of subjects who obtained that score.*

In Figure 3.2.10, the explained variance is shown as a function of the number of considered principal components after sorting them according to their FDR. Data are shown for the classification of AD versus CN (blue), MCIc versus CN (green) and MCIc versus MCInc (red) when using the optimal configuration highlighted (in bold) for each classification in Table 3.2.1. For the classification of AD versus CN, the percentage of variance explained by the first 127 components was 98%; for the classification of MCIc versus CN, the percentage of variance explained by the first 67 components was 74%; for the classification of MCIc versus MCInc, the percentage of variance explained by the first 34 components was 50%.

In the paper published by Cuingnet and coworkers (2011), the same group of ADNI subjects employed in the work described in this section was used with the aim of evaluating and comparing the performances of different ML methods for the classification of AD. In that case, the dataset was half-splitted in order to use one half of the original dastaset to estimate the optimal hyperparameters of the classifiers and the remaining half of the original dataset to evaluate the classification performances. It must be noted that, among the 28 algorithm configurations tested by Cuingnet and colleagues, only one reported a Balanced Accuracy higher than 0.66 (that is, the one obtained in our work) for the classification of MCIc vs. MCInc.

Table 3.2.1 Classification error and optimal parameters (Tissue map, Smoothing, Number of PCA coefficients) for each of the 20 rounds of the inner training-and-validation loop (best configuration in bold).

Comparison	E	Tissue map	Smoothing FWHM [mm^3]	PCA coefficients
AD vs. CN	0.10	GM	6	6
	0.08	**GM**	**10**	**127**
	0.12	GM	10	41
	0.11	GM	4	62
	0.11	GM	6	75
	0.15	GM	2	64
	0.13	GM	8	69
	0.12	GM	4	32
	0.12	GM	2	67
	0.11	GM	2	50
	0.12	GM	4	48
	0.09	GM	4	54
	0.13	GM	8	35
	0.12	GM	2	118
	0.12	GM	4	46
	0.13	GM	2	22
	0.12	GM	2	135
	0.15	GM	6	49
	0.11	GM	6	54
	0.12	GM	12	30
MCIc vs. CN	0.19	GM	8	26
	0.17	GM	2	25
	0.20	GM	2	94
	0.19	GM	4	53
	0.22	WB	2	14
	0.20	WB	10	57
	0.15	GM	4	62
	0.15	GM	10	22
	0.21	GM	10	75
	0.19	GM	10	32
	0.14	**GM**	**6**	**67**
	0.19	GM	4	13
	0.19	WB	6	64
	0.17	GM	10	80
	0.22	GM	8	28
	0.18	GM	12	21
	0.19	GM	12	16
	0.19	GM	10	81
	0.19	GM	8	76
	0.19	GM	8	101
MCIc vs. MCInc	0.30	GM	2	9
	0.31	GM	10	19
	0.33	WB	12	34
	0.34	GM	4	34
	0.32	GM	8	16
	0.30	GM	6	17
	0.33	WB	2	21
	0.28	GM	6	10
	0.27	**GM**	**2**	**34**
	0.31	WM	4	4
	0.31	GM	2	16
	0.32	GM	8	31
	0.32	GM	8	23
	0.30	GM	4	46
	0.34	GM	8	33
	0.33	GM	8	2
	0.32	WB	4	34
	0.28	GM	10	5
	0.30	GM	8	8
	0.30	GM	2	84

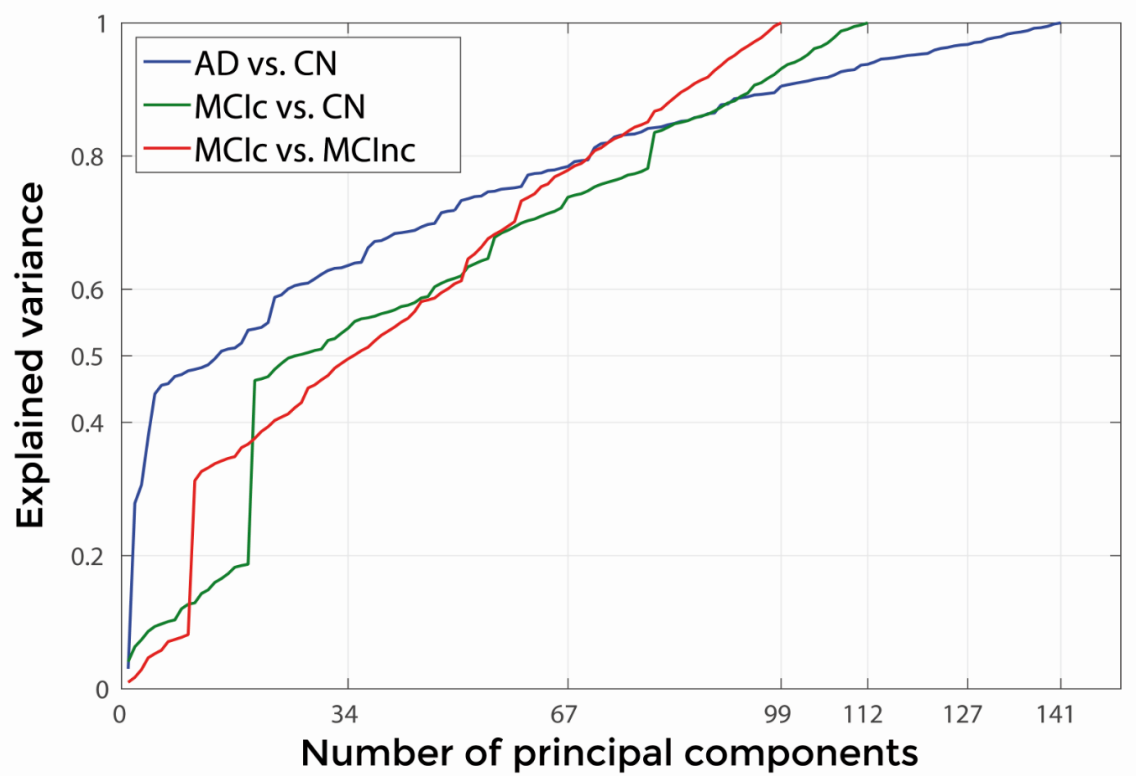


Figure 3.2.10 *Explained Variance, when using the best set of optimal parameters, as a function of the number of considered Principal Components sorted in accordance to their FDR, for the following comparisons: AD vs. CN, MCIc vs. CN, MCIc vs. MCInc.*

3.2.5 Diagnostic MR-related biomarkers

Figures 3.2.11-3.2.13 show voxel-based pattern distribution maps for the three following classification: 1) AD versus CN, 2) MCIc versus CN, 3) MCIc versus MCInc. The pattern of differences (normalized to a range between 0 and 1) is expressed according to the color scales.

Figure 3.2.11 shows the pattern of differences between AD and CN. As it can be seen, voxels influencing the classification of AD with respect to CN are localized in the temporal pole, superior and medial temporal cortex, including hippocampus and entorhinal cortex, amygdala, thalamus, putamen, caudate, insular cortex, gyrus rectus, lateral orbitofrontal cortex, superior and inferior frontal cortex, anterior cingulate cortex, precuneus and in the posterior cerebellar lobule.

Considering the classification of MCIc versus CN, the major part of voxel-based pattern distribution was similar to the one previously found in AD, as shown in Figure 3.2.12.

In the direct classification of the two MCI groups (MCIc versus MCInc), we only detected voxels influencing classification of MCIc with respect to MCInc, as it can be seen in Figure 3.2.13. In other words, in the brain of the patients with MCInc no anatomical changes useful to increase the accuracy of classification were detected. Overall, the major part of voxel-based pattern distribution was similar to the one detected in the previous MCIc versus CN classification.

The anatomical features detected in this application of the proposed ML method to the classification of AD are in line with previous research showing the precedence of pathologic changes in the temporal and parietal cortex (Braak and Braak, 1991; Schroeter et al., 2009), both in terms of spatial localization and extent.

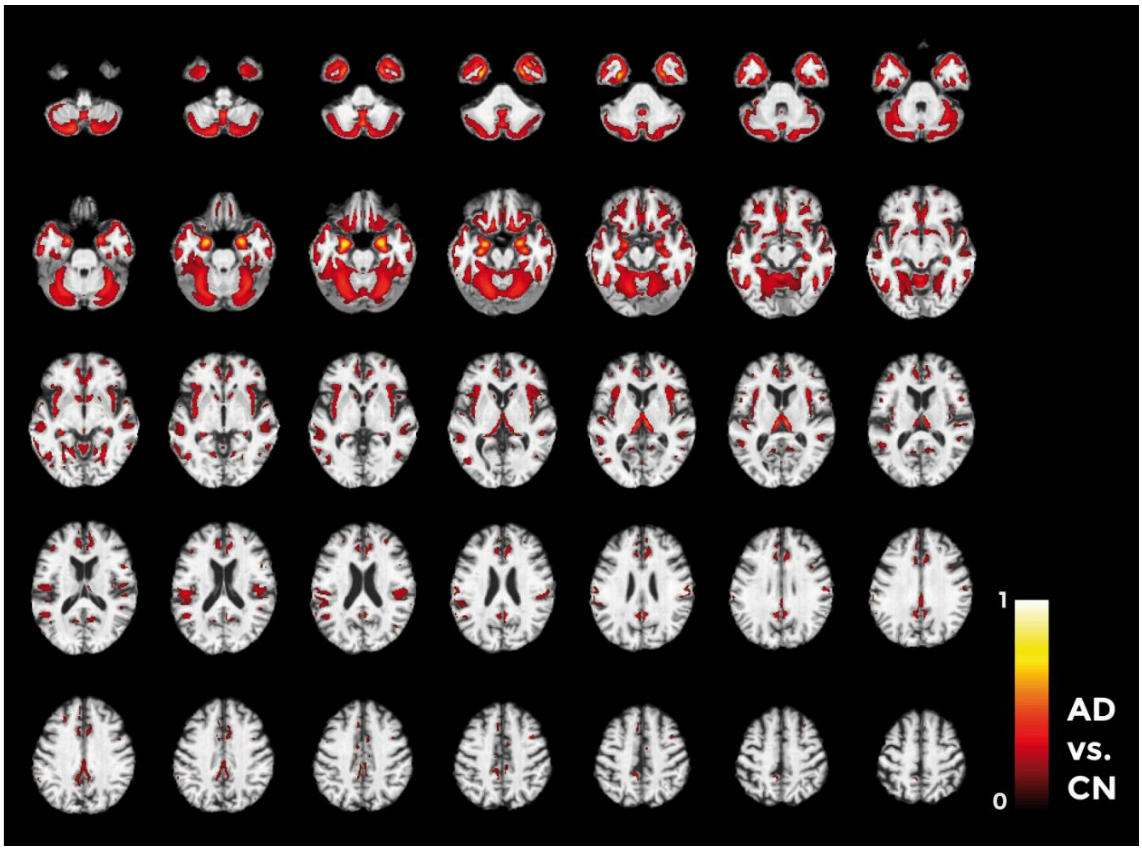


Figure 3.2.11 *Voxel-based pattern distribution map (axial view) for the classification between AD and CN. Voxel-based pattern distribution (normalized to a range between 0 and 1) is expressed according to the color scale (threshold = 50%) and superimposed on a standard stereotactic brain for spatial localization.*

A recent meta-analysis (Schroeter et al., 2009) focusing on the characterization of the prototypical neural substrates of AD and its prodromal stage amnesty MCI using neuroimaging data, showed that the following features resulted to be important for the discrimination of AD versus CN and MCI versus CN, respectively:

(a) reduction in glucose utilization and perfusion in the inferior parietal lobules, posterior superior temporal sulcus, precuneus, posterior cingulate cortex, anterior medial frontal cortex, anterior cingulate gyrus and right inferior temporal sulcus; hypometabolism in the right frontal pole, left posterior middle frontal gyrus and left hippocampal head; gray matter atrophy in both amygdalae, both anterior hippocampal formations, entorhinal areas, medial thalamus, posterior insula, left middle temporal gyrus and superior temporal sulcus, when comparing AD (826 patients) versus CN (1097 subjects);

(b) reduction of glucose utilization and perfusion in the inferior parietal lobules and the posterior cingulate cortex and precuneus; hypometabolism in the left anterior superior insula; gray matter atrophy in the left temporal pole, anterior superior temporal sulcus, right amygdala and gyrus rectus, when comparing MCI (525 patients) versus CN (1087 subjects).

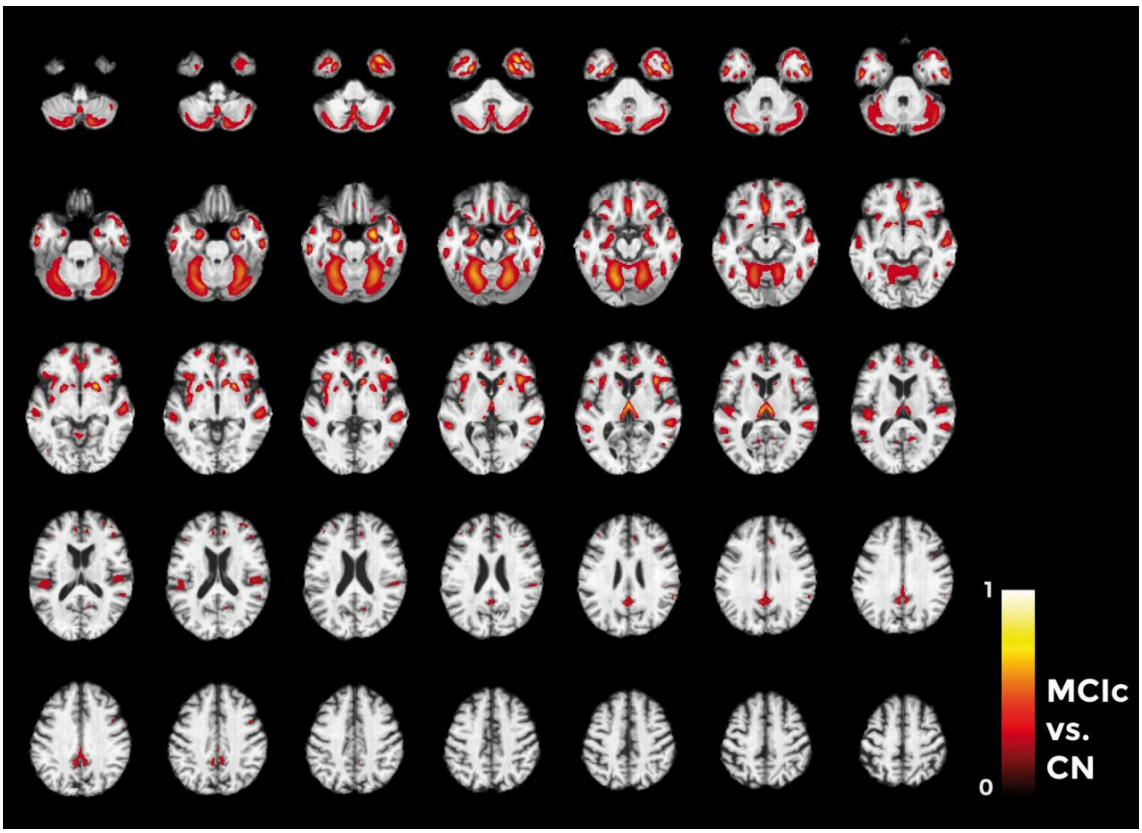


Figure 3.2.12 *Voxel-based pattern distribution map (axial view) for the classification between MCIc and CN. Voxel-based pattern distribution (normalized to a range between 0 and 1) is expressed according to the color scale (threshold = 45%) and superimposed on a standard stereotactic brain for spatial localization.*

In our study, the computation of diagnostic MR-related biomarkers showed only one brain region that seems to be not usually related to AD, i.e., the cerebellum. Specifically, our method identified the posterior lobule of the cerebellum as important area for the classification of AD and MCI. Atrophy of the cerebellum has been pointed out as important neuroimaging marker for AD in very few studies (Thomann et al., 2008a; Nigro et al., 2014). However, many studies based on histo-pathological analysis showed the presence of degenerative changes in the cerebellum of patients with AD compared to CN (Li et al., 1999; Wegiel et al., 2000; Wang et al., 2002). These degenerative changes include reduced Purkinje cell density, atrophy of the molecular and granular cell layer, and a higher number of amyloid plaques in the cerebellar cortex of patients with AD with respect to CN.

Moreover, in a paper by Thomann and colleagues (Thomann et al., 2008b) cognitive performance of patients with AD was found to be significantly correlated

with volumes of posterior cerebellar lobes. This result seems to be in line with the one found in this work, as our ML method detected anatomical changes only in the posterior lobule of the cerebellum.

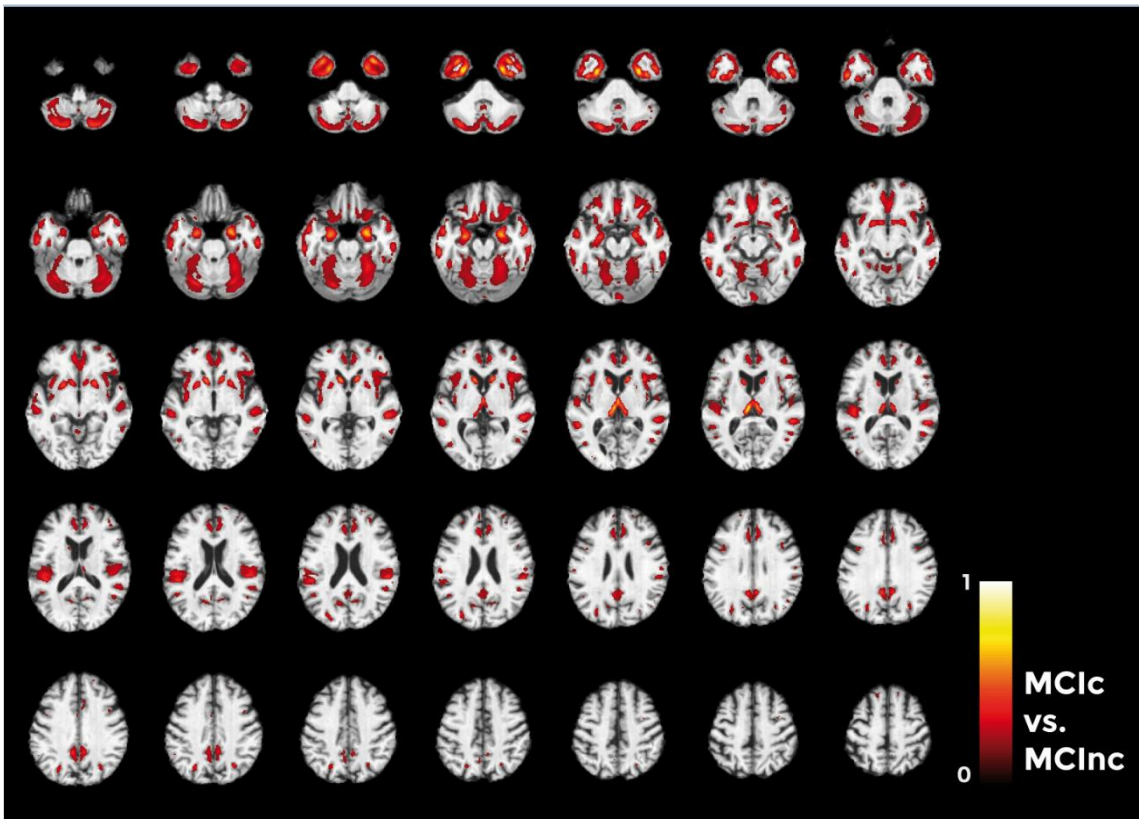


Figure 3.2.13 *Voxel-based pattern distribution map (axial view) for the classification between MCIc and MCInc. Voxel-based pattern distribution (normalized to a range between 0 and 1) is expressed according to the color scale (threshold = 35%) and superimposed on a standard stereotactic brain for spatial localization.*

In general, the pattern detected by the method developed in this work was similar for the investigated classification tasks (i.e., AD vs. CN, MCIc vs. CN and MCIc vs. MCInc). This findings suggest that the use of a DSS such as the implemented ML method may impact on the sensitivity (rather than on the specificity) in the detection of AD-related features. As a consequence of this, we could make the further hypothesis that the problem of how to perform diagnosis of AD at a very early stage (MCI) by MRI data seems to be a matter of increasing the detectability of structural MR biomarkers. If this hypothesis was confirmed by other studies, then the use of high-resolution MRI systems combined with advanced ML algorithms for image processing would allow to move the diagnostic role of MRI from the clinical to the preclinical stage of AD. This should encourage the development and implementation of more advanced pattern recognition algorithms, but also of MRI systems with improved sensitivity, increased resolution and better S/N ratio.

3.3 Application III: Eating Disorders

3.3.1 Participants and psychiatric assessment

Compared with age-/sex-/BMI-matched CN, ED patients did not show global anatomical atrophies in white or gray matter brain volumetry. At a behavioral level, ED group was characterized by a well-known psychopathological profile, as it reported in Table 3.3.1 and in Supplementary Materials of the paper by Cerasa et al. (2015). In particular, as demonstrated by EDI-2, ED patients had higher scores for (a) drive for thinness scale ($t= 4.45$; p-level < 0.00001); (b) bulimia scale ($t= 2.69$; p-level $= 0.01$); (c) interoceptive awareness scale ($t= 3.81$; p-level $= 0.0006$); (d) asceticism scale ($t= 3.81$; p-level $= 0.0006$); (e) body dissatisfaction ($t=3.5$; p-level $= 0.001$); (f) interpersonal distrust scale ($t= 2.07$; p-level $= 0.04$) and (g) impulse regulation scale ($t= 2.46$; p-level $= 0.02$). No significant differences were detected for Perfectionism, Ineffectiveness, Maturity Fears and Social Insecurity scales, in agreement with previous studies (Amianto et al., 2013).

3.3.2 The classifier

None of the acquired MR images was excluded from this study due to problems with image quality or problems occurred during pre-processing (especially, normalization to the template).

In Figure 3.3.1a, 1-st and 2-nd extracted PCA coefficients that showed the highest FDR for the classification of ED versus CN are plotted. Data are shown from a single round of CV as a representative example. In this case, the number of subject involved was equal to 31 (16 ED and 15 CN) and the total number of extracted PCA coefficients was equal to 30. The analysis of variance showed that the percentage of variance retained by the first principal component was equal to 27.0%, while the number of extracted principal components accounting for 50% and 95% of the whole variance was 6 and 27, respectively.

In Table 3.3.2, FDR values of the 30 features (PCA coefficients) used for the classification of ED versus CN are reported. Data from a single round of CV are shown, as a representative example. As it can be seen, in this case the highest FDR value was reached by 8^{th} PCA coefficient, which accordingly resulted the most important feature for the classification of ED versus CN.

In Figure 3.3.1b, 1-st and 2-nd extracted PCA coefficients that showed the highest FDR (i.e., after FDR raking) are plotted jointly with 1-st and 2-nd extracted PCA coefficients considered before FDR ranking. As it can be seen from this plot, FDR allows improving binary group separation by identifying the most discriminative features for a given classification task (in this case, ED versus CN).

Figure 3.3.2 shows the predictive (or decision) function for the classification of ED versus CN resulting from the training phase of the classifier (1-st and 2-nd components with highest FDR are plotted).

Table 3.3.1 Demographic characteristics

Demographical data			
Variables	ED (n°17)	CN (n°17)	*P-level*
Age (years)	30.2 ± 5.6	30.1 ± 5.5	0.95
Educational level (years)	17 (13-21)	17 (13-21)	0.88
BMI	23.6 ± 8.2	24.1 ± 4.8	0.79
MRI data			
Total GM Volume	587.3 ± 37.5	608.88 ± 42.1	0.11
Total WM Volume	486.5 ± 63.1	489.6 ± 41.6	0.86
Total CSF Volume	188.3 ± 28.7	187 ± 23.2	0.88
Clinical Data			
HAMA	**14.6 ± 13**	**4 ± 2.2**	**0.04**
BDI	**16.8 ± 10.1**	**6.3 ± 4.7**	**0.0004**
DES	**14.32 ± 12.4**	**5.12 ± 4**	**0.007**
EAT-26	**23.3 ± 14.4**	**6.35 ± 3.2**	**0.00004**
SDQ-20	**28.64 ± 14.8**	**20.6 ± 1.1**	**0.03**
BIDA	29.9 ± 19.4	19.9 ± 11	0.24
Clinical Data EDI-2 scale			
DRIVE FOR THINNESS	**9.4 ± 6.3**	**1.2 ± 1.3**	**0.0001**
BULIMIA	**3.47 ± 4.5**	**0.1 ± 0.5**	**0.01**
INTEROCEPTIVE AWARENESS	**7.9 ± 6.2**	**0.7 ± 1.2**	**0.0006**
ASCETICISM	**5.6 ± 3.8**	**2 ± 1.1**	**0.0006**
BODY DISSATISFACTION	**12.9 ± 7.2**	**6.1 ± 2.9**	**0.001**
PERFECTIONISM	4.3 ± 3.9	3.3 ± 3.1	0.41
INTERPERSONAL DISTRUST	**3.6 ± 3.1**	**1.4 ± 1.2**	**0.04**
IMPULSE REGULATION	**3.67 ± 4.9**	**0.6 ± 1.4**	**0.02**
INEFFECTIVENESS	3.5 ± 5.2	1.2 ± 2.6	0.12
MATURITY FEARS	5.2 ± 3	3.94 ± 2.6	0.13
SOCIAL INSECURITY	3.53 ± 3.2	2.1 ± 2	0.22

Data are given as mean values (SD) or median values (range) when appropriate. BMI: Body-Mass Index; GM: Gray Matter; WM: White Matter; CSF: cerebrospinal fluid ; PBI: Parental bonding instrument; STAI: State-Trait Anxiety Inventory; HAMA: Hamilton rating scale for anxiety; BDI: Beck Depression Inventory; DES: Dissociative Experiences Scale; EAT-26: Eating attitude test-26; SDQ-20: Somatoform Dissociation Questionnaire-2; BIDA: Body Image Dimensional Assessment; EDI-2: Eating Disorder Inventory-2. Total brain MRI parameters have been calculated using VBM8 tool.

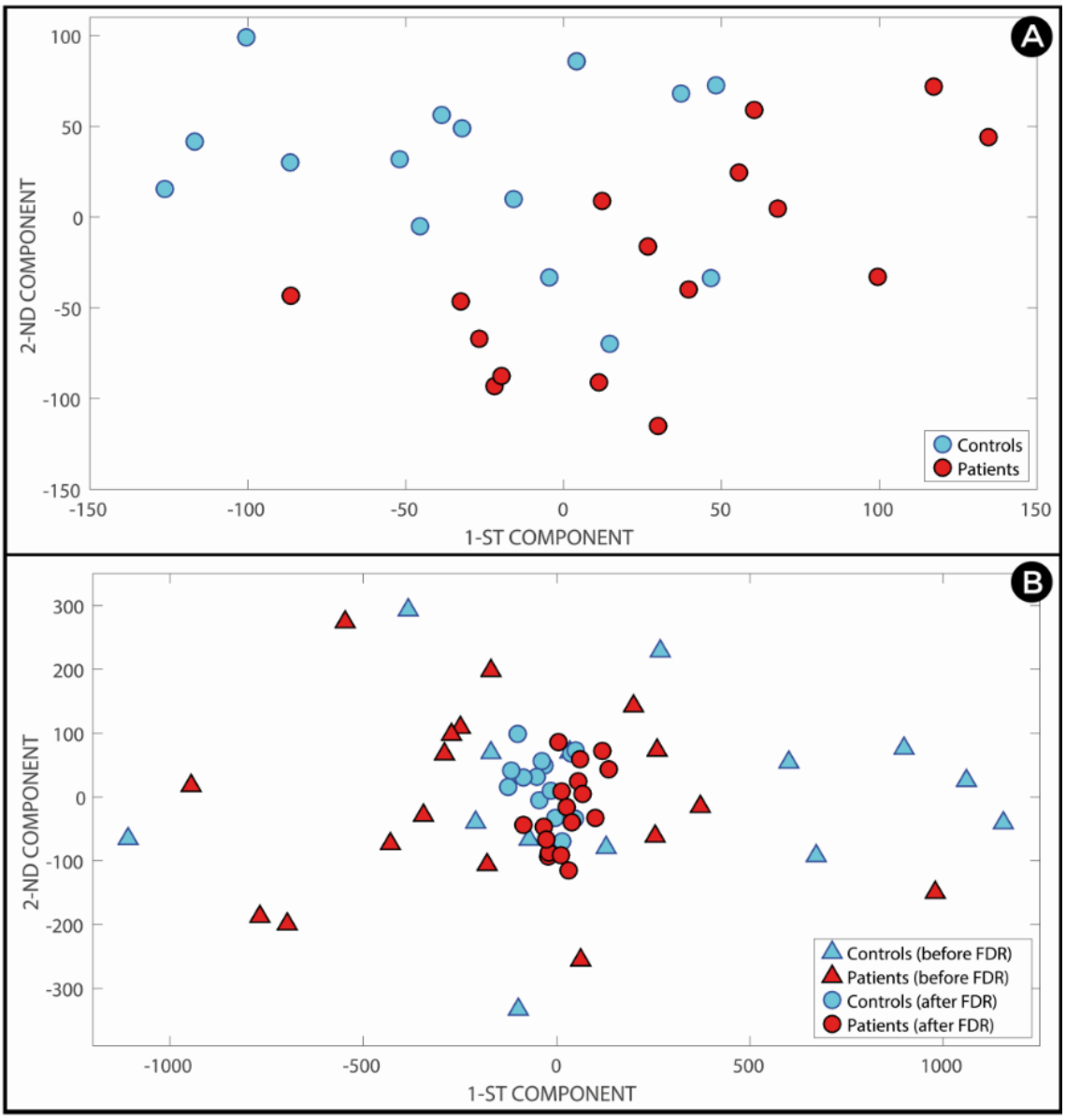


Figure 3.3.1 *Plot of the PCA coefficients that showed the highest FDR (A) and joint plot of the PCA coefficients before (triangles) and after (circles) FDR ranking (B) for the ED versus CN group discrimination (1-st and 2-nd components). Data from a single round of CV are shown as a representative example.*

3.3.3 Performance evaluation

As explained in subsection 2.4.4, classification performance in terms of accuracy, specificity and sensitivity for ED versus CN were obtained through 10-fold and 20-fold CV.

When considering 20-fold CV, classification performance were computed over a number of PCA coefficients ranging from 1 to 32. Accuracy, specificity and sensitivity reached their best values of 0.85, 0.73 and 0.93, respectively, when using 31 PCA coefficients.

When considering 10-fold CV, classification performance were computed over a number of PCA coefficients ranging from 1 to 30. Accuracy, specificity and sensitivity

reached their best values of 0.80, 0.72 and 0.96, respectively, when using 21 PCA coefficients.

Table 3.3.2 FDR values of the 30 features (PCA coefficients) used for the ED versus CN discrimination. *Data from a single round of CV are shown as a representative example.*

PCA coefficient (#)	FDR	PCA coefficient (#)	FDR
1	0.2052	16	0.0176
2	0.0172	17	0.0279
3	0.0021	18	0.0188
4	0.1286	19	0.0206
5	0.0005	20	0.0511
6	0.0786	21	0.0369
7	0.1484	22	0.0001
8	0.3923	23	0.0200
9	0.0354	24	0.0052
10	0.0137	25	0.1839
11	0.0919	26	0.0431
12	0.3376	27	0.0015
13	0.1057	28	0.0250
14	0.0002	29	0.0321
15	0.0128	30	0.0171

In Figure 3.3.3, accuracy, specificity and sensitivity rates are plotted as a function of the number of employed PCA coefficients for the classification of ED versus CN. As expected, the performance of the classification algorithm increases with the number of employed PCA coefficients.

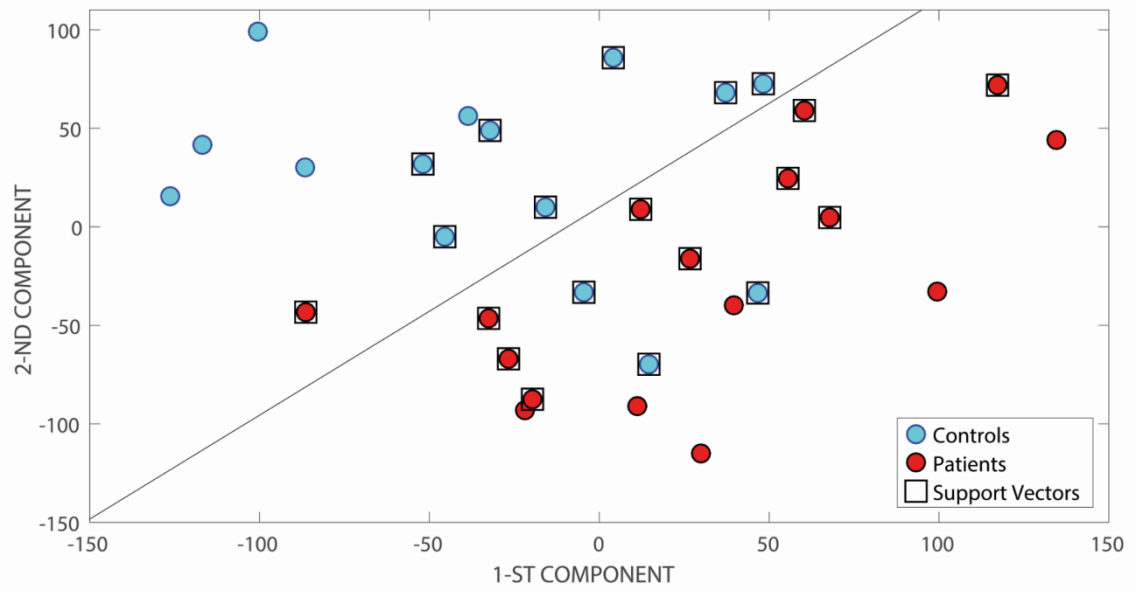


Figure 3.3.2 *Decision function for the ED versus CN group discrimination (1-st and 2-nd components with highest FDR). Data from a single round of CV are shown as a representative example.*

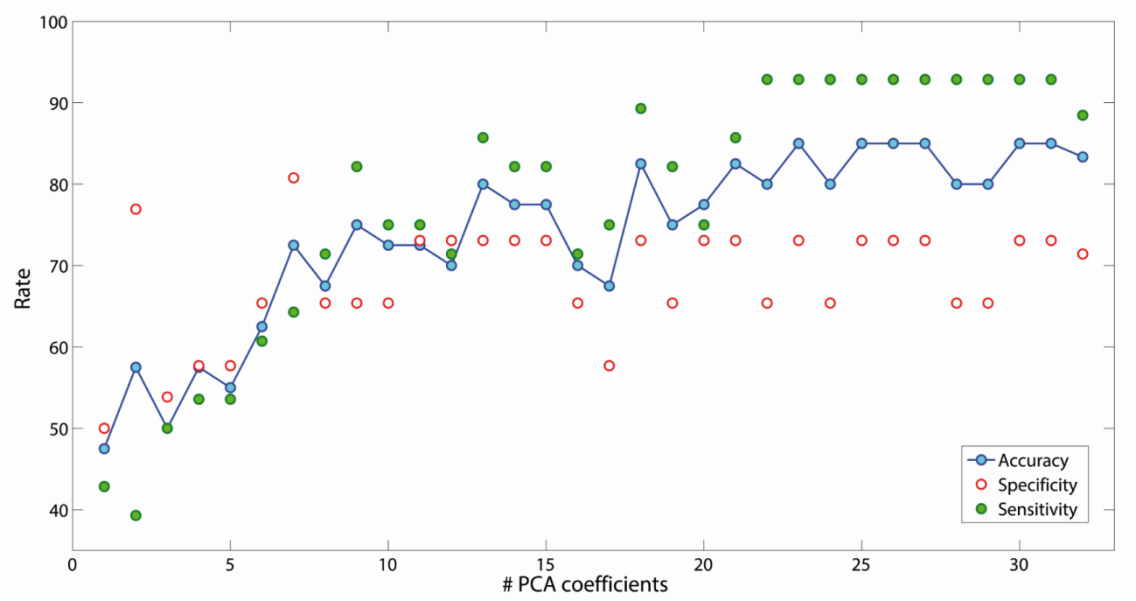


Figure 3.3.3 *Accuracy, Specificity and Sensitivity of classification as a function of the number of employed PCA coefficients for the ED versus CN group discrimination (20-fold CV).*

The classification accuracy reported above as obtained by the implemented ML algorithm for the automatic classification of ED vs. CN resulted to be comparable to or higher than those reported in published studies aiming at the classification of psychiatric disorders by means of ML: for example, 80–85% reported by Castellani et al. (2012) for the classification of schizophrenic patients; 81% reported by Almeida et al. (2013) for the classification of depression disorders; 75% reported by Pettersson-Yeo et al. (2013) for the classification of first-episode psychosis.

3.3.4 Diagnostic MR-related biomarkers

Figure 3.3.4 below shows the voxel-based pattern distribution map of brain structural differences for the classification of ED versus CN. The pattern of differences was mainly localized in the occipital cortex and the posterior cerebellar lobule. Other brain regions involved were in regulation of emotional processing known to be damaged in ED patients were detected: precuneus, sensorimotor and premotor cortices as well as the ACC and OFC.

Among the regions highlighted above, the cerebellum is a multidimensional brain structure that is involved in various functions, including motor, cognitive and emotional functions. The role of the cerebellum in visceral and autonomic regulation has been highlighted (Mahler et al., 1993; Zhu et al., 2008). In particular, the cerebellar vermis has a role in appetite regulation and in feeding behavior. This region is connected with limbic brain structures in an extensive way. Connected regions include hippocampus, parahippocampal gyrus, amygdala, thalamus, cingulate and prefrontal cortices (Middleton et al., 2001). Several structural neuroimaging studies (e.g. Boghi et al, 2011; Suchan et al, 2010; Husain et al, 1992) described the presence of GM volume loss mainly in AN, thus demonstrating the involvement of the cerebellum in ED, with a particular role played by the vermis subregion. A more recent study, conducted by Amianto and colleagues (Amianto et al., 2013) using resting state fMRI, demonstrated

the presence of altered intrinsic connectivity of the cerebellar vermis in both patients with AN and BN. As it can be read in this paper, there might be a relation between the resulting dysfunctional neural pattern and some psychopathological aspects that are altered in ED patients (e.g., drive for thinness).

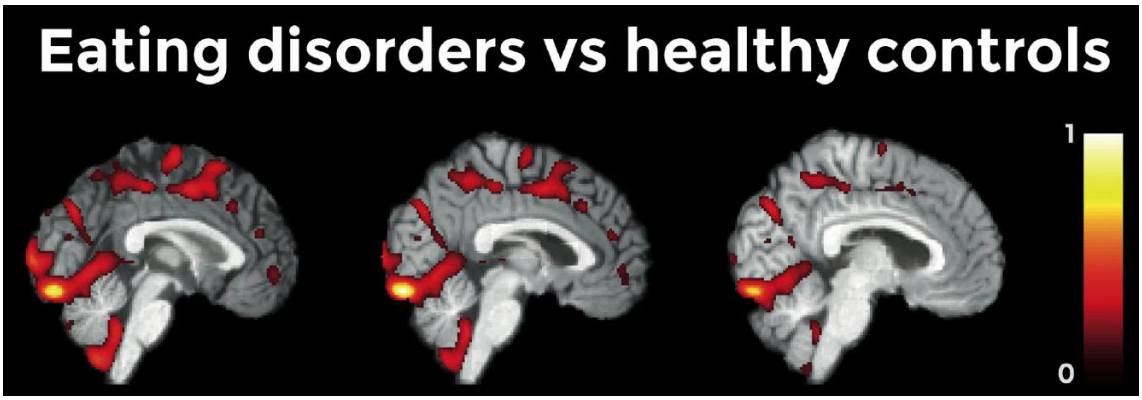


Figure 3.3.4 *Voxel-based pattern distribution map of brain structural differences between ED patients and CN (sagittal view, threshold = 50%). Voxel-based pattern distribution (normalized to a range between 0 and 1) is expressed according to the color scale and superimposed on a standard stereotactic brain for spatial localization.*

The ACC, together with the OFC, are two regions taking part in the ventral limbic circuit, together with the amygdala, insula and ventral striatum, which are important for identifying the emotional significance of appetizing stimuli for inhibiting impulsive behaviors (Marsch et al., 2015) and regulating reward systems (Avena and Bocarsly, 2012). The current neuroimaging literature mainly highlights the role of this neural network in pathophysiological mechanisms of BN. Indeed, the lack of control and the impulsivity in BN could be explained by the alterations of mesolimbic reward response mechanisms, which are neurophysiologically expressed through dysfunctional activities in the ACC and OFC regions (Brooks et al, 2012; Friederich et al, 2013). However, fronto-striatal neural circuit dysfunctions related to altered reward processing were also described in patients with AN (Keating et al, 2012), thus raising a different perspective in which stimuli that are otherwise aversive for healthy controls (e.g., self-starvation, emaciated body image), are considered as rewarding stimuli able to activate relevant reward linked brain regions in patients with AN.

The involvement of the visual cortex is another key site associated with ED. Indeed, altered functional activity of the occipital lobe has been reported in both AN and BN individuals (Brooks et al., 2011). However, body image disturbance is considered one of the core characteristics of AN. Several neuroimaging studies have described the neurobiological correlates of this symptom, defining the presence of a specific neural network involved in body processing: the fusiform area, the inferior temporal sulcus and the primary visual cortex. Recent evidences (Suchan et al, 2013) demonstrated altered effective connectivity between these regions in patients with AN during the viewing of bodies.

Finally, abnormal neural changes in the precuneus and sensorimotor/premotor cortices have already been described in both patients with AN and BN (Gaudio et al., 2011; Amianto et al, 2013; Suchan et al, 2013). Using body images of slim fashion models to induce a self-other body shape comparison, in the paper by Friederich et al. Friederich et al., 2013) it was shown that AN patients had a higher activation of the premotor cortex. Amianto et al. (2013) found altered gray matter volume in the paracentral lobule, precuneus and somatosensory regions when comparing AN and BN patients, as well as the whole ED group, with respect to controls. Altered neural changes in brain areas involved in sensorimotor functions and visuo-proprioceptive information processing may either represent the physiological consequence of physical hyperactivity typical of ED patients or as a dysfunction related to the body awareness. Body awareness is a complex cognition underpinned by aspects of visual perception, proprioception, and touch (Berlucchi and Aglioti, 2013). The processing of the body image concept requires integration of the different types of body-related perceptual experience and processing of information related to peripersonal space. The presence of altered anatomical changes in these regions together with visual cortex has been interpreted as a dysfunctional processing of somatosensory information about the perceived body size (Gaudio et al., 2011; Favaro et al., 2012).

3.4 Other applications

3.4.1 Autism Spectrum Disorder

3.4.1.1 Participants

Table 3.4.1 summarizes demographic, cognitive, and clinical characteristics of the participants.

3.4.1.2 Data analysis

The analysis confirmed the validity of mental age matching ($p > 0.05$). Gender was also balanced between ASD and TD groups, as there were 3 girls in the ASD group and 2 girls in the healthy control group ($\chi^2(1) = .240$; $p > 0.05$). IQ and chronological age were not balanced across groups (both $p < 0.001$), as expected.

In Table 3.4.2 kinematic feature values of the two groups of children (ASD and TD) included in the study are reported jointly with the results of ANCOVA calculated on all kinematic measures. Even after controlling for between-participant differences in IQ and chronological age, several significant group differences were identified for the kinematic variables.

3.4.1.3 The classifier

Figure 3.4.1 shows, as a representative example, the optimal separating hyper-plane for the classification of ASD vs. TD participants as resulting from the training phase of the ML method.

Table 3.4.1 Demographics of the participants

	ASD	TD	t(1,28)	p
N	15	15		
Females : Males	3:12	2:13		
Chronological Age[a]	3;5 ± 7,7 (2;8 - 4;6)	2;6 ± 5,2 (1;7 - 2;9)	-4.55	< .001
Mental Age[a]	2;6 ± 5,7 (1;7 - 3;4)	2,7 ± 5,9 (1;6 - 3;2)	.513	n.s.
IQ[b]	75 ± 13,4 (51 - 96)	105 ± 12,7 (81 - 119)	6.52	< .001
ADOS[c] Social Communication SBRI[d]	 11 ± 2,2 7 ± 1,5 2 ± 1,6	 – – –		

ASD = autism group; TD = typically developing group; IQ and mental age were assessed using the Griffiths Mental Development Scales (Griffiths, 1970).
a Mean years; months ± standard deviation (range)
b Mean ± standard deviation (range)
c ADOS autism diagnostic observation schedule, Lord et al. (2000)
d Stereotyped Behavior and Restricted Interests scale.

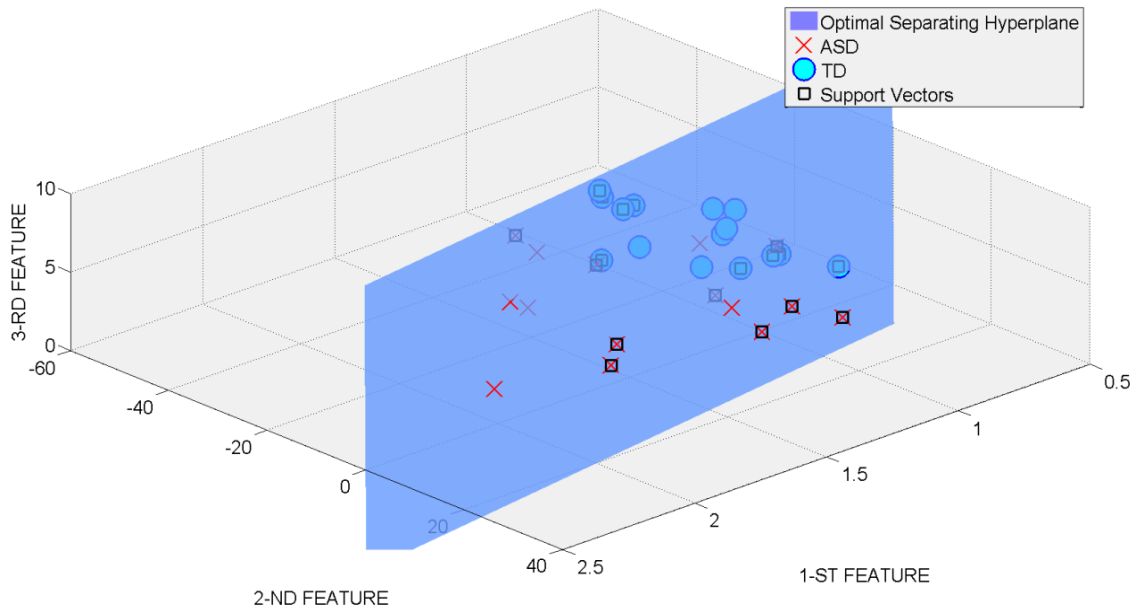


Figure 3.4.1 *Optimal separating hyper-plane for the Autism group (ASD) versus typically developing groups (TD) (1st, 2nd and 3rd components) is shown as a representative example of the training phase of the machine-learning method*

Table 3.4.2 Kinematic data were initially analyzed through an ANCOVA with Group (ASD vs. TD) as a between-participant factor, and with IQ and chronological age as covariates. The alpha level was set to .05 for all data analyses. Table depicts group means and standard deviations for kinematic variables, values of F test, p values and effect sizes reported using partial eta squared (η_p^2).

Submovement 1		ASD	TD	F (1,26)	Sig.	η_p^2
Movement Units	M (SD)	1.91 (0.62)	1.70 (0.37)	<1.0	n.s.	.012
Total Movement Duration	M (SD)	0.69 (0.14)	0.66 (0.12)	<1.0	n.s.	.010
Peak Velocity	M (SD)	0.46 (0.12)	0.59 (0.17)	5.626	**<0.05**	.178
Time of Peak Velocity	M (SD)	0.34 (0.07)	0.31 (0.04)	<1.0	n.s.	.036
Peak Acceleration	M (SD)	3.18 (0.93)	4.26 (1.52)	7.884	**<0.01**	233
Time of Peak Acceleration	M (SD)	0.21 (0.07)	0.16 (0.05)	<1.0	n.s.	.031
Peak Deceleration	M (SD)	-3.59 (1.28)	-3.93 (1.44)	<1.0	n.s.	.067
Time of Peak Deceleration	M (SD)	0.47 (0.08)	0.44 (0.06)	<1.0	n.s.	.017
Submovement 2						
Movement Units	M (SD)	3.45 (1.78)	1.76 (0.39)	4.408	**<0.05**	.145
Total Movement Duration	M (SD)	1.35 (0.44)	0.79 (0.15)	13.832	**=0,001**	.347
Peak Velocity	M (SD)	0.61 (0.15)	0.76 (0.16)	13.475	**=0,001**	.341
Time of Peak Velocity	M (SD)	0.41 (0.14)	0.31 (0.05)	18.501	**<0,001**	.416
Peak Acceleration	M (SD)	3.85 (1.13)	5.58 (1.94)	12.416	**<0,01**	.323
Time of Peak Acceleration	M (SD)	0.23 (0.20)	0.13 (0.04)	6.303	**<0.05**	.195
Pick Deceleration	M (SD)	-3.29 (1.15)	-4.27 (1.88)	2.632	n.s.	.092
Time of Peak Deceleration	M (SD)	0.75 (0.24)	0.51 (0.11)	26.652	**<0,001**	.506
Wrist Angle	M (SD)	-4.25 (16.34)	-25 (12.40)	6.604	**<0.05**	.203

ASD = autism group; TD = typically developing group;

3.4.1.4 Performance evaluation and extraction of the most discriminant features

Table 3.4.3 shows accuracy, specificity, and sensitivity of the machine-learning method for the comparison of ASD versus TD.

Table 3.4.3 Accuracy, Specificity and Sensitivity rates of SVM using LOO validation.

	Maximum Accuracy (%) (# selected features)	Maximum Specificity (%) (# selected features)	Maximum Sensitivity (%) (# selected features)
	Overall Mean Accuracy (%)	Overall Mean Specificity (%)	Overall Mean Sensitivity (%)
ASD vs. TD	96.7 (7)	93.8 (7)	100.0 (7)
	84.9	89.1	82.2

ASD = autism group; TD = typically developing group. The maximum values reached by Accuracy, Specificity and Sensitivity were referred to as Maximum Accuracy, Specificity and Sensitivity rates. Accuracy, Specificity and Sensitivity reached their maximum values using 7 features, all related to the second part of the movement -Sub movement 2-: (1) Total Duration; (2) delta Wrist Angle; (3) number of Movement Units; (4) time of Peak Deceleration; (5) Peak Acceleration; (6) time of Peak Velocity; (7) Peak Velocity.

The implemented ML method was able to classify participants by diagnosis reaching a maximum accuracy of 96.7% (specificity 93.8% and sensitivity 100%) when 7 features selected by the FDR-based technique were given as input to the classifier. Overall mean accuracy, specificity, and sensitivity rates were calculated over a number of selected features ranging from one to 17 (the whole number of features). The overall mean classification accuracy (specificity/sensitivity) for ASD vs. TD was 84.9% (89.1%/ 82.2%).

The dependence of the metrics on the number of considered features is shown in Figure 3.4.2. The resulting data are shown for a number of features ranging from 1 to 17. As it can be seen, accuracy, specificity and sensitivity increase with the number of selected features, reaching their maximum values when considering the 7 selected features by FDR.

Although diagnosis of ASD is particularly difficult in young, low-functioning children (even through the standard diagnostic procedure), classification performance achieved in this work suggest the validity of the proposed method to classify preschool this disorder on the basis of a motor task.

As explained in subsection 2.5.1.5, besides evaluating the classification performance of the implemented method, our analysis allowed us to identify which

kinematic features gave the greatest contribution to the classification of ASD versus TD.

In order to classify ASD versus TD, 7 out of 17 features were sufficient, reaching an accuracy of 96.7%. These 7 features are (in descending order) (1) total duration, (2) delta wrist angle, (3) number of movement units, (4) time of peak deceleration, (5) peak acceleration, (6) time of peak velocity and (7) peak velocity of the *sub-movement 2*. It is worth noting that all these 7 kinematic features are related to the *sub-movement 2*, i.e., the second part of the movement, aimed to transport the ball from a support to the target hole before dropping it. In particular, the top 3 features resulting from this analysis indicate that the movements of children with ASD are slower and more fragmented, leading to inappropriate hand inclination for ball-drops during the final phases of hand transport. These findings are in line and extend previously published results, showing the difficulty of translating intention into a motor chain aiming at an action goal in children with ASD (Cattaneo et al., 2007; Fabbri-Destro et al., 2009; Forti et al., 2011).

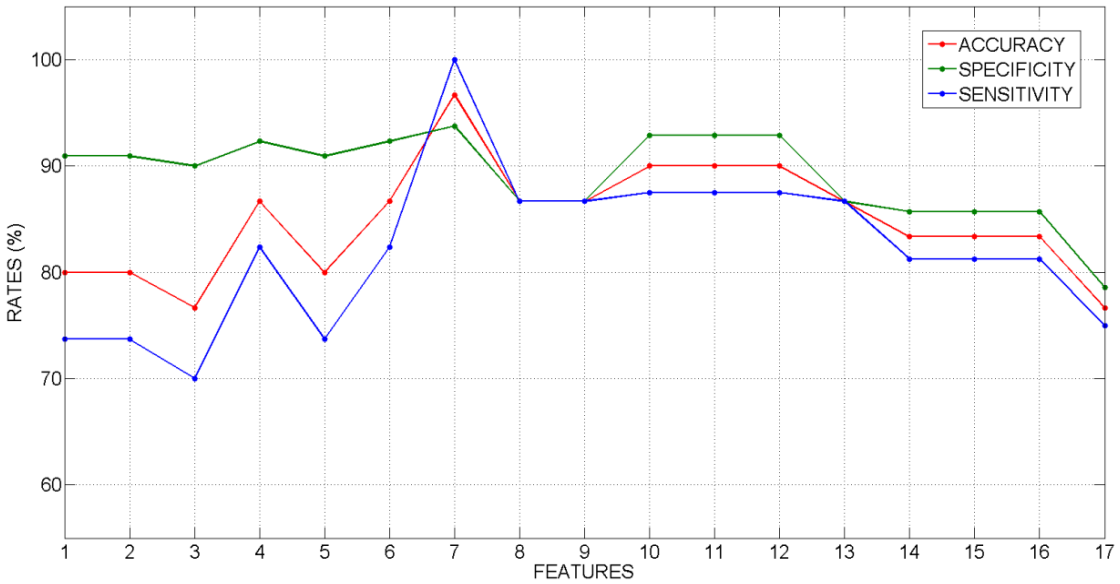


Figure 3.4.2 *Graph showing classification Accuracy, Specificity and Sensitivity rates (%) of SVM (Y-axis) in relation of the number of considered features (X-axis). As expected, Accuracy, Specificity and Sensitivity rates increased with the number of selected features. The classification accuracy reached a Maximum Accuracy of 96.7% (Specificity 93.8%, and Sensitivity 100%) utilizing seven features. All of these 7 kinematic features are related to the second part of the movement Sub movement 2 (i.e., the movement to transport the ball from a support to the target hole in which the ball was to be dropped), suggesting that goal oriented movements may be critical in separating children with ASD from typically developing children.*

When considering all of the N rounds (30) of the LOO validation approach, the most discriminative features between the two groups are (in descending order): Total Duration sub movement 2, Delta Wrist Angle, Movement Units sub movement 2, time of Peak Deceleration sub movement 2, Peak Acceleration sub movement 2, time of Peak Velocity sub movement 2, Peak Velocity sub movement 2, Peak Velocity sub

movement 1, time of Peak Acceleration sub movement 1, Peak Acceleration sub movement 1, time of Peak Acceleration sub movement 2, Peak Deceleration sub movement 2, time of Peak Velocity sub movement 1, Movement Units sub movement 1, time of Peak Deceleration sub movement 1, Peak Deceleration sub movement 1, Total Duration sub movement 1.

Findings reported in this subsection may suggest that (automatic) identification of children with ASD could be performed by using a limited set of kinematic features. Furthermore, this could even lead to the hypothesis of a possible motor signature of ASD related to disrupted planning movement sequences.

3.5 Scalability, computational efficiency and use of cloud computing

The operational time required by the whole pre-processing and training of the classifier (including feature extraction and feature selection) for the classification of AD versus CN (299 subjects), MCIc versus CN (238 subjects) and MCIc versus MCInc (210 subjects) was 31.7, 21.7 and 21.2 seconds, respectively.

The operational time required by the testing phase (including preprocessing and classification of the new dataset) was 1.5 seconds per subject on average.

Both training and testing were performed using the optimal parameters computed in subsection 3.2.4 and reported in Table 3.2.1.

One of the critical aspects of advanced algorithms in medicine, as highlighted in Chapter 1, regards the translation from bench to bedside, i.e., from research to clinical practice. Two questions from the data reported just above may arise: *what would be the operational costs using larger (big data) datasets?* and *what would be the operational costs in the daily clinical practice, i.e., if no advanced computing systems are available?* These questions may remain unanswered, unless new efforts for improving the computational power of both advanced (for big data research) and common (for use in clinical practice) systems are made.

However, it is worth to underline here that the use of centralized web-based services in general may solve the problem of the translation of advanced competence (i.e., ML algorithms) from research to clinic, this problem consisting of two main issues: 1) the high operational costs; 2) the need of experience in developing and implementing these advanced algorithms. All this would be bypassed by the use of a unique centralized system 1) able to deal with big data and 2) managed by experienced researchers in this field.

In order to partially answer to these needs, during this thesis I have been developing a web-based service for the early diagnosis of AD. This service is entirely based on data and methods used in Application II (Sections 2.3 and 3.2), and it is able to perform single-subject prediction of membership to the pathological (AD, MCIc) or healthy (MCInc, CN) class using structural MRI data.

The web-based service is hosted by the servers of our laboratory, which consist of a system with 32 CPUs running at 2.00 GhZ, as described in Section 2.5. The user-side workflow of the service consists of 3 steps: a) login, b) patient information entry and c) upload of MRI data. A schematic user-side workflow of the service is shown in Figure 3.5.1, while a screenshot of the *Upload* page is reported in Figure 3.5.2.

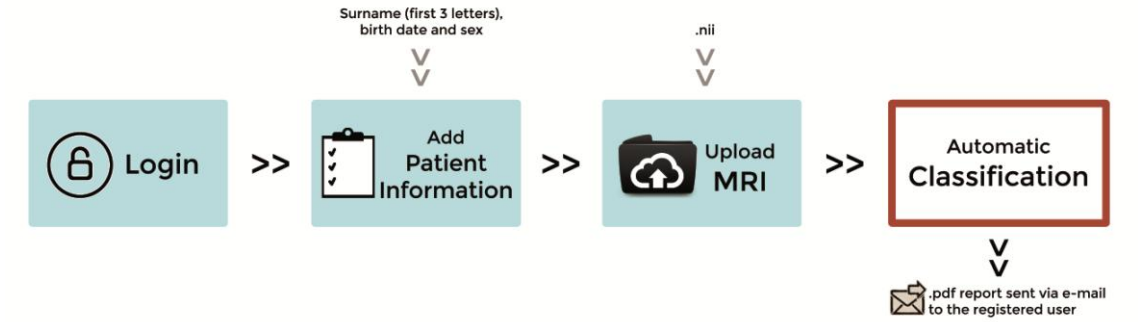


Figure 3.5.1 Schematic user-side workflow of the service

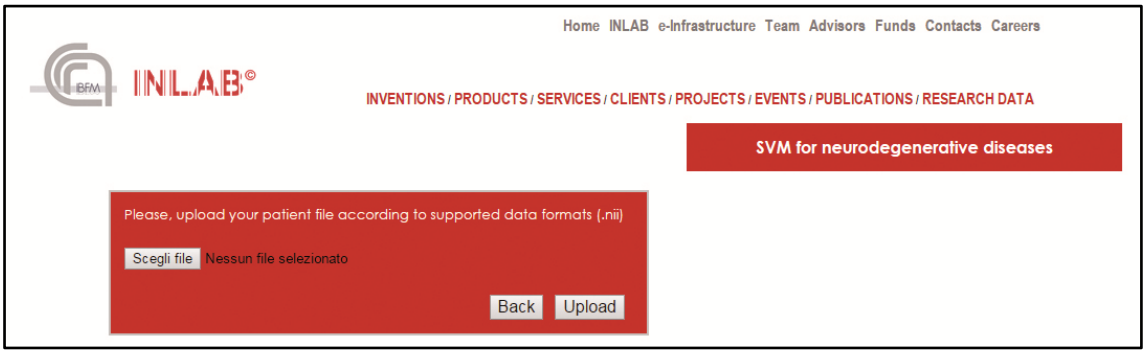


Figure 3.5.2 Screenshot of the *Upload* page

The service was tested by clinicians of the Institute of Neurology, University *Magna Graecia*, Catanzaro, and it can be found on the website of our laboratory (http://inlab.ibfm.cnr.it/svm_service.php), where it is available, after registration, for use in clinical practice.

Chapter 4.
CONCLUSIONS AND OUTLOOK

During this thesis I developed and implemented a ML method able to classify patients by means of their structural MRI data. The efficacy of this method was tested on different pathologies.

Specifically, we were able to perform differential diagnosis of PD (see Application I, Sections 2.2 and 3.1), early diagnosis and prognosis of AD (see Application II, Section 2.3 and 3.2) and diagnosis of ED (see Application III, Sections 2.4 and 3.3). Classification performances achieved in these applications resulted to be significantly higher than chance, but also higher than or comparable to those achieved by other automatic classification methods (see Chapter 3 for details).

A distinctive characteristic of the implemented ML method was the possibility of computing which of the features (in this case, which voxels) given as input resulted to be the most significant for group discrimination, thus allowing the generation of image maps of voxel-based pattern distribution of brain structural differences between two given groups or conditions. The importance of this further analysis is given by the possibility of identifying potential biomarkers for the diagnosis of a pathology. This analysis was conducted for all reported applications (PD, AD and ED), returning results in line with previously published studies (Application I and Application II), but also highlighting new possible biomarkers for the diagnosis of the considered pathologies (Application II and Application III), as detailed in Chapter 3.

Furthermore, this method was adapted to be used for the automatic diagnosis of ASD through kinematic features collected during a movement task. Also in this case, classification performance resulted to be high.

At the end of this work, I have also been developing a web-based service for the early diagnosis of AD, that is entirely based on the methods developed in this thesis and that makes use of data collected in Application II (Sections 2.3 and 3.2). This service is hosted by the servers of our laboratory and it is already available on the website http://inlab.ibfm.cnr.it for use in clinical practice.

The implemented ML methods described in this whole thesis achieved high performance in general, but it cannot surely be the best classifier for all situations and for all pathologies. In this sense, it would be of great interest to study the applications of this method to other pathologies. Another point of interest would be the adaptation of the implemented method to work with data coming from other modalities than MRI, in order to explore its discriminative power when coupled with different sets of data, as already performed for ASD.

As mentioned in Chapters 1 and 2, the implemented ML method is only able to perform binary classification, this feature directly descending from a core characteristic of the SVM algorithm. Because of this, another possible way to improve the implemented ML algorithm would be the introduction of the possibility to perform classification among more than two groups (e.g. direct classification among PD, PSP and CN). This could be achieved through the construction and validation of decision algorithm based on the One-Versus-One (OVO) or One-Versus-All (OVA) strategies. A derivation of this technique was already used during the implementation of the web-based service.

Further improvements could be performed on the whole algorithm of classification by improving single phases of the ML algorithm separately. For example, it could be useful to explore other methods to extract or select significant features from data.

Finally, given the implemented ML method for the prediction of the belonging class of unseen subjects, it would be very interesting to implement the possibility of extracting the belonging probability of each subject to his predicted class. In this way, such an algorithm would not only be able to perform (early) diagnosis and prognosis, but it would also allow to study the staging of a pathology in a patient, this being one of the main issues of personalized medicine.

ACKNOWLEDGEMENTS

I would like to take this opportunity to express my thanks to the supervisors of my work during these years, Isabella Castiglioni and Marco Paganoni. I would also like to thank the director of the Institute of Molecular Bioimaging and Physiology of the National Research Council (IBFM-CNR), Maria Carla Gilardi, as well as the coordinator of the PhD school in Physics of the University of Milano-Bicocca, Giberto Chirico.

Many thanks to my colleagues at the Laboratory of Innovation and Integration in Molecular Imaging (INLAB), and especially to Claudia, Francesca, Matteo and Petronilla.

Thanks also to Antonio Cerasa and Alessandro Crippa for their useful support to my work during the whole period of thesis, and to Miriam López for her help in the implementation of the SVM classifier.

BIBLIOGRAPHY

Almeida, J. R. C., Mourao-Miranda, J., Aizenstein, H. J., Versace, A., Kozel, F. A., Lu, H., ... & Phillips, M. L. (2013). Pattern recognition analysis of anterior cingulate cortex blood flow to classify depression polarity. *The British Journal of Psychiatry*, *203*(4), 310-311.

Álvarez, I., Górriz, J. M., Ramírez, J., Salas-Gonzalez, D., López, M., Segovia, F., ... & Prieto, B. (2009). Alzheimer's diagnosis using eigenbrains and support vector machines. In *Bio-Inspired Systems: Computational and Ambient Intelligence* (pp. 973-980). Springer Berlin Heidelberg.

Amaral, D. G., Schumann, C. M., & Nordahl, C. W. (2008). Neuroanatomy of autism. *Trends in neurosciences*, *31*(3), 137-145.

Ament, K., Mejia, A., Buhlman, R., Erklin, S., Caffo, B., Mostofsky, S., & Wodka, E. (2015). Evidence for specificity of motor impairments in catching and balance in children with autism. *Journal of autism and developmental disorders*, *45*, 742-751.

Amianto, F., D'Agata, F., Lavagnino, L., Caroppo, P., Abbate-Daga, G., Righi, D., ... & Fassino, S. (2013). Intrinsic connectivity networks within cerebellum and beyond in eating disorders. *The Cerebellum*, *12*(5), 623-631.

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

Avena, N. M., & Bocarsly, M. E. (2012). Dysregulation of brain reward systems in eating disorders: neurochemical information from animal models of binge eating, bulimia nervosa, and anorexia nervosa. *Neuropharmacology*, *63*(1), 87-96.

Barnett, A. L., Guzzetta, A., Mercuri, E., Henderson, S. E., Haataja, L., Cowan, F., & Dubowitz, L. (2004). Can the Griffiths scales predict neuromotor and perceptual-motor impairment in term infants with neonatal encephalopathy?. *Archives of disease in childhood*, *89*(7), 637-643.

Berlucchi, G., & Aglioti, S. M. (2010). The body in the brain revisited.*Experimental brain research*, *200*(1), 25-35.

Bernstein, E. M., & Putnam, F. W. (1986). Development, reliability, and validity of a dissociation scale. *The Journal of nervous and mental disease*, *174*(12), 727-735.

Blennow, K., de Leon, M. J., & Zetterberg, H. (2006). Alzheimer's disease. *The Lancet. 530 368(9533), 387-403*.

Boghi, A., Sterpone, S., Sales, S., D'Agata, F., Bradac, G. B., Zullo, G., & Munno, D. (2011). In vivo evidence of global and focal brain alterations in anorexia nervosa. *Psychiatry Research: Neuroimaging*, *192*(3), 154-159.

Bonneville, M., Meunier, J., Bengio, Y., & Soucy, J. P. (1998). Support vector machines for improving the classification of brain PET images. In*Medical Imaging'98* (pp. 264-273). International Society for Optics and Photonics.

Braak, H., Ghebremedhin, E., Rüb, U., Bratzke, H., & Del Tredici, K. (2004). Stages in the development of Parkinson's disease-related pathology. *Cell and tissue research*, *318*(1), 121-134.

Braak, H., Del Tredici, K., Rüb, U., de Vos, R. A., Steur, E. N. J., & Braak, E. (2003). Staging of brain pathology related to sporadic Parkinson's disease.*Neurobiology of aging*, *24*(2), 197-211.

Braak, H., & Braak, E. (1991). Neuropathological stageing of Alzheimer-related changes. *Acta neuropathologica*, *82*(4), 239-259.

Brian, J., Bryson, S. E., Garon, N., Roberts, W., Smith, I. M., Szatmari, P., & Zwaigenbaum, L. (2008). Clinical assessment of autism in high-risk 18-month-olds. *Autism*, *12*(5), 433-456.

Brooks, S. J., Rask-Andersen, M., Benedict, C., & Schiöth, H. B. (2012). A debate on current eating disorder diagnoses in light of neurobiological

findings: is it time for a spectrum model?. *BMC psychiatry*, *12*(1), 76.

Brooks, S. J., O'Daly, O. G., Uher, R., Friederich, H. C., Giampietro, V., Brammer, M., ... & Campbell, I. C. (2011). Differential neural responses to food images in women with bulimia versus anorexia nervosa. *PLoS One*, *6*(7), e22259.

Busatto, G. F., Garrido, G. E., Almeida, O. P., Castro, C. C., Camargo, C. H., Cid, C. G., ... & Bottino, C. M. (2003). A voxel-based morphometry study of temporal lobe gray matter reductions in Alzheimer's disease. *Neurobiology of aging*, *24*(2), 221-231.

Canu, E., Agosta, F., Baglio, F., Galantucci, S., Nemni, R., & Filippi, M. (2011). Diffusion tensor magnetic resonance imaging tractography in progressive supranuclear palsy. *Movement Disorders*, *26*(9), 1751-1755.

Castellani, U., Rossato, E., Murino, V., Bellani, M., Rambaldelli, G., Perlini, C., ... & Brambilla, P. (2012). Classification of schizophrenia using feature-based morphometry. *Journal of neural transmission*, *119*(3), 395-404.

Cattaneo, L., Fabbri-Destro, M., Boria, S., Pieraccini, C., Monti, A., Cossu, G., & Rizzolatti, G. (2007). Impairment of actions chains in autism and its possible role in intention understanding. *Proceedings of the National Academy of Sciences*, *104*(45), 17825-17830.

Cerasa, A., Castiglioni, I., Salvatore, C., Funaro, A., Martino, I., Alfano, S., & Quattrone, A. (2015). Biomarkers of Eating Disorders Using Support Vector Machine Analysis of Structural Neuroimaging Data: Preliminary Results. *Behavioural Neurology, 2015*. doi:10.1155/2015/924814.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, *20*(3), 273-297.

Crippa, A., Salvatore, C., Perego, P., Forti, S., Nobile, M., Molteni, M., & Castiglioni, I. (2015). Use of Machine Learning to Identify Children with Autism and Their Motor Abnormalities. *Journal of autism and developmental disorders*, 1-11.

Crippa, A., Forti, S., Perego, P., & Molteni, M. (2013). Eye-hand coordination in children with high functioning autism and Asperger's disorder using a gap-overlap paradigm. *Journal of autism and developmental disorders*, *43*(4), 841-850.

Cuingnet, R., Gerardin, E., Tessieras, J., Auzias, G., Lehéricy, S., Habert, M. O., ... & Alzheimer's Disease Neuroimaging Initiative. (2011). Automatic classification of patients with Alzheimer's disease from structural MRI: a comparison of ten methods using the ADNI database. *neuroimage*, *56*(2), 766-781.

Davatzikos, C., Fan, Y., Wu, X., Shen, D., & Resnick, S. M. (2008). Detection of prodromal Alzheimer's disease via pattern classification of magnetic resonance imaging. *Neurobiology of aging*, *29*(4), 514-523.

Del Tredici, K., Rüb, U., de Vos, R. A., Bohl, J. R., & Braak, H. (2002). Where does parkinson disease pathology begin in the brain?. *Journal of Neuropathology & Experimental Neurology*, *61*(5), 413-426.

Dowd, A. M., McGinley, J. L., Taffe, J. R., & Rinehart, N. J. (2012). Do planning and visual integration difficulties underpin motor dysfunction in autism? A kinematic study of young children with autism. *Journal of autism and developmental disorders*, *42*(8), 1539-1548.

DSM-4 American Psychiatric Association. (2000). Diagnostic and statistical manual of mental disorders (4th ed., text rev.). *Washington, DC: American Psychiatric Publishing.*

DSM-5 American Psychiatric Association. (2013). Diagnostic and statistical manual of mental disorders. *Arlington: American Psychiatric Publishing*.

Dubois, B., Feldman, H. H., Jacova, C., DeKosky, S. T., Barberger-Gateau, P., Cummings, J., ... & Scheltens, P. (2007). Research criteria for the diagnosis of Alzheimer's disease: revising the NINCDS–ADRDA criteria. *The Lancet Neurology*, *6*(8), 734-746.

Eynde, F., Suda, M., Broadbent, H., Guillaume, S., Eynde, M., Steiger, H., ... & Schmidt, U. (2012). Structural Magnetic Resonance Imaging in Eating Disorders: A Systematic Review of Voxel-Based Morphometry Studies.*European Eating Disorders Review*, *20*(2), 94-105.

Fabbri-Destro, M., Cattaneo, L., Boria, S., & Rizzolatti, G. (2009). Planning actions in autism. *Experimental Brain Research*, *192*(3), 521-525.

Fairburn, C. G., & Harrison, P. J. (2003). Eating disorders. *The Lancet*,*361*(9355), 407-416.

Favaro, A., Santonastaso, P., Manara, R., Bosello, R., Bommarito, G., Tenconi, E., & Di Salle, F. (2012). Disruption of visuospatial and somatosensory functional connectivity in anorexia nervosa. *Biological Psychiatry*, *72*(10), 864-870.

Focke, N. K., Helms, G., Scheewe, S., Pantel, P. M., Bachmann, C. G., Dechent, P., ... & Trenkwalder, C. (2011). Individual voxel-based subtype prediction can differentiate progressive supranuclear palsy from idiopathic parkinson syndrome and healthy controls. *Human brain mapping*, *32*(11), 1905-1915.

Forti, S., Valli, A., Perego, P., Nobile, M., Crippa, A., & Molteni, M. (2011). Motor planning and control in autism. A kinematic analysis of preschool children. *Research in Autism Spectrum Disorders*, *5*(2), 834-842.

Fournier, K. A., Hass, C. J., Naik, S. K., Lodha, N., & Cauraugh, J. H. (2010). Motor coordination in autism spectrum disorders: a synthesis and meta-

analysis. *Journal of autism and developmental disorders*, *40*(10), 1227-1240.

Fox, N. C., & Schott, J. M. (2004). Imaging cerebral atrophy: normal ageing to Alzheimer's disease. *The Lancet*, *363*(9406), 392-394.

Freitag, C. M., Kleser, C., Schneider, M., & von Gontard, A. (2007). Quantitative assessment of neuromotor function in adolescents with high functioning autism and Asperger syndrome. *Journal of autism and developmental disorders*, *37*(5), 948-959.

Friederich, H. C., Wu, M., Simon, J. J., & Herzog, W. (2013). Neurocircuit function in eating disorders. *International Journal of Eating Disorders*, *46*(5), 425-432.

Garg, A. X., Adhikari, N. K., McDonald, H., Rosas-Arellano, M. P., Devereaux, P. J., Beyene, J., ... & Haynes, R. B. (2005). Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: a systematic review. *Jama*, *293*(10), 1223-1238.

Garner, D. M. (1991). Eating Disorder Inventory-2: Professional manual. Florida: Psychological Assessment Resources.

Garner, D. M., & Garfinkel, P. E. (1979). The Eating Attitudes Test: An index of the symptoms of anorexia nervosa. *Psychological medicine*, *9*(02), 273-279.

Gaser, C., Volz, H. P., Kiebel, S., Riehemann, S., & Sauer, H. (1999). Detecting structural changes in whole brain based on nonlinear deformations—application to schizophrenia research. *Neuroimage*, *10*(2), 107-113.

Gaudio, S., & Quattrocchi, C. C. (2012). Neural basis of a multidimensional model of body image distortion in anorexia nervosa. *Neuroscience & Biobehavioral Reviews*, *36*(8), 1839-1847.

Gelb, D. J., Oliver, E., & Gilman, S. (1999). Diagnostic criteria for Parkinson disease. *Archives of neurology*, *56*(1), 33-39.

Glazebrook, C. M., Elliott, D., & Lyons, J. (2006). A kinematic analysis of how young adults with and without autism plan and control goal-directed movements. *MOTOR CONTROL-CHAMPAIGN-*, *10*(3), 244.

Glazebrook, C., Gonzalez, D., Hansen, S., & Elliott, D. (2009). The role of vision for online control of manual aiming movements in persons with autism spectrum disorders. *Autism*, *13*(4), 411-433.

Grabner, G., Janke, A. L., Budge, M. M., Smith, D., Pruessner, J., & Collins, D. L. (2006). Symmetric atlasing and model based segmentation: an application to the hippocampus in older adults. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2006* (pp. 58-66). Springer Berlin Heidelberg.

Griffiths, R. (1970). The ability of young children. A study in mental measurement.

Gunn, S. R. (1998). Support vector machines for classification and regression.*ISIS technical report*, *14*.

Habeck, C., Foster, N. L., Perneczky, R., Kurz, A., Alexopoulos, P., Koeppe, R. A., ... & Stern, Y. (2008). Multivariate and univariate neuroimaging biomarkers of Alzheimer's disease. *Neuroimage*, *40*(4), 1503-1515.

Haller, S., Badoud, S., Nguyen, D., Barnaure, I., Montandon, M. L., Lovblad, K. O., & Burkhard, P. R. (2013). Differentiation between Parkinson disease and other forms of Parkinsonism using support vector machine analysis of susceptibility-weighted imaging (SWI): initial results. *European radiology*, *23*(1), 12-19.

Haller, S., Badoud, S., Nguyen, D., Garibotto, V., Lovblad, K. O., & Burkhard, P. R. (2012). Individual detection of patients with Parkinson disease using support vector machine analysis of diffusion tensor imaging data: initial results.*American Journal of Neuroradiology*, *33*(11), 2123-2128.

Hathaway, S. R., & McKinley, J. C. (1943). The Minnesota multiphasic personality inventory (Rev. ed., 2nd printing.).

Haufe, S., Meinecke, F., Görgen, K., Dähne, S., Haynes, J. D., Blankertz, B., & Bießmann, F. (2014). On the interpretation of weight vectors of linear models in multivariate neuroimaging. *Neuroimage*, *87*, 96-110.

Hidalgo-Muñoz, A. R., Ramírez, J., Górriz, J. M., & Padilla, P. (2014). Regions of interest computed by SVM wrapped method for Alzheimer's disease examination from segmented MRI. *Frontiers in aging neuroscience*, *6*.

Hoek, H. W., & Van Hoeken, D. (2003). Review of the prevalence and incidence of eating disorders. *International Journal of eating disorders*, *34*(4), 383-396.

Husain, M. M., Black, K. J., Doraiswamy, P. M., Shah, S. A., Rockwell, W. K., Ellinwood, E. H., & Krishnan, K. R. R. (1992). Subcortical brain anatomy in anorexia and bulimia. *Biological Psychiatry*, *31*(7), 735-738.

Hyman, B. T., & Trojanowski, J. Q. (1997). Editorial on consensus recommendations for the postmortem diagnosis of Alzheimer disease from the National Institute on Aging and the Reagan Institute Working Group on diagnostic criteria for the neuropathological assessment of Alzheimer disease.*Journal of Neuropathology & Experimental Neurology*, *56*(10), 1095-1097.

Izawa, J., Pekny, S. E., Marko, M. K., Haswell, C. C., Shadmehr, R., & Mostofsky, S. H. (2012). Motor Learning Relies on Integrated Sensory Inputs in ADHD, but Over-Selectively on Proprioception in

Autism Spectrum Conditions.*Autism Research*, *5*(2), 124-136.

Jack, C. R., Bernstein, M. A., Fox, N. C., Thompson, P., Alexander, G., Harvey, D., ... & Weiner, M. W. (2008). The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods. *Journal of Magnetic Resonance Imaging*, *27*(4), 685-691.

Jagust, W., Gitcho, A., Sun, F., Kuczynski, B., Mungas, D., & Haan, M. (2006). Brain imaging evidence of preclinical Alzheimer's disease in normal aging. *Annals of neurology*, *59*(4), 673-681.

Jarrold, C., & Brock, J. (2004). To match or not to match? Methodological issues in autism-related research. *Journal of autism and developmental disorders*, *34*(1), 81-86.

Jenkinson, M., Beckmann, C. F., Behrens, T. E., Woolrich, M. W., & Smith, S. M. (2012). Fsl. *Neuroimage*, *62*(2), 782-790.

Jovicich, J., Czanner, S., Greve, D., Haley, E., van der Kouwe, A., Gollub, R., ... & Dale, A. (2006). Reliability in multi-site structural MRI studies: effects of gradient non-linearity correction on phantom and human data. *Neuroimage*,*30*(2), 436-443.

Jubault, T., Brambati, S. M., Degroot, C., Kullmann, B., Strafella, A. P., Lafontaine, A. L., ... & Monchi, O. (2009). Regional brain stem atrophy in idiopathic Parkinson's disease detected by anatomical MRI. *PloS one*, *4*(12), e8247.

Kawamoto, K., Houlihan, C. A., Balas, E. A., & Lobach, D. F. (2005). Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success. *Bmj*,*330*(7494), 765.

Kaye, W. H., Fudge, J. L., & Paulus, M. (2009). New insights into symptoms and neurocircuit function of anorexia nervosa. *Nature Reviews Neuroscience*,*10*(8), 573-584.

Keating, C., Tilbrook, A. J., Rossell, S. L., Enticott, P. G., & Fitzgerald, P. B. (2012). Reward processing in anorexia nervosa. *Neuropsychologia*, *50*(5), 567-575.

King, J. A., Geisler, D., Ritschel, F., Boehm, I., Seidel, M., Roschinski, B., ... & Ehrlich, S. (2015). Global cortical thinning in acute anorexia nervosa normalizes following long-term weight restoration. *Biological psychiatry*, *77*(7), 624-632.

Klöppel, S., Stonnington, C. M., Chu, C., Draganski, B., Scahill, R. I., Rohrer, J. D., ... & Frackowiak, R. S. (2008). Automatic classification of MR scans in Alzheimer's disease. *Brain*, *131*(3), 681-689.

Knake, S., Belke, M., Menzler, K., Pilatus, U., Eggert, K. M., Oertel, W. H., ... & Höglinger, G. U. (2010). In vivo demonstration of microstructural brain pathology in progressive supranuclear palsy: a DTI study using TBSS.*Movement Disorders*, *25*(9), 1232-1238.

Knopman, D. S., DeKosky, S. T., Cummings, J. L., Chui, H., Corey–Bloom, J., Relkin, N., ... & Stevens, J. C. (2001). Practice parameter: Diagnosis of dementia (an evidence-based review) Report of the Quality Standards Subcommittee of the American Academy of Neurology. *Neurology*, *56*(9), 1143-1153.

Lai, M. C., Lombardo, M. V., & Baron-Cohen, S. (2014). Autism. *The Lancet, 383(9920), 896–910*.

Leary, M. R., & Hill, D. A. (1996). Moving on: autism and movement disturbance. *Mental retardation*, (34), 39-53.

Li, Q., Van Antwerp, D., Mercurio, F., Lee, K. F., & Verma, I. M. (1999). Severe liver degeneration in mice lacking the IκB kinase 2 gene. *Science*,*284*(5412), 321-325.

Litvan, I., Agid, Y., Calne, D., Campbell, G., Dubois, B., Duvoisin, R. C., ... & Zee, D. S. (1996). Clinical research criteria for the diagnosis of progressive supranuclear palsy (Steele-Richardson-Olszewski syndrome) Report of the NINDS-SPSP International Workshop*. *Neurology*, *47*(1), 1-9.

Lord, C., Risi, S., Lambrecht, L., Cook Jr, E. H., Leventhal, B. L., DiLavore, P. C., ... & Rutter, M. (2000). The Autism Diagnostic Observation Schedule—Generic: A standard measure of social and communication deficits associated with the spectrum of autism. *Journal of autism and developmental disorders*,*30*(3), 205-223.

Lord, C., Rutter, M., & Le Couteur, A. (1994). Autism Diagnostic Interview-Revised: a revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. *Journal of autism and developmental disorders*, *24*(5), 659-685.

Magnin, B., Mesrob, L., Kinkingnéhun, S., Pélégrini-Issac, M., Colliot, O., Sarazin, M., ... & Benali, H. (2009). Support vector machine-based classification of Alzheimer's disease from whole-brain anatomical MRI.*Neuroradiology*, *51*(2), 73-83.

Mahler, P., Guastavino, J. M., Jacquart, G., & Strazielle, C. (1993). An unexpected role of the cerebellum: involvement in nutritional organization.*Physiology & behavior*, *54*(6), 1063-1067.

Mahmoudi, A., Takerkart, S., Regragui, F., Boussaoud, D., & Brovelli, A. (2012). Multivoxel pattern analysis for fMRI data: A review. *Computational and mathematical methods in medicine*, *2012*.

Mari, M., Castiello, U., Marks, D., Marraffa, C., & Prior, M. (2003). The reach–to–grasp movement in children with autism spectrum disorder. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *358*(1430), 393-403.

Marsh, R., Stefan, M., Bansal, R., Hao, X., Walsh, B. T., & Peterson, B. S. (2013). Anatomical

characteristics of the cerebral surface in bulimia nervosa.*Biological psychiatry*.

Martin, C. R., Preedy, V. R., & Hunter, R. J. (Eds.). (2012). *Nanomedicine and the Nervous System*. CRC Press.

Massey, L. A., Jäger, H. R., Paviour, D. C., O'Sullivan, S. S., Ling, H., Williams, D. R., ... & Micallef, C. (2013). The midbrain to pons ratio A simple and specific MRI sign of progressive supranuclear palsy. *Neurology*, *80*(20), 1856-1861.

McKhann, G., Drachman, D., Folstein, M., Katzman, R., Price, D., & Stadlan, E. M. (1984). Clinical diagnosis of Alzheimer's disease Report of the NINCDS-ADRDA Work Group* under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease. *Neurology*, *34*(7), 939-939.

Messina, D., Cerasa, A., Condino, F., Arabia, G., Novellino, F., Nicoletti, G., ... & Quattrone, A. (2011). Patterns of brain atrophy in Parkinson's disease, progressive supranuclear palsy and multiple system atrophy. *Parkinsonism & related disorders*, *17*(3), 172-176.

Middleton, F. A., & Strick, P. L. (2001). Cerebellar projections to the prefrontal cortex of the primate. *The Journal of Neuroscience*, *21*(2), 700-712.

Miller, R. A. (1994). Medical diagnostic decision support systems--past, present, and future: a threaded bibliography and brief commentary. *Journal of the American Medical Informatics Association*, *1*(1), 8.

Minshew, N. J., Sung, K., Jones, B. L., & Furman, J. M. (2004). Underdevelopment of the postural control system in autism. *Neurology*, *63*(11), 2056-2061.

Mitchell, A. J., & Shiri-Feshki, M. (2009). Rate of progression of mild cognitive impairment to dementia–meta-analysis of 41 robust inception cohort studies.*Acta Psychiatrica Scandinavica*, *119*(4), 252-265.

Moradi, E., Pepe, A., Gaser, C., Huttunen, H., Tohka, J., & Alzheimer's Disease Neuroimaging Initiative. (2015). Machine learning framework for early MRI-based Alzheimer's conversion prediction in MCI subjects. *NeuroImage*,*104*, 398-412.

Narayana, P. A., Brey, W. W., Kulkarni, M. V., & Sievenpiper, C. L. (1988). Compensation for surface coil sensitivity variation in magnetic resonance imaging. *Magnetic resonance imaging*, *6*(3), 271-274.

Naselaris, T., Kay, K. N., Nishimoto, S., & Gallant, J. L. (2011). Encoding and decoding in fMRI. *Neuroimage*, *56*(2), 400-410.

Nielsen, S. (2001). Epidemiology and mortality of eating disorders. *Psychiatric Clinics of North America*, *24*(2), 201-214.Nielsen, S. (2001). Epidemiology and mortality of eating disorders. *Psychiatric Clinics of North America*, *24*(2), 201-214.

Nigro, S., Cerasa, A., Zito, G., Perrotta, P., Chiaravalloti, F., Donzuso, G., ... & Quattrone, A. (2014). Fully automated segmentation of the pons and midbrain using human T1 MR brain images. *PloS one*, *9*(1), e85618.

Nijenhuis, E. R., Spinhoven, P., Van Dyck, R., Van Der Hart, O., & Vanderlinden, J. (1996). The development and psychometric characteristics of the Somatoform Dissociation Questionnaire (SDQ-20). *The Journal of nervous and mental disease*, *184*(11), 688-694.

Nobile, M., Perego, P., Piccinini, L., Mani, E., Rossi, A., Bellina, M., & Molteni, M. (2011). Further evidence of complex motor dysfunction in drug naive children with autism using automatic motion analysis of gait. *Autism*, 1362361309356929.

Oba, H., Yagishita, A., Terada, H., Barkovich, A. J., Kutomi, K., Yamauchi, T., ... & Suzuki, S. (2005). New and reliable MRI diagnosis for progressive supranuclear palsy. *Neurology*, *64*(12), 2050-2055.

Orrù, G., Pettersson-Yeo, W., Marquand, A. F., Sartori, G., & Mechelli, A. (2012). Using support vector machine to identify imaging biomarkers of neurological and psychiatric disease: a critical review. *Neuroscience & Biobehavioral Reviews*, *36*(4), 1140-1152.

Osuna, E., Freund, R., & Girosi, F. (1997, June). Training support vector machines: an application to face detection. In *Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference on*(pp. 130-136). IEEE.

Parker, G., Tupling, H., & Brown, L. B. (1979). A parental bonding instrument.*British Journal of Medical Psychology*, *52*(1), 1-10.

Pereira, F., Mitchell, T., & Botvinick, M. (2009). Machine learning classifiers and fMRI: a tutorial overview. *Neuroimage*, *45*(1), S199-S209.

Pettersson-Yeo, W., Benetti, S., Marquand, A. F., Dell'Acqua, F., Williams, S. C. R., Allen, P., ... & Mechelli, A. (2013). Using genetic, cognitive and multi-modal neuroimaging data to identify ultra-high-risk and first-episode psychosis at the individual level. *Psychological medicine*, *43*(12), 2547-2562.

Quattrone, A., Nicoletti, G., Messina, D., Fera, F., Condino, F., Pugliese, P., ... & Gallo, O. (2008). MR imaging index for differentiation of progressive supranuclear palsy from parkinson disease and the parkinson variant of multiple system atrophy 1. *Radiology*, *246*(1), 214-221.

Rinehart, N., & McGINLEY, J. (2010). Is motor dysfunction core to autism spectrum disorder?. *Developmental Medicine & Child Neurology*,*52*(8), 697-697.

Salas-Gonzalez, D., Górriz, J. M., Ramírez, J., Illán, I. A., López, M., Segovia, F., ... & Puntonet, C. G. (2010). Feature selection using factor analysis for Alzheimer's diagnosis using F18-FDG PET images. *Medical physics*, *37*(11), 6084-6095.

Salvatore, C., Cerasa, A., Battista, P., Gilardi, M. C., Quattrone, A., Castiglioni, I., & Alzheimer's Disease Neuroimaging Initiative. (2015a). Magnetic resonance imaging biomarkers for the early diagnosis of Alzheimer's disease: a machine learning approach. *Frontiers in neuroscience*, *9*.

Salvatore, C., Battista, P., & Castiglioni, I. (2015b). Frontiers for the early diagnosis of AD by means of MRI brain imaging and Support Vector Machines. *Current Alzheimer Research*, *in press.*

Salvatore, C., Cerasa, A., Castiglioni, I., Gallivanone, F., Augimeri, A., Lopez, M., ... & Quattrone, A. (2014). Machine learning on brain MRI data for differential diagnosis of Parkinson's disease and Progressive Supranuclear Palsy. *Journal of neuroscience methods*, *222*, 230-237.

Schölkopf, B., Akritas, M. G., & Politis, D. N. (2003). An introduction to support vector machines.

Schroeter, M. L., Stein, T., Maslowski, N., & Neumann, J. (2009). Neural correlates of Alzheimer's disease and mild cognitive impairment: a systematic and quantitative meta-analysis involving 1351 patients. *Neuroimage*, *47*(4), 1196-1206.

Schrouff, J., Rosa, M. J., Rondina, J. M., Marquand, A. F., Chu, C., Ashburner, J., ... & Mourão-Miranda, J. (2013). PRoNTo: pattern recognition for neuroimaging toolbox. *Neuroinformatics*, *11*(3), 319-337.

Schulz, J. B., Skalej, M., Wedekind, D., Luft, A. R., Abele, M., Voigt, K., ... & Klockgether, T. (1999). Magnetic resonance imaging–based volumetry differentiates idiopathic Parkinson's syndrome from multiple system atrophy and progressive supranuclear palsy. *Annals of neurology*, *45*(1), 65-74.

Segura-García, C., Papaianni, M. C., Rizza, P., Flora, S., & De Fazio, P. (2012). The development and validation of the Body Image Dimensional Assessment (BIDA). *Eating and Weight Disorders-Studies on Anorexia, Bulimia and Obesity*, *17*(3), e219-e225.

Seitz, J., Bühren, K., von Polier, G. G., Heussen, N., Herpertz-Dahlmann, B., & Konrad, K. (2015). Morphological changes in the brain of acutely ill and weight-recovered patients with anorexia nervosa. *Zeitschrift für Kinder-und Jugendpsychiatrie und Psychotherapie*.

Shi, H. C., Zhong, J. G., Pan, P. L., Xiao, P. R., Shen, Y., Wu, L. J., ... & Li, H. Y. (2013). Gray matter atrophy in progressive supranuclear palsy: meta-analysis of voxel-based morphometry studies. *Neurological Sciences*, *34*(7), 1049-1055.

Smith, S. M., Jenkinson, M., Woolrich, M. W., Beckmann, C. F., Behrens, T. E., Johansen-Berg, H., ... & Matthews, P. M. (2004). Advances in functional and structural MR image analysis and implementation as FSL. *Neuroimage*, *23*, S208-S219.

Sperling, R. A., Aisen, P. S., Beckett, L. A., Bennett, D. A., Craft, S., Fagan, A. M., ... & Phelps, C. H. (2011). Toward defining the preclinical stages of Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's & dementia*, *7*(3), 280-292.

Steele, J. C., Richardson, J. C., & Olszewski, J. (1964). Progressive supranuclear palsy: a heterogeneous degeneration involving the brain stem, basal ganglia and cerebellum with vertical gaze and pseudobulbar palsy, nuchal dystonia and dementia. *Archives of neurology*, *10*(4), 333-359.

Suchan, B., Bauser, D. S., Busch, M., Schulte, D., Grönemeyer, D., Herpertz, S., & Vocks, S. (2013). Reduced connectivity between the left fusiform body area and the extrastriate body area in anorexia nervosa is associated with body image distortion. *Behavioural brain research*, *241*, 80-85.

Suchan, B., Busch, M., Schulte, D., Grönermeyer, D., Herpertz, S., & Vocks, S. (2010). Reduction of gray matter density in the extrastriate body area in women with anorexia nervosa. *Behavioural brain research*, *206*(1), 63-67.

Taki, Y., Kinomura, S., Sato, K., Inoue, K., Goto, R., Okada, K., ... & Fukuda, H. (2008). Relationship between body mass index and gray matter volume in 1,428 healthy individuals. *Obesity*, *16*(1), 119-124.

Teitelbaum, P., Teitelbaum, O., Nye, J., Fryman, J., & Maurer, R. G. (1998). Movement analysis in infancy may be useful for early diagnosis of autism.*Proceedings of the National Academy of Sciences*, *95*(23), 13982-13987.

Thomann, P. A., Schläfer, C., Seidl, U., Dos Santos, V., Essig, M., & Schröder, J. (2008a). The cerebellum in mild cognitive impairment and Alzheimer's disease–a structural MRI study. *Journal of psychiatric research*,*42*(14), 1198-1202.

Thomann, P. A., Toro, P., Dos Santos, V., Essig, M., & Schröder, J. (2008b). Clock drawing performance and brain morphology in mild cognitive impairment and Alzheimer's disease. *Brain and cognition*, *67*(1), 88-93.

Titova, O. E., Hjorth, O. C., Schiöth, H. B., & Brooks, S. J. (2013). Anorexia nervosa is linked to reduced brain structure in reward and somatosensory regions: a meta-analysis of VBM studies. *BMC psychiatry*, *13*(1), 110.

Tolosa, E., Wenning, G., & Poewe, W. (2006). The diagnosis of Parkinson's disease. *The Lancet Neurology*, *5*(1), 75-86.

Treasure, J., Claudino, A. M. & Zucker, N. (2010). Eating disorders. *The Lancet, vol. 375, no. 9714, pp. 583–593*.

Tufail, A. B., Abidi, A., Siddiqui, A. M., & Younis, M. S. (2012). Automatic classification of initial categories of Alzheimer's disease from structural MRI phase images: a comparison of PSVM, KNN and ANN methods. *Age*, *75*(8.96), 76-13.

Van den Eynde, F., & Treasure, J. (2009). Neuroimaging in eating disorders and obesity: implications for research. *Child and adolescent psychiatric clinics of North America*, *18*(1), 95-115.

Van Kuyck, K., Gérard, N., Van Laere, K., Casteels, C., Pieters, G., Gabriëls, L., & Nuttin, B. (2009). Towards a neurocircuitry in anorexia nervosa: evidence from functional neuroimaging studies. *Journal of psychiatric research*, *43*(14), 1133-1145.

Van Waelvelde, H., Oostra, A., Dewitte, G., Van den Broeck, C., & Jongmans, M. J. (2010). Stability of motor problems in young children with or at risk of autism spectrum disorders, ADHD, and or developmental coordination disorder.*Developmental Medicine & Child Neurology*, *52*(8), e174-e178.

Wang, J., Xu, G., Gonzales, V., Coonfield, M., Fromholt, D., Copeland, N. G., ... & Borchelt, D. R. (2002). Fibrillar inclusions and motor neuron degeneration in transgenic mice expressing superoxide dismutase 1 with a disrupted copper-binding site. *Neurobiology of disease*, *10*(2), 128-138.

Wechsler, D. (1987). *WMS-R: Wechsler memory scale-revised*. Psychological Corporation.

Wegiel, J., Wang, K. C., Tarnawski, M., & Lach, B. (2000). Microglial cells are the driving force in fibrillar plaque formation, whereas astrocytes are a leading factor in plaque degradation. *Acta neuropathologica*, *100*(4), 356-364.

Whyatt, C., & Craig, C. (2013). Sensory-motor problems in Autism. *Frontiers in integrative neuroscience*, *7*.

Zhu, J. N., & Wang, J. J. (2008). The cerebellum in feeding control: possible function and mechanism. *Cellular and molecular neurobiology*, *28*(4), 469-478.

PUBLICATIONS

The work described in this thesis (from 2012 to 2015) led to the publications listed below. For ISI International Papers, the whole manuscripts are attached at the end of this list.

ISI International Papers

1. Gallivanone, F., Canevari, C., Gianolli, L., **Salvatore, C.**, Della Rosa, P. A., Gilardi, M. C., & Castiglioni, I. (2013). A partial volume effect correction tailored for 18 F-FDG-PET oncological studies. *BioMed research international*, *2013*.
2. **Salvatore, C.**, Cerasa, A., Castiglioni, I., Gallivanone, F., Augimeri, A., Lopez, M., ... & Quattrone, A. (2014). Machine learning on brain MRI data for differential diagnosis of Parkinson's disease and Progressive Supranuclear Palsy. *Journal of neuroscience methods*, *222*, 230-237.
3. Canini, M., Battista, P., Della Rosa, P. A., Catricalà, E., **Salvatore, C.**, Gilardi, M. C., & Castiglioni, I. (2014). Computerized Neuropsychological Assessment in Aging: Testing Efficacy and Clinical Ecology of Different Interfaces. *Computational and mathematical methods in medicine*, 2014.
4. Crippa, A., **Salvatore, C.**, Perego, P., Forti, S., Nobile, M., Molteni, M., & Castiglioni, I. (2015). Use of Machine Learning to Identify Children with Autism and Their Motor Abnormalities. *Journal of autism and developmental disorders*, 45(7), 2146-2156. doi: 10.1007/s10803-015-2379-8.
5. **Salvatore, C.**, Cerasa, A., Battista, P., Gilardi, M.C., Quattrone, A., & Castiglioni, I. (2015). Magnetic Resonance Imaging biomarkers for the early diagnosis of Alzheimer's Disease: a machine learning approach. *Frontiers in Neuroscience*. doi: 10.3389/fnins.2015.00307.
6. Cerasa, A., Castiglioni, I., **Salvatore, C.**, Funaro, A., Martino, I., Alfano, S., & Quattrone, A. (2015). Biomarkers of Eating Disorders Using Support Vector Machine Analysis of Structural Neuroimaging Data: Preliminary Results. *Behavioural Neurology*, *2015*. doi:10.1155/2015/924814.
7. **Salvatore, C.**, Battista, P., & Castiglioni, I. (2015). Frontiers for the early diagnosis of AD by means of MRI brain imaging and Support Vector Machines. *Current Alzheimer Research, in press.*

Indexed International Papers

8. Grosso, E., López, M., **Salvatore, C.**, Gallivanone, F., Di Grigoli, G., Valtorta, S., ... & Castiglioni, I. (2012, August). A Decision Support System for the assisted diagnosis of brain tumors: A feasibility study for 18 F-FDG PET preclinical studies. *In Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE*, pp. 6255-6258.
9. Gallivanone, F., Di Grigoli, G., **Salvatore, C.**, Valtorta, S., Gilardi, M. C., Moresco, R. M., & Castiglioni, I. (2012, October). Acute stress studies in rats by 18 FDG PET and SPM. *In Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC), 2012 IEEE,* pp. 2886-2889.
10. Cava, C., Zoppis, I., Mauri, G., Ripamonti, M., Gallivanone, F., **Salvatore, C.**, ... & Castiglioni, I. (2013, July). Combination of gene expression and genome copy number alteration has a prognostic value for breast cancer. *In Engineering in Medicine and Biology Society (EMBC), 2013 35th Annual International Conference of the IEEE*, pp. 608-611.

International Book Chapters

11. Cava, C., Gallivanone, F., **Salvatore, C.**, Della Rosa, P. and Castiglioni, I. (2014). Bioinformatics clouds for high-throughput technologies. *Handbook of Research on Cloud Infrastructures for Big Data Anaytics*. IGI Global: 489-507. doi: 10.4018/978-1-4666-5864-6.ch020.

International Conference Proceedings

12. Gallivanone, F., Di Grigoli, G., **Salvatore, C.**, Valtorta, S., Grosso, E., Gilardi, M.C., Moresco, R.M. and Castiglioni, I. (2012). SPM for activation studies in rats on stress conditions. *6th Hot Topics in Molecular Imaging Conference (TOPIM),* p.44.
13. Gallivanone, F., Di Grigoli, G., **Salvatore, C.**, Belloli, S., Valtorta, S., Raccagni, I., Gilardi, M.C., Castiglioni, I. and Moresco, R.M. (2013). Feasibility of supervised machine learning technique for assisted diagnosis of cancer: application to PET studies in small animals. *European Molecular Imaging Meeting (EMIM), 8th annual meeting of the European Society for Molecular Imaging (ESMI)*, Turin, Italy.
14. Castiglioni, I., Cerasa, A., **Salvatore, C.**, Gallivanone, F., Augimeri, A., Lopez, M., ... & Quattrone, A. (2013, June). Machine learning performs differential individual diagnosis of PD and PSP by brain MRI studies. *In MOVEMENT DISORDERS*, vol. 28, pp. S71-S72. 111 RIVER ST, HOBOKEN 07030-5774, NJ USA: WILEY-BLACKWELL.
15. **Salvatore, C.**, Cerasa, A., Battista, P., Gilardi, M.C., Quattrone, A., & Castiglioni, I. (2015). Neuroimaging biomarkers predicting conversion to AD. *2nd meeting of the Scientific Council of NeuroMI*, Milan.
16. Battista, P., **Salvatore, C.**, Gilardi, M.C., & Castiglioni, I. (2015). Neuropsychological testing and artificial intelligence for early and differential diagnosis of dementia. *2nd meeting of the Scientific Council of NeuroMI*, Milan.

National Conference Proceedings

17. Gallivanone, F., Grosso, E., Di Grigoli, G., **Salvatore, C.**, Valtorta, S., Gilardi, M.C., Moresco, R.M. and Castiglioni, I. (2012). Statistical Parametric Mapping for activation studies in rats. *Atti del Congresso Nazionale di Bioingegneria*, p.147. ISBN: 978 88 555 3182-5147.

PUBLICATION I

Application of the implemented ML method to the differential diagnosis of PD (Application I)

Journal of Neuroscience Methods 222 (2014) 230–237

Contents lists available at ScienceDirect

Journal of Neuroscience Methods

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

Clinical Neuroscience

Machine learning on brain MRI data for differential diagnosis of Parkinson's disease and Progressive Supranuclear Palsy

C. Salvatore[a], A. Cerasa[b], I. Castiglioni[c,*], F. Gallivanone[c], A. Augimeri[b], M. Lopez[d], G. Arabia[e], M. Morelli[e], M.C. Gilardi[c], A. Quattrone[b,e]

[a] Department of Physics, University of Milan – Bicocca, Piazza della Scienza 3, 20126 Milan, Italy
[b] Neuroimaging Research Unit, Institute of Neurological Sciences, National Research Council, Germaneto, CZ, Italy
[c] Institute of Molecular Bioimaging and Physiology, National Research Council (IBFM-CNR), via F.lli Cervi 93, 20090 Segrate, MI, Italy
[d] DITEN, University of Genoa, Via Opera Pia 11A, 16145 Genoa, Italy
[e] Institute of Neurology, University "Magna Graecia", Germaneto, CZ, Italy

HIGHLIGHTS

- The algorithm allows individual differential diagnosis of PD and PSP by means of MR images.
- The algorithm does not require a priori hypotheses of where useful information may be coded in the images.
- Classification accuracy was significantly higher compared to other published methods.
- The algorithm was able to obtain voxel-based morphological biomarkers of PD and PSP.

ARTICLE INFO

Article history:
Received 1 July 2013
Received in revised form
14 November 2013
Accepted 17 November 2013

Keywords:
Support Vector Machine (SVM)
Parkinson's disease (PD)
Progressive Supranuclear Palsy (PSP)
Magnetic resonance imaging (MRI)
Machine learning

ABSTRACT

Background: Supervised machine learning has been proposed as a revolutionary approach for identifying sensitive medical image biomarkers (or combination of them) allowing for automatic diagnosis of individual subjects. The aim of this work was to assess the feasibility of a supervised machine learning algorithm for the assisted diagnosis of patients with clinically diagnosed Parkinson's disease (PD) and Progressive Supranuclear Palsy (PSP).
Method: Morphological T1-weighted Magnetic Resonance Images (MRIs) of PD patients (28), PSP patients (28) and healthy control subjects (28) were used by a supervised machine learning algorithm based on the combination of Principal Components Analysis as feature extraction technique and on Support Vector Machines as classification algorithm. The algorithm was able to obtain voxel-based morphological biomarkers of PD and PSP.
Results: The algorithm allowed individual diagnosis of PD versus controls, PSP versus controls and PSP versus PD with an Accuracy, Specificity and Sensitivity > 90%. Voxels influencing classification between PD and PSP patients involved midbrain, pons, corpus callosum and thalamus, four critical regions known to be strongly involved in the pathophysiological mechanisms of PSP.
Comparison with existing methods: Classification accuracy of individual PSP patients was consistent with previous manual morphological metrics and with other supervised machine learning application to MRI data, whereas accuracy in the detection of individual PD patients was significantly higher with our classification method.
Conclusions: The algorithm provides excellent discrimination of PD patients from PSP patients at an individual level, thus encouraging the application of computer-based diagnosis in clinical practice.

1. Introduction

Parkinson's disease (PD) is the second most common neurodegenerative disease affecting millions of people worldwide. The primary objective in the clinical practice of PD is to achieve an individual differential diagnosis, in order to tailor the best individual treatment. PD clinical diagnosis is particularly prone to errors (Tolosa et al., 2006), as an array of motor symptoms can also

* Corresponding author at: IBFM-CNR, via F.lli Cervi 93, 20090 Segrate, MI, Italy. Tel.: +39 02 21717511; fax: +39 02 21717558.
E-mail addresses: christian.salvatore@unimib.it (C. Salvatore), a.cerasa@unicz.it (A. Cerasa), isabella.castiglioni@ibfm.cnr.it (I. Castiglioni), miriamlp@ugr.es (M. Lopez).

0165-0270/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jneumeth.2013.11.016

be present in other parkinsonian conditions such as Progressive Supranuclear Palsy (PSP). For instance, PSP patients are clinically similar to PD patients, however, are less responsive to treatment and they have a more rapid disease progression. Among the various forms of parkinsonisms, PSP is one of the most difficult to clinically disentangle from idiopathic PD, particularly in early disease stages, when the typical clinical signs are not clearly evident (Gelb et al., 1999; Litvan et al., 1996).

To date, individual diagnosis of PSP is predominantly based on patients' clinical history where standard brain magnetic resonance imaging (MRI) protocols are routinely employed only to exclude concomitant diseases, resulting in poor diagnostic accuracy, sensitivity and specificity, given that images are only visually inspected (Tolosa et al., 2006; Jankovic, 2008). In the past 20 years, a considerable effort has been put into the development of advanced neuroimaging processing techniques in order to identify neuroimaging biomarkers which could be then used for enhancing the diagnostic confidence of clinical diagnosis. Although significant results have been obtained (Shi et al., 2013), most studies have reported differences between patients and controls only at a group level, thus with a very limited translation to an individual diagnosis in more clinical settings. For this reason, attention has recently been directed toward alternative approaches to the analyses of neuroimaging data.

In the last few years, there has been a growing interest within the neuroimaging community in classification methods, including machine-learning algorithms. These techniques are based on algorithms able to automatically extract multiple information from image sets without requiring a priori hypotheses of where this information may be coded in the images. The main aim of these methods is to classify individual structural or functional brain images by maximizing the distance between groups of images. Several studies have assessed the diagnostic value of these techniques, e.g. for the diagnosis of the Alzheimer's disease (Kloppel et al., 2008; Magnin et al., 2009) and Mild Cognitive Impairment (Teipel et al., 2007), and have showed promising findings.

The aim of this study is to implement a supervised machine-learning method able to perform individual differential diagnosis of PD and PSP by means of structural T1-weighted MRIs. This method was based on Principal Component Analysis (PCA) underlying the feature extraction technique (Habeck et al., 2008; Salas-Gonzalez et al., 2010) and on Support Vector Machines (SVMs) for classification purposes (Scholkopf and Smola, 2001; López et al., 2011). In order to identify potential MRI-related biomarkers useful for the diagnosis of PD and PSP, we also generated image maps of pattern distribution of brain structural differences, which reflect the importance of each image voxel for the SVM classification.

2. Materials and methods

2.1. Clinical and MRI studies

In this retrospective study we enrolled 56 patients and 28 healthy control subjects. All study procedures and ethical aspects were approved by the institutional review board. Written informed consent was obtained from all subjects.

All patients and the healthy control subjects were examined by neurologists with more than ten years of experience in movement disorders. Age at onset, disease duration and severity of symptoms, as assessed by the Unified Parkinson's Disease Rating Scale (UPDRS), and the Hoehn–Yahr (H&Y) stage, were recorded. The Mini Mental State Examination (MMSE) was used to assess general cognitive status. The group of 56 patients consisted of 28 patients with clinically diagnosed PD (Gelb et al., 1999) and 28 patients with clinical diagnosis of probable or possible PSP (Litvan et al., 1996). The healthy control subjects had no history of neurologic or psychiatric diseases, with normal neurological examinations. The 28 healthy control subjects were of similar age as both patient groups.

One brain structural MRI study was performed for each subject by a 1.5-T unit (Signa NV/I; GE Medical Systems, USA). MRI data were acquired using a 3D T1-weighted spoiled gradient echo sequence with the following parameters: TR = 15.2 ms; TE = 6.7 ms; flip angle = 15°; FOV = 24 cm. Slice thickness was of 1.2 mm and each slice had a resolution of 256 × 256 pixels. A T1-weighted 3D dataset was obtained for each subject. Motion artifacts were negligible for all scans by visual inspection.

2.2. The machine-learning method

A machine-learning method able to perform individual differential diagnosis of PD and PSP by means of structural T1-weighted Magnetic Resonance Images (MRI) was implemented. This method was based on Principal Component Analysis (PCA) as feature extraction technique and on Support Vector Machines (SVMs) as classification algorithm (Scholkopf and Smola, 2001; López et al., 2011). Both these phases were implemented using the Matlab platform (Matlab version R2011b, The MathWorks, Natick, MA).

2.2.1. Image pre-processing

Original datasets were cropped, re-oriented and converted from DICOM format to 3D NIfTI format using the dcm2nii tool included in the MRICron software (http://www.mccauslandcenter.sc.edu/mricro/mricron/). After that, the pre-processing procedure mainly consisted of 2 steps: (1) skull stripping, which was achieved using the BET tool of the FSL 4.1 software (Smith et al., 2004; Jenkinson et al., 2012), and (2) normalization to MNI space, which was performed by co-registration to the MNI template (MNI152_T1_1mm_brain) (Grabner et al., 2006) included in the FSL 4.1 software.

Images were then imported into the Matlab platform using the 'Tools For NIfTI And ANALYZE Image' toolbox (http://www.mathworks.com/matlabcentral/fileexchange/8797). No smoothing or segmentation were applied. Resulting images were limited within a bounding box. Final whole-brain volumes consisted of 145 × 178 × 133 voxels.

2.2.2. Feature extraction

Feature extraction was implemented by applying spatial transformations on the images in order to reduce data dimensions without losing relevant information.

We used PCA to extract the most significant features from our MRI datasets (Habeck et al., 2008; Alvarez et al., 2009; Salas-Gonzalez et al., 2010). PCA is a standard technique which consists in applying an orthogonal transformation to a dataset of (possibly) correlated variables to obtain a set of values of linearly uncorrelated (orthogonal) variables. These values are called 'principal components' of the original dataset.

The application of PCA involves a second step: the projection itself, which reduces the original number of features to a much lower number of so-called PCA coefficients. These coefficients are the ones used for classification.

Mathematically, let us define X to be a dataset of 3D brain images X_i, where i varies from 1 to N, N being the number of images in the dataset. Let's suppose that each image X_i is given in the form of a vector of dimension V (in our study V is the total number of voxels of each image, $V = 145 \times 178 \times 133$), so that the dimension of X is $N \times V$, and that the dataset X has zero mean (in case dataset X had non-zero mean, then the average X_M would be subtracted from each

image X_i). Now, PCA-space is defined as the space which is spanned by the eigenvectors of the covariance matrix C of the dataset X:

$$C = X \cdot X^{T}$$

Finally, PCA coefficients can be extracted by projecting each image onto the PCA-space (Alvarez et al., 2009).

Application of PCA to a given dataset results in a number of principal components which is at most equal to the number of the lower dimension of the data matrix −1. If N is the number of subjects in the dataset, there will only be $N-1$ eigenvectors (principal components) with non-zero eigenvalues. The other eigenvectors have a zero eigenvalue associated, so it does not make sense to consider them.

Once the original data are projected onto PCA coefficients or scores, these coefficients (low dimension representation of the samples) and associated labels can be considered to understand which principal components are more discriminative. For this purpose principal components were ordered in a decreasing order, according to their Fisher Discriminant Ratio (FDR):

$$\mathrm{FDR} = \frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2}$$

where μ_i and σ_i^2 denote the mean and the variance of the i-th class, respectively.

2.2.3. Classification algorithm

We used a classification algorithm based on Support Vector Machines (SVMs) (Scholkopf et al., 2002; Vapnik, 1995, 1998, 1999). Let's suppose that we have a set of training data each one consisting of a pair: an input vector $x_n \in \mathrm{R}^N$, $n = 1, \ldots N$ and the corresponding label or target value $t_n \in \{+/-1\}$. The aim of a SVM is to estimate a decision function which will correctly classify unseen examples (x, t). To do this, SVMs design the optimal separating hyper-plane in terms of distance between classes, that is, a hyper-plane which is maximally distant from the two classes to be separated (this is why it is also called maximal margin hyper-plane). The decision function for an unseen example x is then given by

$$y(x) = \sum_{n=1}^{N} a_n \cdot t_n \cdot k(x, x_n) + b$$

where N is the number of samples in the training set, a_n is a weight constant, $k(x, x_n)$ is a kernel function and b is a threshold parameter. This decision function returns the predicted class $y(x)$ of the unseen example x. In this way, this hyper-plane is able to discriminate binary labeled training datasets and to subsequently classify unseen data as belonging to one or the other of the two training classes.

In our work, the SVM classifier was implemented using algorithms of the biolearning toolbox of Matlab. Datasets were divided into three binary labeled groups: PD versus Controls, PSP versus Controls and PSP versus PD. The training step was performed using the principal components and the associated labels of each dataset as features for the SVM (López et al., 2011). A linear kernel was chosen as the more general form and for the purpose of improving computational efficiency. The training of the classifier was carried out for a number i of principal components ranging from 1 to PC, where PC is the whole number of extracted principal components.

2.3. Validation of the classifier

Validation of the classification algorithm was performed by a cross-validation strategy. In cross-validation, the data is partitioned into complementary subsets: the training set and the testing set. The training of the classifier is performed on the training set; the validation of the classifier is performed on the testing set. Multiple rounds of cross-validation can then be performed using different partitions. In this study we used Leave-One-Out (LOO) validation, that is a particular case of cross-validation in which the testing set is made up of only one sample of the dataset and the training set is made up of the remaining samples of the dataset $(N-1)$. In this way, all samples in the dataset can be tested in turn if the total number of rounds equals the number of samples in the dataset. LOO is a widely used validation approach because it returns an almost unbiased estimate of the probability of test error of a classification algorithm (e.g. Vapnik, 1998; Chapelle et al., 1999).

Moreover, the use of the same number of subjects for training and classification is somewhat controversial. Given the fair amount of subjects available in our study, we also adopted another validation method using $N/2$ randomly chosen subjects for training the classifier and the remaining N/2 subjects for testing.

Validation was carried out by the two methods for each of the three binary labeled groups and for a number i of principal components ranging from 1 to PC, where PC is the whole number of extracted principal components ($N-1$ or $N/2$). Accuracy, Specificity and Sensitivity rates were computed for each binary labeled group over the first i principal components as follows:

$$\mathrm{Accuracy}_i = \frac{T_{\mathrm{RC}}}{T}$$

$$\mathrm{Specificity}_i = \frac{X_{\mathrm{RC}}}{X_{\mathrm{RC}} + Y_{\mathrm{WC}}}$$

$$\mathrm{Sensitivity}_i = \frac{Y_{\mathrm{RC}}}{Y_{\mathrm{RC}} + X_{\mathrm{WC}}}$$

T being the total number of classified images; T_{RC} being the total number of classified images that underwent Right Classification (RC); X_{RC} being the number of images belonging to the first binary labeled group that underwent Right Classification (RC); X_{WC} being the number of images belonging to the first binary labeled group that underwent Wrong Classification (WC); Y_{RC} being the number of classified images belonging to the second binary labeled group that underwent RC; Y_{WC} being the number of images belonging to the second binary labeled group that underwent WC.

Overall Mean Accuracy, Specificity and Sensitivity rates were obtained as mean values calculated over a number of principal components ranging from 1 to PC, where PC is the whole number of extracted principal components.

The dependence of Accuracy, Specificity and Sensitivity on the number of principal components was studied.

$\mathrm{Accuracy}_{>80}$, $\mathrm{Specificity}_{>80}$ and $\mathrm{Sensitivity}_{>80}$ rates (mean, minimum and maximum values) were calculated over a range of principal components for which Accuracy, Specificity and Sensitivity fell above 80% for each of the three binary labeled groups.

2.4. Voxel-based pattern distribution

Using the whole number of subjects in the dataset (28 PD, 28 PSP and 28 Controls) and for each binary labeled group, we generated image maps of voxel-based pattern distribution of brain structural differences, which reflect the importance of each image voxel for the SVM classification. This allowed the identification of MRI-related biomarkers useful for the diagnosis of PD and PSP.

During the training step, the SVM classifier calculates a specific weight for each of the images, reflecting the importance of that image for binary separation. This weight is non-zero only for support vectors, while its sign is positive for support vectors belonging to the first binary labeled group and negative for support vectors belonging to the second binary labeled group.

Table 1
Demographic and clinical data of enrolled subjects.

Variables	HC	PD	PSP	*p* values
N°	28	28	28	
Gender (% males)	54%	54%	64%	
Age (years)	67.5 ± 7.1	68.2 ± 5	69.4 ± 5.7	n.s.
Disease duration (years)	–	8.0 ± 4.8	3.0 ± 1.6	<0.001[a]
Age at onset (years)	–	60.8 ± 5.6	67.2 ± 3.0	<0.001[a]
UPDRS-ME	–	24 (10–45)	34 (24–47)	<0.001[b]
H&Y	–	3 (1–4)	4 (3–5)	<0.001[b]
MMSE	26.5 ± 2.1	25.7 ± 1.9	24 ± 4.8	<0.001[c]

Note: Data are given as mean values (SD) or median values (range) when appropriate.
UPDRS-ME, Unified Parkinson Disease Rating Scale-Motor Examination in "off" phase (off medications overnight); H&Y, Hoehn-Yahr; MMSE, Mini Mental State Examination.
[a] Unpaired *t* test.
[b] Mann–Whitney test.
[c] One-way ANOVA.

Each image of the training set was multiplied with the corresponding weight and summed on a voxel basis, resulting in a map of values reflecting the importance of each voxel for SVM binary group discrimination (Kloppel et al., 2008; Focke et al., 2011). The resulting map was superimposed onto a standard stereotactic brain for visualization and localization purposes.

3. Results

3.1. Clinical and MRI studies

Data on demographic and clinical features of all patients and control subjects included in the study are listed in Table 1. Although no significant differences were detected among groups in demographical data, as expected, PSP patients were characterized by a more rapid disease progression (with a fatal prognosis after few years) and by a more critical clinical status with respect to PD patients in terms of motor disability, as assessed by higher UPDRS and H&Y scores.

Considering MRI studies, all subjects had no evidence of vascular lesions as evaluated in Fluid Attenuated Inversion Recovery (FLAIR) and by T2-weighted MRI. Healthy control subjects had normal MRI scanning. Both PD and PSP patients showed no evident structural abnormalities.

3.2. The machine-learning method

3.2.1. Features extraction

Fig. 1 shows the 1-st, 2-nd and 3-rd PCA coefficients extracted, as representative examples, for the PSP (28) versus PD (28) binary labeled group (for this comparison, the total number of extracted PCA coefficients was equal to 55).

3.2.2. Classification algorithm

Fig. 2 shows the optimal separating hyper-plane resulting from the SVM classification algorithm trained for the PSP (28) versus PD (28) binary labeled group.

3.3. Validation of the classifier

Table 2 shows Accuracy, Specificity and Sensitivity of the SVM classifier for PD versus Controls, PSP versus Controls and PSP versus PD binary labeled groups obtained using LOO validation. Overall Mean Accuracy, Specificity and Sensitivity rates were calculated over a number of principal components ranging from 1 to 55. Overall Mean Accuracy (Specificity/Sensitivity) were 85.8 (86.0/86.0), 89.1 (89.1/89.5) and 88.9 (88.5/89.5)% for PD versus Controls, PSP versus Controls and PSP versus PD binary labeled groups, respectively.

Fig. 3 shows how the metrics depend on the number of principal components in LOO validation, as a representative example. In this case, results are shown for a number of principal components ranging from 1 to 55 and for the PSP versus PD binary labeled group. As expected, Accuracy, Specificity and Sensitivity rates increase with the number of considered principal components. This occurs because, in our method, the dimension of the feature space is enhanced with a covariance eigenvalue criterion coupled with an FDR criterion, allowing noisy information to be contained only in a small fraction of the eigenvectors (the last eigenvectors). For instance, without an FDR process, the relevant information would be contained in the few first eigenvectors and the rest of the eigenvectors would only contribute with noisy information (Alvarez et al., 2009).

The range of principal components for which Accuracy, Specificity and Sensitivity fell above 80% ($Accuracy_{>80}$, $Specificity_{>80}$ and $Sensitivity_{>80}$) was found from 30 to 52 components in LOO validation. In this range, for each of the binary labeled groups, $Accuracy_{>80}$ was found > 83.9% (mean $Accuracy_{>80}$: 92.7, 97.0 and 98.2% for PD versus Controls, PSP versus Controls and PSP

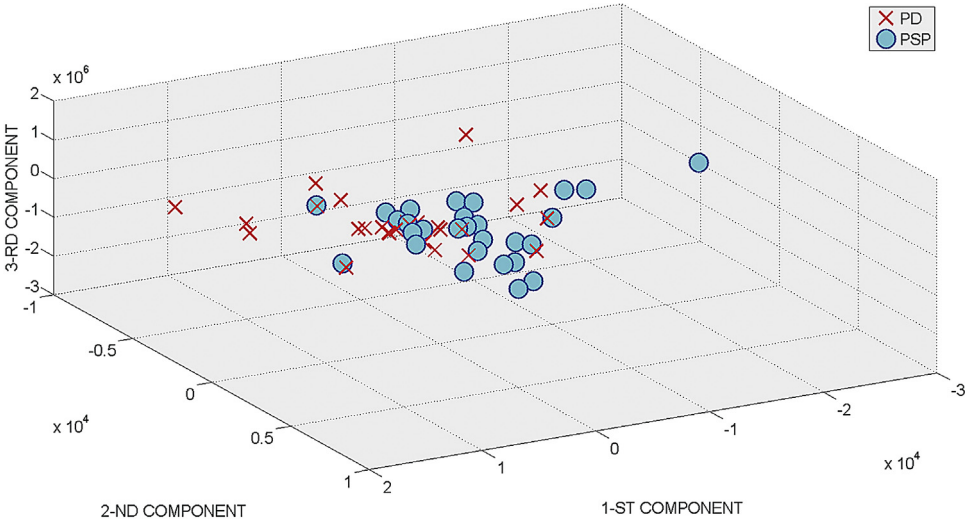


Fig. 1. PCA coefficients for the PSP versus PD binary labeled group (1st, 2nd and 3rd components).

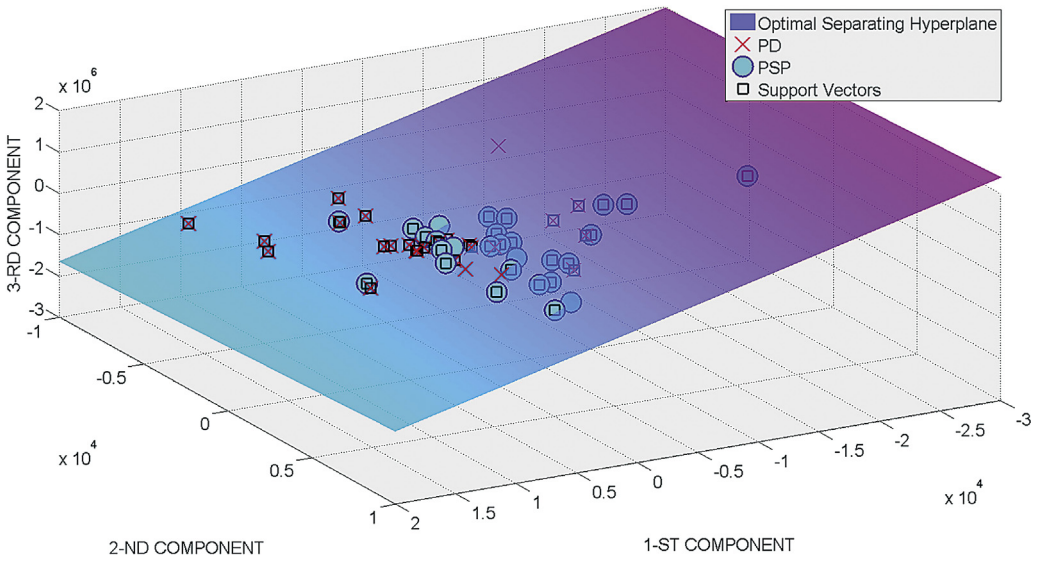


Fig. 2. Optimal separating hyper-plane for the PSP versus PD binary labeled group (1st, 2nd and 3rd components).

Table 2
Accuracy, Specificity and Sensitivity rates of SVM using LOO validation.

	Overall Mean Accuracy (%) $Accuracy_{>80}$ (%) Mean (Min/Max)	Overall Mean Specificity (%) $Specificity_{>80}$ (%) Mean (Min/Max)	Overall Mean Sensitivity (%) $Sensitivity_{>80}$ (%) Mean (Min/Max)
PD vs. Controls	85.8 92.7 (83.9/100.0)	86.0 92.3 (81.3/100.0)	86.0 93.4 (80.6/100.0)
PSP vs. Controls	89.1 97.0 (92.9/100.0)	89.1 98.2 (92.9/100.0)	89.5 95.9 (90.0/100.0)
PSP vs. PD	88.9 98.2 (94.6/100.0)	88.5 98.8 (96.3/100.0)	89.5 97.8 (93.1/100.0)

versus PD binary labeled groups, respectively). $Specificity_{>80}$ was found > 81.3% (mean $Specificity_{>80}$: 92.3, 98.2 and 98.8% for PD versus Controls, PSP versus Controls and PSP versus PD binary labeled groups, respectively). $Sensitivity_{>80}$ was found > 80.6% (mean $Sensitivity_{>80}$: 93.4, 95.9 and 97.8% for PD versus Controls, PSP versus Controls and PSP versus PD binary labeled groups, respectively). Furthermore, there is always at least one number of employed principal components for which $Accuracy_{>80}$, $Specificity_{>80}$ and $Sensitivity_{>80}$ values reach 100%.

Table 3 shows Accuracy, Specificity and Sensitivity of the SVM classifier for PD versus Controls, PSP versus Controls and PSP versus PD binary labeled groups obtained using N/2 randomly chosen subjects for training the classifier and the remaining N/2 subjects for testing. Overall Mean Accuracy, Specificity and Sensitivity

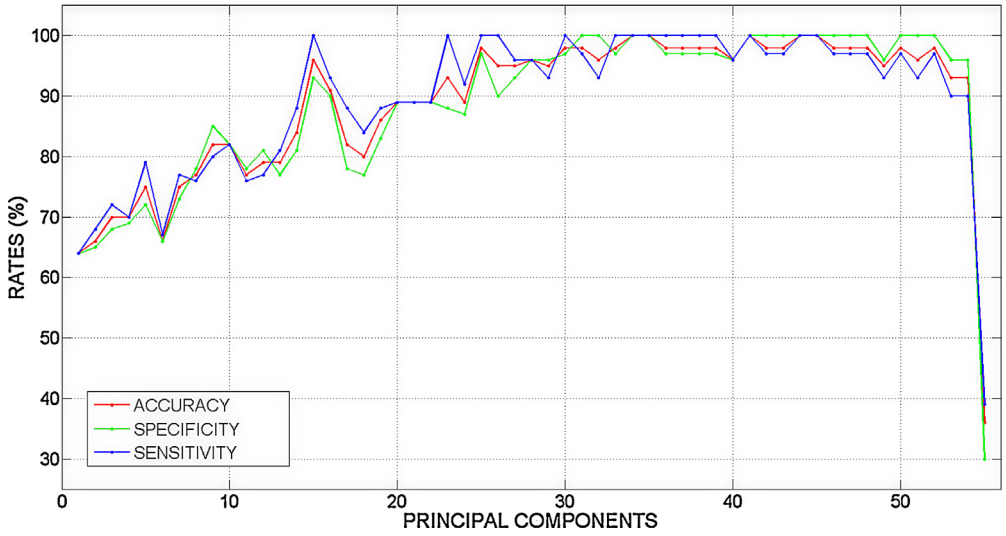


Fig. 3. Accuracy, Specificity and Sensitivity rates (%) of SVM versus Number of PCA components in LOO validation.

Table 3
Accuracy, Specificity and Sensitivity rates of SVM using N/2 subjects for training and N/2 subjects for testing.

	Overall Mean Accuracy (%) $Accuracy_{>80}$ (%) Mean (Min/Max)	Overall Mean Specificity (%) $Specificity_{>80}$ (%) Mean (Min/Max)	Overall Mean Sensitivity (%) $Sensitivity_{>80}$ (%) Mean (Min/Max)
PD vs. Controls	83.2 93.5 (89.3/100)	81.9 90.6 (82.4/100)	85.4 97.4 (92.3/100)
PSP vs. Controls	86.2 92.2 (85.7/96.4)	92.1 92.5 (85.7/100)	82.9 92.4 (85.7/100)
PSP vs. PD	84.7 92.2 (89.3/96.4)	87.5 91.3 (82.4/100)	83.8 94.4 (86.7/100)

rates were calculated over a number of principal components ranging from 1 to 28. Overall Mean Accuracy (Specificity/Sensitivity) were 83.2 (81.9/85.4), 86.2 (92.1/82.9) and 84.7 (87.5/83.8)% for PD versus Controls, PSP versus Controls and PSP versus PD binary labeled groups, respectively.

The range of principal components for which Accuracy, Specificity and Sensitivity fell above 80% ($Accuracy_{>80}$, $Specificity_{>80}$ and $Sensitivity_{>80}$) was found from 16 to 26 components when using N/2 randomly chosen subjects for training the classifier and the remaining N/2 subjects for testing. In this range, for each of the binary labeled groups, $Accuracy_{>80}$ was found > 85.7% (mean $Accuracy_{>80}$: 93.5, 92.2 and 92.2% for PD versus Controls, PSP versus Controls and PSP versus PD binary labeled groups, respectively). $Specificity_{>80}$ was found > 82.4% (mean $Specificity_{>80}$: 90.6, 92.5 and 91.3% for PD versus Controls, PSP versus Controls and PSP versus PD binary labeled groups, respectively). $Sensitivity_{>80}$ was found > 85.7% (mean $Sensitivity_{>80}$: 97.4, 92.4 and 94.4% for PD versus Controls, PSP versus Controls and PSP versus PD binary labeled groups, respectively). All the considered parameters proved the good performance of the classification algorithm with respect to all three binary labeled groups.

3.4. Voxel-based pattern distribution

Fig. 4 shows, maps of voxel-based pattern distribution displaying the influence of each voxel for the classification in the PD (28) versus Controls (28), in the PSP (28) versus Controls (28) and in the PSP (28) versus PD (28) comparisons. The pattern of differences is expressed according to the color scales. The most relevant finding concerned the separation of patients with PD with respect to patients with PSP. As showed in the bottom part of the Fig. 4, voxels influencing the classification between PD and PSP patients are localized in the midbrain, pons, corpus callosum and thalamus. When considering instead the direct comparisons with healthy controls, the voxel-based pattern distribution was similar for both PD and PSP patients, although it is important to highlight that SVM classification revealed significant voxels within the medial part of the midbrain (encompassing the substantia nigra) and the caudal part of the pons only for PD as compared to controls (upper part of the Fig. 4).

4. Discussion

Effective and accurate diagnosis of PD or PSP, a critical clinical variant of PD, by means of MRI biomarkers has recently attracted strong attention. So far, several biomarkers have been shown to be sensitive to the diagnosis of PD as compared to PSP. For instance, morphological abnormalities in the brainstem as well as in the cerebellar peduncles, have been demonstrated to be useful markers in a clinical context (Massey et al., 2013; Quattrone et al., 2008; Schulz et al., 1999). However, the only validated MRI-based measurement employed in clinical practice of PD derives from conventional MRI using manual morphometric quantification. In the above - mentioned studies it has been shown by using different approaches,

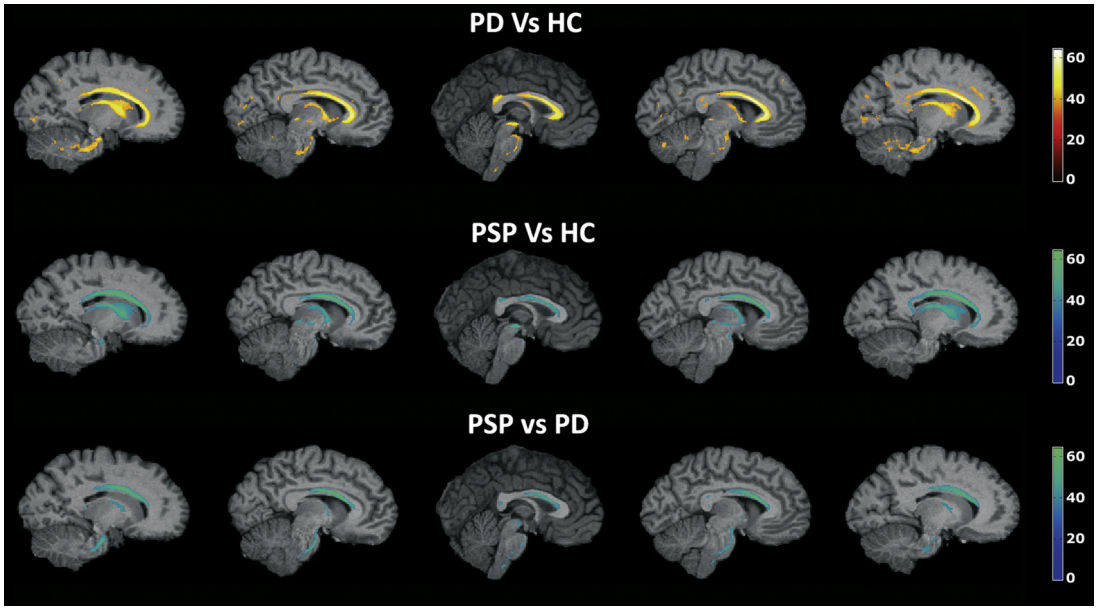


Fig. 4. Maps of voxel-based pattern distribution of brain structural differences (sagittal view, threshold = 60%). The importance of each voxel in the SVM classification is expressed according to the color scale. A, PD versus Controls; B, PSP versus controls; C, PSP versus PD; PD, Parkinson's disease; PSP, Progressive Supranuclear Palsy.

that the diameter or area of the brainstem subregions represent the simplest morphological feature on the basis of which a differential diagnosis of PD with respect to PSP can be performed. Although manual segmentation is currently considered as the gold standard approach to determine the morphology of brain regions, this technique is time-consuming, dependent on the raters' experience and limited to few specific regions-of-interest. For this reason, the implementation of supervised whole-brain automatic classification methods (Haller et al., 2011) is an essential step for improving clinical management of neurological patients, as well as, in longitudinal and prospective studies.

SVM has been proposed as a revolutionary approach for identifying sensitive biomarkers (or a combination between them) that allow for automatic discrimination of individual subjects. In this study, we considered a SVM algorithm using structural neuroimaging data at a whole brain level, reaching an excellent individual classification, both in the comparisons between PSP patients with PD patients (mean accuracy > 92.5) and, in the comparison between PD and healthy controls (mean accuracy > 92.7%). To the best of our knowledge, the classification accuracy in the discrimination of individual PSP patients was consistent with previous manual morphological metrics (Massey et al., 2013; Quattrone et al., 2008; Oba et al., 2005; Schulz et al., 1999) and with other SVM application to MRI data (Focke et al., 2011; Haller et al., 2012), whereas the accuracy in the detection of individual PD patients was significantly higher. For instance, in one first study (Focke et al., 2011), the best classification accuracy achieved ~ 97% (considering only white matter tissue) and was obtained for the comparison between PSP vs. PD using T1-weighted MRI data, while separation from the PD and the control cohort was only marginally better than chance. Successively, Haller and colleagues (2012, 2013) applying SVM to other MRI parameters (DTI and susceptibility-weighted images) reported classification accuracies in the individual diagnosis of PD comparable to our findings. However, it is important to bear in mind that in these later works, the clinical classification was made without considering healthy controls and including a heterogeneous cohort of parkinsonisms in the same group of patients, i.e. joining PSP cases with patients affected by multiple system atrophy (MSA) without considering that these disorders are characterized by distinct patterns of brain atrophy (Messina et al., 2011).

Our study has several strengths that may have improved our discrimination analysis. First, the larger sample of PSP patients investigated (n° 28) with respect to other studies (n° 10 in Focke et al., 2011; n° 1 in Haller et al., 2012 and n° 1 in Haller et al., 2013). Second, the clinical consistency of our group selection where only PSP patients were enrolled without conflating our analysis with the inclusion of other parkinsonian variants, such as MSA, dementia with Lewy Body, vascular Parkinsonism and atypical tremor (Focke et al., 2011; Haller et al., 2012, 2013). Third, the balance between classes (i.e., we choose the same number of samples in each class), allowing the system to learn without biases due to unequal samples.

Another point of relevance of our work is that we also studied the relevance of each brain voxel with respect to the classification analysis, thus allowing to identify regions critically involved in the pathophysiological mechanisms of PD and PSP. Indeed, brain regions that allowed to perform the best discrimination between PD and PSP were: midbrain, pons, corpus callosum and thalamus. These features are highly consistent with typical neuropathological (Steele et al., 1964) and imaging findings described in patients with PSP (Shi et al., 2013; Messina et al., 2011), where a key role is played by the volumetric atrophy of the brainstem, which represents a hallmark of PSP (Oba et al., 2005). More recently, studies using diffusion tensor imaging (DTI) for quantifying white matter pathology in PSP, highlighted the involvement of the corpus callosum as well. This finding is of particular relevance since corpus callosum is the largest white matter tract in the brain, enabling interhemispheric communication, particularly with respect to motor coordination, and is one of the tract that is known to be damaged in PSP (Knake et al., 2010; Canu et al., 2011).

As far as the spatial pattern of the SVM classification between PD and controls is concerned, more widespread patterns involving several cerebral regions were found in our analysis, thus confirming that PD is a more heterogeneous clinical phenotype that might be characterized by several and topographically separated neural pathologies. Of note, SVM revealed significant voxels within the medial part of the midbrain (encompassing the substantia nigra) and the caudal part of the pons. This finding is consistent with the Braak's neuroanatomical model of the PD (Braak et al., 2003). Post-mortem studies by Braak and colleagues (Del Tredici et al., 2002), based on the analysis of Lewy neuritis and Lewy bodies accumulation, a proteic hallmark of PD, have shown that various cerebral structures are damaged before substantia nigra in a consistent and repeated pattern. In a six-stage model (Braak et al., 2004), PD would initially begin in the medulla oblongata (stage 1) and in the olfactory bulb, and progresses in a caudo-rostral pattern (stage 2), affecting substantia nigra in stage 3 only, corresponding to the onset of the motor symptoms, often revealed by the first visit of the patient to a neurologist. However, previous structural neuroimaging studies investigating the neural basis of PD did not describe the occurrence of anatomical changes in this region (a part from a single study, Jubault et al., 2009). It must be acknowledged, that most of the studies that have investigated structural abnormalities in PD, as those previously reported, are based on standard mass-univariate analytical methods, applied in structural neuroimaging (i.e. voxel based morphometry). The SVM technique has one main advantage over these techniques: it takes inter-regional correlations into account and therefore is sensitive to differences that are subtle and spatially distributed; as such, it provides an ideal framework for investigating neurological disorders that affect a distributed network of regions.

In conclusion, our findings, together with those provided by other colleagues (Focke et al., 2011; Haller et al., 2012, 2013) offer new avenues for encouraging the application of computer-based diagnosis in clinical practice of PD.

References

Alvarez I, Gorriz JM, Ramirez J, Salas-Gonzalez D, Lopez M, Puntonet CG, et al. Alzheimer's diagnosis using eigenbrains and support vector machines. Electron Lett 2009;45:342–3.

Braak H, Del Tredici K, Rub U, De Vos RA, Jansen Steur EN, Braak E. Staging of brain pathology related to sporadic Parkinson's disease. Neurobiol Aging 2003;24:197–211.

Braak H, Ghebremedhin E, Rub U, Bratzke H, Del Tredici K. Stages in the development of Parkinson's disease-related pathology. Cell Tissue Res 2004;318:121–34.

Canu E, Agosta F, Baglio F, Galantucci S, Nemni R, Filippi M. Diffusion tensor magnetic resonance imaging tractography in progressive supranuclear palsy. Mov Disord 2011;26:1752–5.

Chapelle O, Vapnik V. Model selection for support vector machines. Adv Neural Inform Process Syst 1999;12:230–6.

Del Tredici K, Rub U, De Vos RA, Bohl JR, Braak H. Where does parkinson disease pathology begin in the brain? J Neuropathol Exp Neurol 2002;61:413–26.

Focke NK, Helms G, Scheewe S, Pantel PM, Bachmann CG, Dechent P, et al. Individual voxel-based subtype prediction can differentiate progressive supranuclear palsy from idiopathic parkinson syndrome and healthy controls. Hum Brain Mapp 2011;32:1905–15.

Gelb DJ, Oliver E, Gilman S. Diagnostic criteria for Parkinson disease. Arch Neurol 1999;56:33–9.

Grabner G, Janke AL, Budge MM, Smith D, Pruessner J, Collins DL. Symmetric atlasing and model based segmentation: an application to the hippocampus in older adults. Med Image Comput Comput Assist Interv 2006;9(Pt 2):58–66.

Habeck C, Foster NL, Perneczky R, Kurz A, Alexopoulos P, Koeppe RA, et al. Multivariate and univariate neuroimaging biomarkers of Alzheimer's disease. Neuroimage 2008;40:1503–15.

Haller S, Badoud S, Nguyen D, Barnaure I, Montandon ML, Lovblad KO, et al. Differentiation between Parkinson disease and other forms of Parkinsonism using support vector machine analysis of susceptibility-weighted imaging (SWI): initial results. Eur Radiol 2013;23(1):12–9.

Haller S, Badoud S, Nguyen D, Garibotto V, Lovblad KO, Burkhard PR. Individual detection of patients with Parkinson disease using support vector machine analysis of diffusion tensor imaging data: initial results. AJNR Am J Neuroradiol 2012;33(11):2123–8.
Haller S, Lovblad KO, Giannakopoulos P. Principles of classification analyses in mild cognitive impairment (MCI) and Alzheimer disease. J Alzheimers Dis 2011;26(Suppl 3):389–94.
Jankovic J. Parkinson's disease: clinical features and diagnosis. J Neurol Neurosurg Psychiatry 2008;79:368–76.
Jenkinson M, Beckmann CF, Behrens TE, Woolrich MW, Smith SM. FSL. Neuroimage 2012;62:782–90.
Jubault T, Brambati SM, Degroot C, Kullmann B, Strafella AP, Lafontaine AL, et al. Regional brain stem atrophy in idiopathic Parkinson's disease detected by anatomical MRI. PLoS One 2009;4:e8247.
Kloppel S, Stonnington CM, Chu C, Draganski B, Scahill RI, Rohrer JD, et al. Automatic classification of MR scans in Alzheimer's disease. Brain 2008;131:681–9.
Knake S, Belke M, Menzler K, et al. In vivo demonstration of microstructural brain pathology in progressive supranuclear palsy: a DTI study using TBSS. Mov Disord 2010;25:1232–8.
Litvan I, Agid Y, Calne D, et al. Clinical research criteria for the diagnosis of progressive supranuclear palsy (Steele-Richardson-Olszewski syndrome): report of the NINDS-SPSP international workshop. Neurology 1996;47:1–9.
López M, Ramírez J, Górriz JM, Álvarez I, Salas-Gonzalez D, Segovia F, et al. Principal component analysis-based techniques and supervised classification schemes for the early detection of Alzheimer's disease. Neurocomputing 2011;74:1260–71.
Magnin B, Mesrob L, Kinkingnéhun S, Pélégrini-Issac M, Colliot O, Sarazin M, et al. Support vector machine-based classification of Alzheimers disease from whole-brain anatomical MRI. Neuroradiology 2009;51:73–83.
Massey LA, Jäger HR, Paviour DC, O'Sullivan SS, Ling H, Williams DR, et al. The midbrain to pons ratio: a simple and specific MRI sign of progressive supranuclear palsy. Neurology 2013;80:1856–61.
Messina D, Cerasa A, Condino F, Arabia G, Novellino F, Nicoletti G, et al. Patterns of brain atrophy in Parkinson's disease, progressive supranuclear palsy and multiple system atrophy. Parkinsonism Relat Disord 2011;17(3):172–6.
Oba H, Yagishita A, Terada H, Barkovich AJ, Kutomi K, Yamauchi T, et al. New and reliable MRI diagnosis for progressive supranuclear palsy. Neurology 2005;64:2050–5.
Quattrone A, Nicoletti G, Messina D, Fera F, Condino F, Pugliese P, et al. MR imaging index for differentiation of progressive supranuclear palsy from Parkinson disease and the Parkinson variant of multiple system atrophy1. Radiology 2008;246:214–21.
Salas-Gonzalez D, Górriz JM, Ramírez J, Illán IA, López M, Segovia F, et al. Feature selection using factor analysis for Alzheimer's diagnosis using 18F-FDG PET images. Med Phys 2010;37:6084–95.
Scholkopf B, Smola AJ. Learning with kernels: support vector machines, regularization, optimization, and beyond. 1st ed. Cambridge: MIT Press; 2001.
Schulz JB, Skalej M, Wedekind D, Luft AR, Abele M, Voigt K, et al. Magnetic resonance imaging based volumetry differentiates idiopathic Parkinson's syndrome from multiple system atrophy and progressive supranuclear palsy. Ann Neurol 1999;45:65–74.
Shi HC, Zhong JG, Pan PL, Xiao PR, Shen Y, Wu LJ, et al. Gray matter atrophy in progressive supranuclear palsy: meta-analysis of voxel-based morphometry studies. Neurol Sci 2013;34(7):1049–55, http://dx.doi.org/10.1007/s10072-013-1406-9.
Smith SM, Jenkinson M, Woolrich MW, Beckmann CF, Behrens TEJ, Johansen-Berg H, et al. Advances in functional and structural MR image analysis and implementation as FSL. Neuroimage 2004;23(S1):208–19.
Steele JC, Richardson JC, Olszewski J. Progressive supranuclear palsy. A heterogeneous degeneration involving the brain stem, basal ganglia and cerebellum with vertical gaze and pseudobulbar palsy, nuchal dystonia and dementia. Arch Neurol 1964;10:333–59.
Teipel SJ, Born C, Ewers M, Bokde AL, Reiser MF, Moller HJ, et al. Multivariate deformation-based analysis of brain atrophy to predict Alzheimer's disease in mild cognitive impairment. Neuroimage 2007;38:13–24.
Tolosa E, Wenning G, Poewe W. The diagnosis of Parkinson's disease. Lancet Neurol 2006;5:75–86.
Vapnik V. An overview of statistical learning theory. IEEE Trans Neural Netw 1999;10:988–1000.
Vapnik V. Statistical learning theory. New York: John Wiley and Sons, Inc; 1998.
Vapnik V. The nature of statistical learning theory. 2nd ed. New York: Springer-Verlag; 1995.

PUBLICATIONS II – III

Application of the implemented ML method to the early diagnosis of AD (Application II)

frontiers in Neuroscience

ORIGINAL RESEARCH
published: 01 September 2015
doi: 10.3389/fnins.2015.00307

Magnetic resonance imaging biomarkers for the early diagnosis of Alzheimer's disease: a machine learning approach

Christian Salvatore[1], *Antonio Cerasa*[2], *Petronilla Battista*[1], *Maria C. Gilardi*[1], *Aldo Quattrone*[3], *Isabella Castiglioni*[1]* *and the Alzheimer's Disease Neuroimaging Initiative*[†]

[1] Institute of Molecular Bioimaging and Physiology, National Research Council (IBFM-CNR), Milan, Italy, [2] Neuroimaging Research Unit, Institute of Molecular Bioimaging and Physiology, National Research Council (IBFM-CNR), Catanzaro, Italy, [3] Department of Medical Sciences, Institute of Neurology, University "Magna Graecia", Catanzaro, Italy

Determination of sensitive and specific markers of very early AD progression is intended to aid researchers and clinicians to develop new treatments and monitor their effectiveness, as well as to lessen the time and cost of clinical trials. Magnetic Resonance (MR)-related biomarkers have been recently identified by the use of machine learning methods for the *in vivo* differential diagnosis of AD. However, the vast majority of neuroimaging papers investigating this topic are focused on the difference between AD and patients with mild cognitive impairment (MCI), not considering the impact of MCI patients who will (MCIc) or not convert (MCInc) to AD. Morphological T1-weighted MRIs of 137 AD, 76 MCIc, 134 MCInc, and 162 healthy controls (CN) selected from the Alzheimer's disease neuroimaging initiative (ADNI) cohort, were used by an optimized machine learning algorithm. Voxels influencing the classification between these AD-related pre-clinical phases involved hippocampus, entorhinal cortex, basal ganglia, gyrus rectus, precuneus, and cerebellum, all critical regions known to be strongly involved in the pathophysiological mechanisms of AD. Classification accuracy was 76% AD vs. CN, 72% MCIc vs. CN, 66% MCIc vs. MCInc (nested 20-fold cross validation). Our data encourage the application of computer-based diagnosis in clinical practice of AD opening new prospective in the early management of AD patients.

Keywords: Alzheimer's disease, mild cognitive impairment, magnetic resonance imaging, support vector machine, structural neuroimaging biomarkers, machine learning, automatic classification, artificial intelligence

OPEN ACCESS

Edited by:
Stephen C. Strother, University of Toronto, Canada

Reviewed by:
Delia Cabrera DeBuc, University of Miami, USA
Li-Wei Kuo, National Health Research Institutes, Taiwan

***Correspondence:**
Isabella Castiglioni, Institute of Molecular Bioimaging and Physiology, National research Council (IBFM-CNR), Via F.lli Cervi, 93, 20090 Segrate, Milan, Italy
isabella.castiglioni@ibfm.cnr.it

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

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

Received: *20 March 2015*
Accepted: *13 August 2015*
Published: *01 September 2015*

Citation:
Salvatore C, Cerasa A, Battista P, Gilardi MC, Quattrone A and Castiglioni I (2015) Magnetic resonance imaging biomarkers for the early diagnosis of Alzheimer's disease: a machine learning approach. Front. Neurosci. 9:307. doi: 10.3389/fnins.2015.00307

Introduction

The increase in life expectancy and the prevalence of age-related cognitive disorders have led to great interest in studying normal and pathological aging with the aim to individuate early predictors of degenerative disorders, differential diagnosis, and efficacies of pharmacological and cognitive approaches in the treatment of these disorders. Indeed, considering the great burden of degenerative diseases on national healthcare systems in terms of cost and therapies, research aimed at improving the early and differential diagnosis of these pathologies is mandatory.

Alzheimer's Disease (AD) is the first most common neurodegenerative disease affecting millions of people worldwide (Martin et al., 2012). Determination of sensitive and specific markers of very early AD progression is intended to aid researchers and clinicians to develop new treatments and monitor their effectiveness, as well as to lessen the time and cost of clinical trials. To date, individual diagnosis of AD is predominantly based on the clinical examination and neuropsychological assessment (Knopman et al., 2001; Blennow et al., 2006), but definite diagnosis can only be performed by post-mortem analysis.

In the 1980s, the National Institute of Neurologic and Communicative Disorders and Stroke and the Alzheimer's Disease and Related Disorders Association (NINCDS-ADRDA) developed clinical diagnostic criteria for AD by applying a binary approach to diagnosis. According to this approach, a cognitive impairment is necessary for the diagnosis of AD, with definite, probable and possible categories (McKhann et al., 1984). Neuropathological data based on senile plaques and neurofibrillary tangles were afterwards introduced (Hyman and Trojanowski, 1997).

In 2011, revised diagnostic criteria for AD have been developed by the National Institute on Aging-Alzheimer's Association workgroup. These revised diagnostic criteria have replaced the binary approach for a more biological definition of AD: additional supportive features can be obtained by neurogenetic testing, measurement of cerebrospinal fluid (CSF), amyloid and tau, and by neuronal injury biomarkers as measured by neuroimaging studies, including Positron Emission Tomography (PET) and Magnetic Resonance Imaging (MRI). PET and MR changes provide measurements of metabolism/amyloid markers (Fox and Schott, 2004; Jagust et al., 2006) and of atrophic regions, respectively, in order to identify AD, even before dementia is apparent (Albert et al., 2011; Sperling et al., 2011).

Due to the non- invasiveness of MR modality, a considerable effort has been put into the development of advanced MR image processing techniques in order to identify MR-related biomarkers which could be used for enhancing the accuracy of clinical diagnosis of AD. Most studies which were focused on the identification of MR image differences between patients with a clinical diagnosis of AD and healthy subjects were based on a priori-defined regions of interest or on mass univariate image analysis methods (e.g., Voxel Based Morphometry, Busatto et al., 2003; Karas et al., 2003; Ishii et al., 2005). However, both approaches are not able to detect spatially distributed pattern of brain anatomy.

In order to overcome these limitations, in the last few years, there has been a growing interest within the neuroimaging community toward alternative approaches to the analyses of neuroimaging data by considering multivariate pattern analysis, including machine-learning algorithms. Due to their multivariate properties, machine-learning techniques are able to automatically extract multiple information from image sets without requiring a priori hypotheses of where this information may be coded in the images. Several studies have assessed the diagnostic value of these techniques in the classification of AD by cerebral MRI studies (Davatzikos et al., 2008; Klöppel et al., 2008; Gerardin et al., 2009; Cuingnet et al., 2011; Hidalgo-Muñoz et al., 2014), showing promising results also for the prediction of conversion in the early stages of disease (Tufail et al., 2012; Moradi et al., 2015). Among these studies, Klöppel et al. (2008) used machine learning classification and structural MR images for the extraction of spatially-distributed multivariate diagnostic biomarkers. Specifically, the authors were able to identify MR-related biomarkers useful for the differential diagnosis of AD with respect to Fronto-Temporal Lobar Degeneration and normality.

However, early diagnosis of AD by structural MR imaging studies is currently an open challenge due to the difficulty of quantifying patterns of structural change during early stages of AD or during clinically normal stages (Davatzikos et al., 2008). Patients suffering from AD at a prodromal stage are often clinically classified as Mild Cognitive Impairment (MCI), but not all MCI patients convert into AD. A meta-analysis of research and clinical reports suggests that the rate of conversion of MCI to AD is around 5–10% per year (Mitchell and Shiri-Feshki, 2009). Criteria for MCI have been developed (Albert et al., 2011) and various forms have been described (Petersen et al., 1999). Detecting the transition from the asymptomatic phase to symptomatic pre-dementia phase or from the symptomatic pre-dementia phase to dementia onset in the clinical setting is a non-trivial issue (Albert et al., 2011). This causes a diagnostic uncertainty for the early stage of disease.

For this objective, it seems crucial to identify multivariate MR-related diagnostic biomarkers that are able to accurately diagnose MCI converter (MCIc) and MCI non converter (MCInc) with respect to AD and normality. Therefore, different morphological characteristics between normal aging and MCI may be identified by a specific and sensitive analysis of MR images, by revealing which are the most informative image features supporting an early diagnosis (Davatzikos et al., 2008).

In this work we propose a machine learning method able to extract spatially distributed multivariate diagnostic biomarkers from structural MR brain images to be used for the early and accurate diagnosis of AD. In particular, our method is able to identify MRI-related biomarkers of MCI subjects which will convert into AD, opening new prospective in the early management of AD patients.

Materials and Methods

Participants

Subjects included in this study were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). We enrolled 162 cognitively normal elderly controls (CN), 137 patients with diagnosis of AD, 76 patients with diagnosis of MCI who converted to AD within 18 months (MCIc) and 134 patients with diagnosis of MCI who did not convert to AD within 18 months (MCInc). MCI patients who had been followed less than 18 months were not considered. Demographic and clinical data (sex, age and mini-mental score) for each group are shown in **Table 1** (see http://adni.loni.usc.edu/study-design/background-rationale/ for further description of groups). A total of 509 subjects from 41 different radiology

TABLE 1 | Demographic and clinical data for the considered groups of participants.

Group type	# Subjects	Age mean ± std [range]	Gender # Males/# Females	MMSE score mean ± std [range]	# Centers
CN	162	76.3 ± 5.4 [60–90]	76 M/86 F	29.2 ± 1.0 [25–30]	40
MCInc	134	74.5 ± 7.2 [58–88]	84 M/50 F	27.2 ± 1.7 [24–30]	36
MCIc	76	74.8 ± 7.4 [55–88]	43 M/33 F	26.5 ± 1.9 [23–30]	30
AD	137	76.0 ±7.3 [55–91]	67 M/70 F	23.2 ± 2.0 [18–27]	39

centers were considered. Identification Numbers (IDs) of each subject involved in this study are reported in Supplementary Tables S1–S4. The ADNI was launched in 2003 by the National Institute on Aging (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the Food and Drug Administration (FDA), private pharmaceutical companies and non-profit organizations, as a $60 million, 5-year public private partnership. The primary goal of ADNI has been to test whether serial MR, PET, other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of MCI and early AD.

According to the ADNI inclusion criteria, enrolled subjects were all between 55 and 90 years of age and spoke either English or Spanish. Each subject was willing, able to perform all test procedures described in the protocol and had a study partner able to provide an independent evaluation of functioning. Inclusion criteria for CN were: Mini Mental State Examination (MMSE) scores between 24 and 30; Clinical Dementia Rating (CDR) (Morris, 1993) of zero; absence of depression, MCI and dementia. Inclusion criteria for MCI were: MMSE scores between 24 and 30; CDR of 0.5; objective memory loss, measured by education adjusted scores on Wechsler Memory Scale Logical Memory II (Wechsler, 1987), absence of significant levels of impairment in other cognitive domains; absence of dementia. Inclusion criteria for AD were: MMSE scores between 20 and 26; CDR of 0.5 or 1.0; NINCDS/ADRDA criteria for probable AD (McKhann et al., 1984; Dubois et al., 2007). Detailed description of inclusion/exclusion criteria can be found in the ADNI protocol (http://www.adni-info.org/Scientists/ADNIStudyProcedures.aspx).

MR Images

T1-weighted structural MR images of all selected subjects were obtained from the ADNI dataset. In order to allow standardization of images from different sites and platforms, we only used images which had undergone: (1) geometry correction for gradient nonlinearity, by 3D gradwarp correction (Jovicich et al., 2006); and (2) intensity correction for non-uniformity, by B1 non-uniformity correction (Narayana et al., 1988). T1-weighted structural MR images of each subject were acquired according to the ADNI acquisition protocol (Jack et al., 2008). MR imaging examinations were performed at 1.5 T. Scans from the baseline visit (when available) or from the screening visit. According to the ADNI protocol, MR imaging examination was performed twice per visit. Scans were then rated by the ADNI investigators of the ADNI MR imaging quality control center at the Mayo Clinic on the basis of blurring/ghosting, flow artifact, intensity, and homogeneity, signal-to-noise ratio (SNR), susceptibility artifacts, and gray-white/cerebrospinal fluid contrast (Jack et al., 2008). In this work, we used the image which was rated as the *best quality scan* for each subject. 3D MR images were downloaded from the ADNI dataset in 3D NIfTI format.

A pre-processing procedure, which mainly aimed at the spatial normalization of all MR images by co-registration to a standard template was applied. All pre-processing procedures were applied to MR images by means of the VBM8 software package (Ashburner and Friston, 2000). First steps of pre-processing consisted in: (1) image re-orientation; (2) cropping; (3) skull-stripping; (4) image normalization to MNI standard space, which was performed by co-registration to the MNI template (MNI152 T1 1 mm brain) (Grabner et al., 2006; O'Hanlon et al., 2013). After co-registration to the MNI template, MR images had a size of 121 × 145 × 121 voxels. Each image was then segmented into Gray Matter (GM) and White Matter (WM) tissue probability maps. Resulting images (whole-brain, GM and WM) were smoothed using an isotropic Gaussian kernel with Full Width at Half Maximum (FWHM) ranging from 2 to 12 mm^3 (step: 2 mm^3).

The Classifier

In order to classify the different groups of subjects by means of their T1-weighted structural (whole-brain, GM and WM) we used a machine learning classifier previously implemented by our group (Salvatore et al., 2014). The whole process consists of 2 steps: (1) feature extraction and selection from the MR images of the subjects, which aimed at the selection of the most discriminative features by Principal Components Analysis (PCA) coupled with a Fisher Discriminant Ratio (FDR) criterion (López et al., 2011), and (2) single-subject classification, which aimed at the classification of the subjects on the basis of a predictive model generated for the separation of the different subject groups by means of the most discriminative features (Klöppel et al., 2008; Salvatore et al., 2014).

Feature Extraction and Selection

In order to identify the most discriminative features among groups, an automatic feature extraction technique was applied to MR images (whole-brain, GM and WM). This technique also allowed to reduce the number of features to be handled without losing relevant information for discrimination, and thus to enhance computational performances of the machine learning algorithm.

PCA was implemented to perform feature extraction (Habeck et al., 2008; López et al., 2011). This technique is based on two consecutive steps: (1) application of an orthogonal transformation to a dataset of (possibly) correlated variables; this operation results in a set of values of orthogonal (uncorrelated) variables, which are referred to as Principal Components of the original dataset and which define the so-called PCA subspace; (2) projection of each variable of the original dataset onto the PCA subspace; this operation results in the reduction of the original set of observed variables into a smaller set of features, which are referred to as PCA coefficients and which can be used in subsequent analyses. The total number of PCA coefficients is equal to the number of Principal Components extracted from the original dataset.

Mathematically, if we consider a dataset A composed of S samples, with each sample being a collection of N variables, then the dimension of the dataset is S × N. By computing the eigenvectors of the covariance matrix of the dataset A, PCA subspace can be defined as the space spanned by these eigenvectors. Application of PCA to the dataset A results in a number of Principal Components (i.e., of eigenvectors) with non-zero eigenvalues which is at most equal to the value of the smaller dimension of the dataset–1. Principal Components are sorted in descending order according to the proportion of variance explained, with the constraint for them to be orthogonal with each other.

In this study, datasets were composed of S samples (MR images), where the dimension N of each pre-processed MR image was 121 × 145 × 121 voxels. Application of PCA to our datasets resulted in a number of Principal Components with non-zero eigenvalues which was at most equal to the number S of samples in each dataset–1. The dimension of each dataset after application of PCA was S × (S – 1).

PCA coefficients resulting from the feature extraction process were then sorted in a descending order according to their FDR, which gives information about the class discriminatory power of a given component. For each component, FDR can be calculated as follows:

$$FDR = \frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2} \quad (1)$$

where μ_i and σ_i^2 are the mean and the variance of the ith class, respectively.

The explained variance was studied as a function of the number of considered principal components before and after sorting them in accordance to their FDR, in order to show the impact of FDR-analysis on PCA coefficients.

Classification

The classification algorithm of the proposed machine learning method was based on Support Vector Machines (SVM) (Schölkopf and Smola, 2002). The aim of SVM is to find a predictive model which is able to perform binary group separation. This predictive model is represented by a hyper-plane which can be computed using a set of data input to SVM for its training (training data). The set of training data consists of: (1) a vector of samples belonging to two different classes and (2) the corresponding vector of labels (two labels, each label identifies one class). SVM is able to compute a predictive model for the classification of a new sample to one or to the other of the two classes. Specifically, the predicted class y for the sample x is given by the following relation:

$$y(x) = \sum_{n=1}^{N} w_n \cdot t_n \cdot k(x, x_n) + b \quad (2)$$

where N is the number of samples included in the training set; w_n is a weight assigned by SVM to each sample n in the training set during the training phase; t_n is the label of the sample n of the training set; $k(x, x_n)$ is a kernel function; b is a threshold parameter. The main difference among SVM and other classification methods is that the hyper-plane computed by SVM is the one which maximizes the separation between the two classes.

In this work, we used the Matlab platform to both implement and optimize the SVM classifier. We used a linear kernel for all analyses. Our code also included algorithms of the biolearning toolbox of Matlab.

Optimization of Classification and Evaluation of Accuracy

An optimization was performed with the purpose of finding the best parameter configuration for the classification of the different groups of subjects. A Nested Cross Validation (Nested CV) was used. In this approach, the original dataset is split into k subsets of (possibly) equal size. An inner training-and-validation loop for parameter estimation and optimization is performed using k-1 subsets, and an outer test loop for performance evaluation is performed using the kth held-out subset. This procedure is then repeated k times, until all k subsets are used once for performance evaluation.

In this study, we performed nested 20-fold CV using 19/20 of the original data for the inner training and validation loop allowing parameter estimation and optimization. For each inner loop, these 19/20 subsets were randomly split in half in order to perform training and validation on two independent datasets. The trained and optimized model and parameter set were then used to predict the held-out 1/20 subset.

For each round, the optimal parameters (which brain tissue, which level of filtering, how many PCA coefficients) were chosen as those for which the classification error (E) was minimized. Specifically, we aimed at minimizing the quantity given by,

$$E = 1 - Balanced\ Accuracy \quad (3)$$

$$Balanced\ Accuracy = \frac{1}{2}\left(Specificity + Sensitivity\right) \quad (4)$$

as a function of the following parameters: (1) tissue map (whole-brain, GM and WM); (2) smoothing (FWHM = 2, 4, 6, 8, 10, 12 mm^3, or no smoothing); (3) number of PCA coefficients (from 1 to PC, where PC is the total number of extracted coefficients).

For each of the 20 separate rounds of the outer loop, Balanced Accuracy was calculated and results were averaged across all 20 rounds (Overall Balanced Accuracy).

Parameter optimization and accuracy evaluation were performed for the three following comparisons: (1) AD vs. CN, (2) MCIc vs. CN, and (3) MCIc vs. MCInc.

It is worth noting that pre-processing, feature extraction and feature selection steps were performed separately on the datasets used in the inner training-and-validation loop and in the outer test loop, in order to avoid over-fitting problems (Kuncheva, 2004).

Diagnostic MR-related Biomarkers

Extraction of MR-related biomarkers was carried out according to the following procedure. For each round of the inner training-and-validation loop, maps of voxel-based pattern distribution of MR image differences among groups of subjects were generated for the optimized configuration (minimum E), thus obtaining 20 maps. These maps were averaged in order to obtain the final map. This procedure was applied to the three following comparisons: (1) AD vs. CN, (2) MCIc vs. CN, and (3) MCIc vs. MCInc.

The importance of each considered sample for the classification was computed on the basis of the predictive model generated by our SVM (Klöppel et al., 2008; Focke et al., 2011; Salvatore et al., 2014). As specified in Equation (2), the weight w_n, assigned by the SVM to the sample n during the training phase of the classification, indicates the importance of that sample for the computation of the separating hyper-plane and, thus, indicates the importance of that sample for the separation of the two considered groups. It is worth noting that the weight w_n assigned by SVM to the sample n is non-zero only for support vectors, being respectively positive or negative depending on the class to which the sample n belongs. Each sample n of the training set was multiplied by the corresponding assigned weight w_n. Resulting weighted samples were added in order to generate a vector representing the weight of each feature for the classification. In order to ensure the correct interpretation of weights assigned by SVM, we then applied the method proposed by Haufe and colleagues to compute activation patterns for backward models as described in Haufe et al. (2014). The computed pattern was finally transformed back from the PCA space to the MR-images space, resulting in a map of voxel-based pattern distribution of MR image differences among groups.

Voxel-based pattern distribution (normalized to a range between 0 and 1) was represented by a proper color scale and superimposed on a standard stereotactic brain for spatial localization. In this way, MR-related diagnostic biomarkers for the diagnosis of AD (AD vs. CN) and for the early diagnosis of AD (MCIc vs. CN, and MCIc vs. MCInc) were identified.

Results

Participants

Groups of participants did not show significant differences for age (Student's t-test with significance level at 0.05) and gender (Pearson's chi-square test with significance level at 0.05). Significant differences for MMSE scores were found between CN and patients (AD, MCIc) (Student's t-test with $p < 0.0001$), consistently with previous studies considering the same groups of ADNI subjects (Cuingnet et al., 2011).

MR Images

Co-registration of all MRI images to the MNI template and segmentation into GM and WM tissue probability maps were performed correctly. **Figure 1** shows results of these procedures for a representative MR image of a MCIc patient. Sagittal view of the original volume (A), the slice co-registered to the MNI space (B), the GM tissue probability map (C) and the WM tissue probability map (D) are shown.

The Classifier

Feature Extraction and Selection

Figure 2 shows a representative example of PCA coefficients resulting from the feature extraction and selection obtained from the comparison between AD and CN. 1st, 2nd, and 3rd components are shown when using GM tissue probability map and an isotropic Gaussian kernel with 10 mm^3 FWHM for smoothing. The number of the extracted PC was 141.

Figures 3, **4** show, as representative examples, the explained variance as a function of the number of considered PCs, before (**Figure 3**) and after (**Figure 4**) sorting them in accordance to their FDR. Plots are shown for the comparisons between AD and CN, MCIc, and CN, MCIc, and MCInc when using GM tissue probability maps and no smoothing. The trend of explained variance as a function of the number of considered PCs was modified by the application of FDR-analysis. In particular, FDR-analysis allowed the most discriminative information for class separation to be contained in the first few principal components. This is shown by the step in the explained variance in correspondence with a low number of components for the comparisons between AD vs. CN and MCIc vs. CN (**Figure 4**).

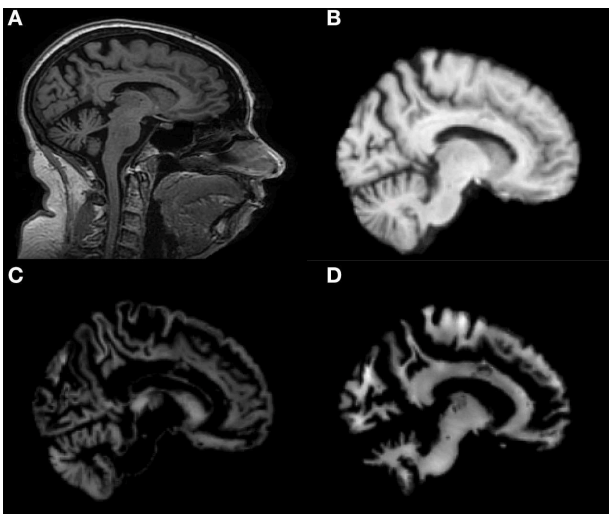


FIGURE 1 | Sagittal image of a MR scan from a MCIc patient: (A) original image; (B) same slice, deskulled and co-registered to the MNI space; same slice, segmented into Gray Matter (GM) (C) and into White Matter (WM) (D).

Classification

In **Figure 5**, a representative example of the hyper-plane separating AD from CN subjects is shown when using 3 PCA coefficients, GM tissue probability map and an isotropic Gaussian kernel with 10 mm^3 FWHM for smoothing. The number of subjects involved was 142 (67 AD, 75 CN) and the total number of extracted PCA coefficients was 141.

Optimization of Classification and Evaluation of Accuracy

Figures 6–8 show E (1 – Balanced Accuracy) as a function of applied smoothing (FWHM – mm^3) and number of PCA coefficients when using GM tissue probability maps. Plots are shown for the comparisons between AD and CN, MCIc, and CN, MCIc, and MCInc.

Optimal parameters resulting from classifier optimization are reported in **Table 2**. For all the comparisons, minimum values of E were obtained mostly when using GM tissue probability maps (frequency of 100% for AD vs. CN, 85% for MCIc vs. CN, 80% for MCIc vs. MCInc). For the comparison between AD and CN, the best set of optimal parameters among the 20 rounds was: GM tissue probability map; 10 mm^3 FWHM of the isotropic Gaussian kernel for smoothing; 127 PCA coefficients. When using these parameters, E reached its minimum value of 0.08. For the comparison between MCIc and CN, the best set of optimal parameters among the 20 rounds was: GM tissue probability map; 6 mm^3 FWHM of the isotropic Gaussian kernel for smoothing; 67 PCA coefficients. When using these parameters, E reached its minimum value of 0.14. For the comparison between MCIc and MCInc, the best set of optimal parameters among the 20 rounds was: GM tissue probability map; 2 mm^3 FWHM of the isotropic Gaussian kernel for smoothing; 34 PCA coefficients. When using these parameters, E reached its minimum value of 0.27.

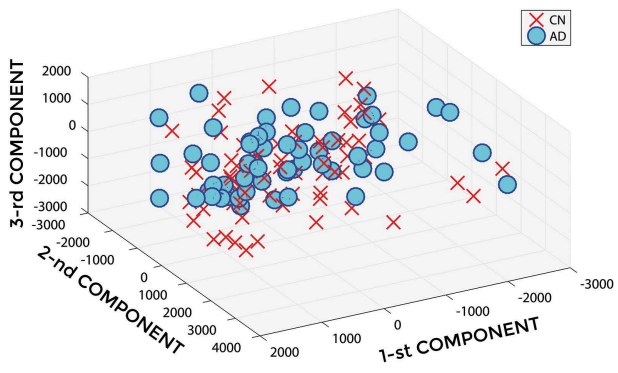


FIGURE 2 | PCA coefficients for the comparison between AD (o symbol) and CN (× symbol) when using GM tissue probability map and an isotropic Gaussian kernel with 10 mm^3 FWHM for smoothing. 1st, 2nd, and 3rd components are shown.

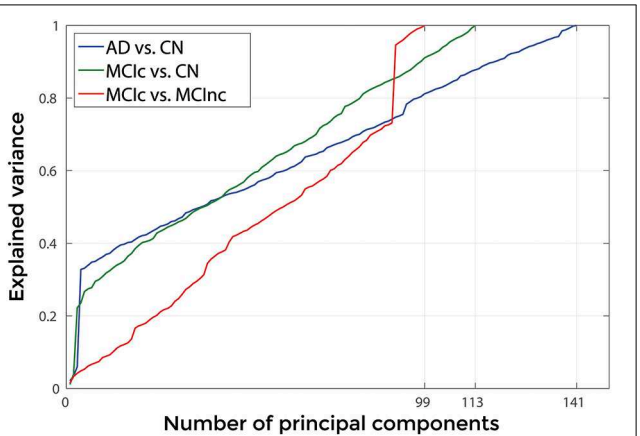


FIGURE 4 | Explained Variance as a function of the number of considered principal components sorted in accordance to their FDR, when using GM tissue probability map and no smoothing, for the following comparisons: AD vs. CN, MCIc vs. CN, MCIc vs. MCInc.

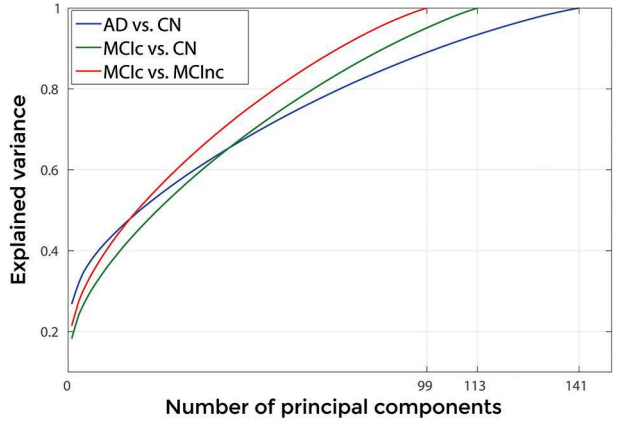


FIGURE 3 | Explained Variance as a function of the number of considered Principal Components, when using GM tissue probability map and no smoothing, for the following comparisons: AD vs. CN, MCIc vs. CN, MCIc vs. MCInc.

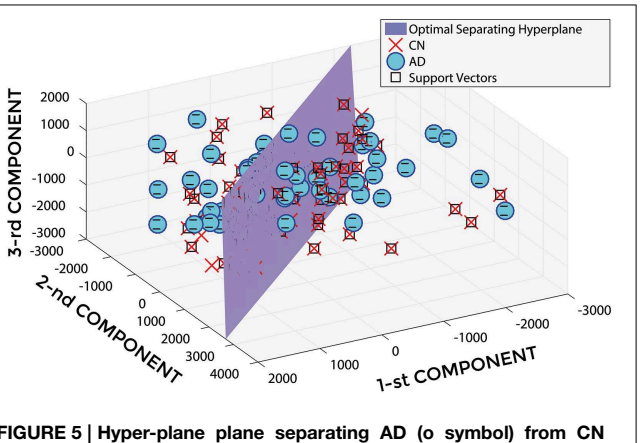


FIGURE 5 | Hyper-plane plane separating AD (o symbol) from CN (× symbol) PCA coefficients (3 PCA coefficients), and defined Support Vectors (□ symbol), when using GM tissue probability map and an isotropic Gaussian kernel with 10 mm^3 FWHM for smoothing. 1st, 2nd, and 3rd components are shown.

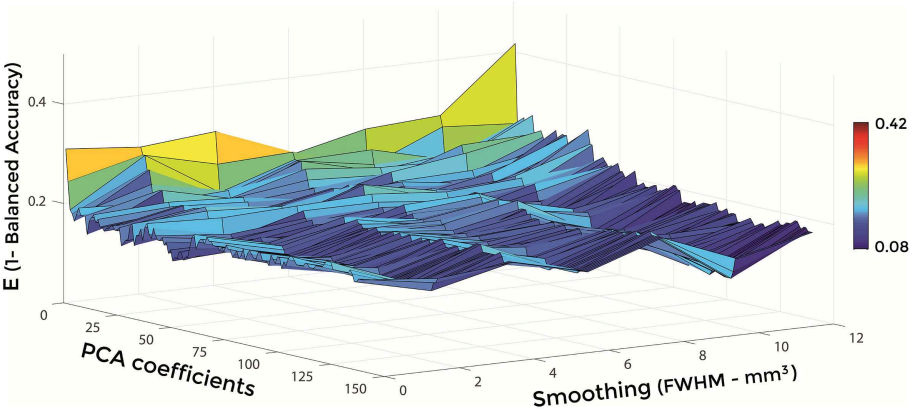


FIGURE 6 | E (1 – Balanced Accuracy) as a function of smoothing (FWHM – mm^3) and number of PCA coefficients for the comparison between AD and CN when using GM.

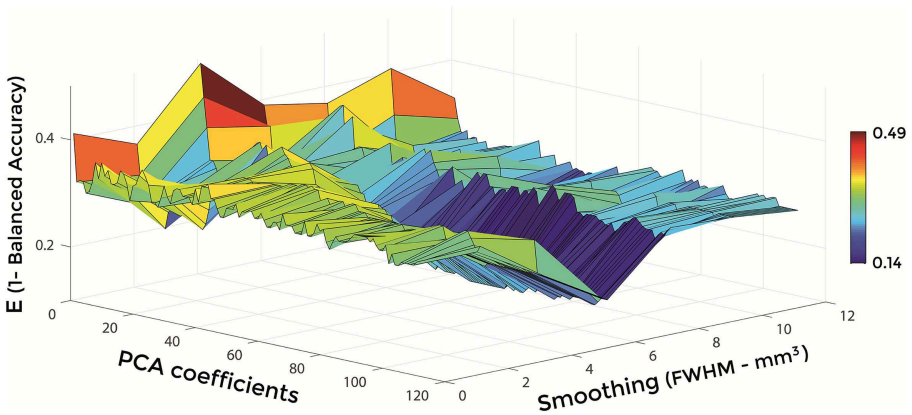


FIGURE 7 | E (1 – Balanced Accuracy) as a function of smoothing (FWHM – mm^3) and number of PCA coefficients for the comparison between MCIc and CN when using GM.

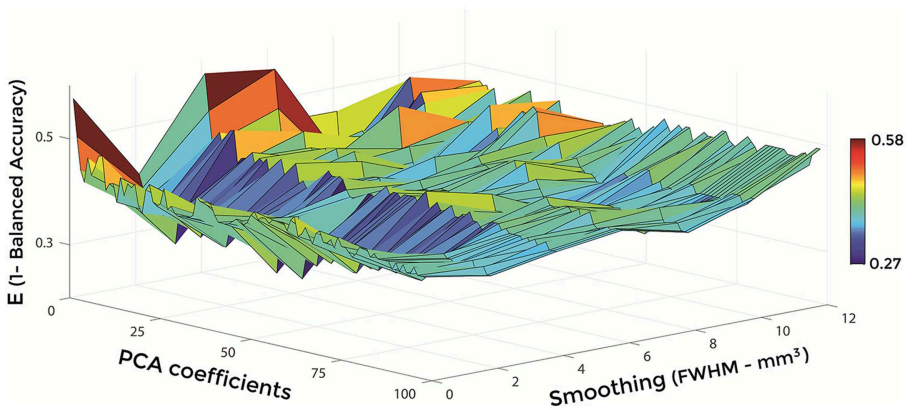


FIGURE 8 | E (1 – Balanced Accuracy) as a function of smoothing (FWHM – mm^3) and number of PCA coefficients for the comparison between MCIc and MCInc when using GM.

TABLE 2 | Classification error and optimal parameters (Tissue map, Smoothing, Number of PCA coefficients) for each of the 20 rounds of the inner training-and-validation loop (best configuration in bold).

Comparison	E	Tissue map	Smoothing FWHM [mm^3]	PCA coefficients
AD vs. CN	0.10	GM	6	6
	0.08	**GM**	**10**	**127**
	0.12	GM	10	41
	0.11	GM	4	62
	0.11	GM	6	75
	0.15	GM	2	64
	0.13	GM	8	69
	0.12	GM	4	32
	0.12	GM	2	67
	0.11	GM	2	50
	0.12	GM	4	48
	0.09	GM	4	54
	0.13	GM	8	35
	0.12	GM	2	118
	0.12	GM	4	46
	0.13	GM	2	22
	0.12	GM	2	135
	0.15	GM	6	49
	0.11	GM	6	54
	0.12	GM	12	30
MCIc vs. CN	0.19	GM	8	26
	0.17	GM	2	25
	0.20	GM	2	94
	0.19	GM	4	53
	0.22	WB	2	14
	0.20	WB	10	57
	0.15	GM	4	62
	0.15	GM	10	22
	0.21	GM	10	75
	0.19	GM	10	32
	0.14	**GM**	**6**	**67**
	0.19	GM	4	13
	0.19	WB	6	64
	0.17	GM	10	80
	0.22	GM	8	28
	0.18	GM	12	21
	0.19	GM	12	16
	0.19	GM	10	81
	0.19	GM	8	76
	0.19	GM	8	101
MCIc vs. MCInc	0.30	GM	2	9
	0.31	GM	10	19
	0.33	WB	12	34
	0.34	GM	4	34
	0.32	GM	8	16
	0.30	GM	6	17
	0.33	WB	2	21
	0.28	GM	6	10
	0.27	**GM**	**2**	**34**
	0.31	WM	4	4
	0.31	GM	2	16
	0.32	GM	8	31
	0.32	GM	8	23
	0.30	GM	4	46
	0.34	GM	8	33
	0.33	GM	8	2
	0.32	WB	4	34
	0.28	GM	10	5
	0.30	GM	8	8
	0.30	GM	2	84

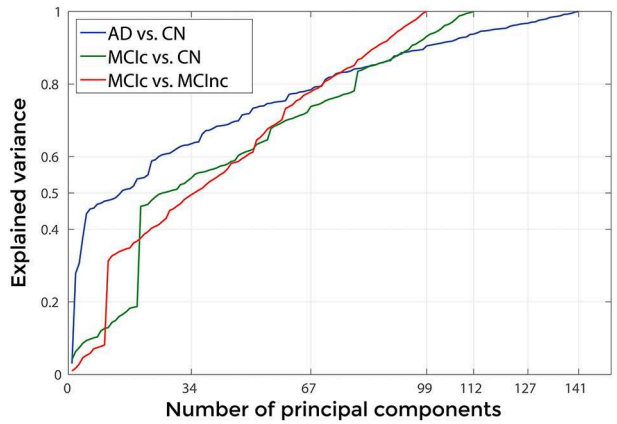


FIGURE 9 | Explained Variance, when using the best set of optimal parameters, as a function of the number of considered principal components sorted in accordance to their FDR, for the following comparisons: AD vs. CN, MCIc vs. CN, MCIc vs. MCInc.

The operational time required by the whole pre-processing and training of the classifier (including feature extraction and selection) using the best set of optimal parameters, as measured by the *tic* and *toc* functions implemented in Matlab (version R2015a) and running on a system with 32 CPUs at 2.00 GHz, was 31.7s for the comparison between AD and CN, 21.7s for the comparison between MCIc and CN and 21.2s for the comparison between MCIc and MCInc. The testing phase, including preprocessing and classification of the new dataset, took 1.5s per subject on average.

The Overall Balanced Accuracy (averaged across all 20 rounds) was 0.76 ± 0.11 for the classification of AD vs. CN, 0.72 ± 0.12 for the classification of MCIc vs. CN, 0.66 ± 0.16 for the classification of MCIc vs. MCInc, respectively.

Since MMSE resulted significantly different between CN and patients (AD, MCIc), we have also tested our classification algorithm after incorporating MMSE as additional feature. Balanced Accuracy resulted to be affected (from 0.76 ± 0.11 to 0.99 ± 0.03 for AD vs. CN, from 0.72 ± 0.12 to 0.78 ± 0.16 for MCIc vs. CN, from 0.66 ± 0.16 to 0.60 ± 0.17 for MCIc vs. MCInc).

Figure 9 shows the explained variance as a function of the number of considered PC sorted in accordance to their FDR. Plots are shown for the comparisons between AD and CN, MCIc and CN, MCIc and MCInc when using the best configuration highlighted in **Table 2**. For AD vs. CN comparison, the percentage of variance explained by the first 127 components was 98%; for MCIc vs. CN comparison, the percentage of variance explained by the first 67 components was 74%; for MCIc vs. MCInc comparison, the percentage of variance explained by the first 34 components was 50%.

Diagnostic MR-related Biomarkers

Figures 10–12 show voxel-based pattern distribution maps for the three following classification: (1) AD vs. CN, (2) MCIc vs. CN, (3) MCIc vs. MCInc. The pattern of differences (normalized

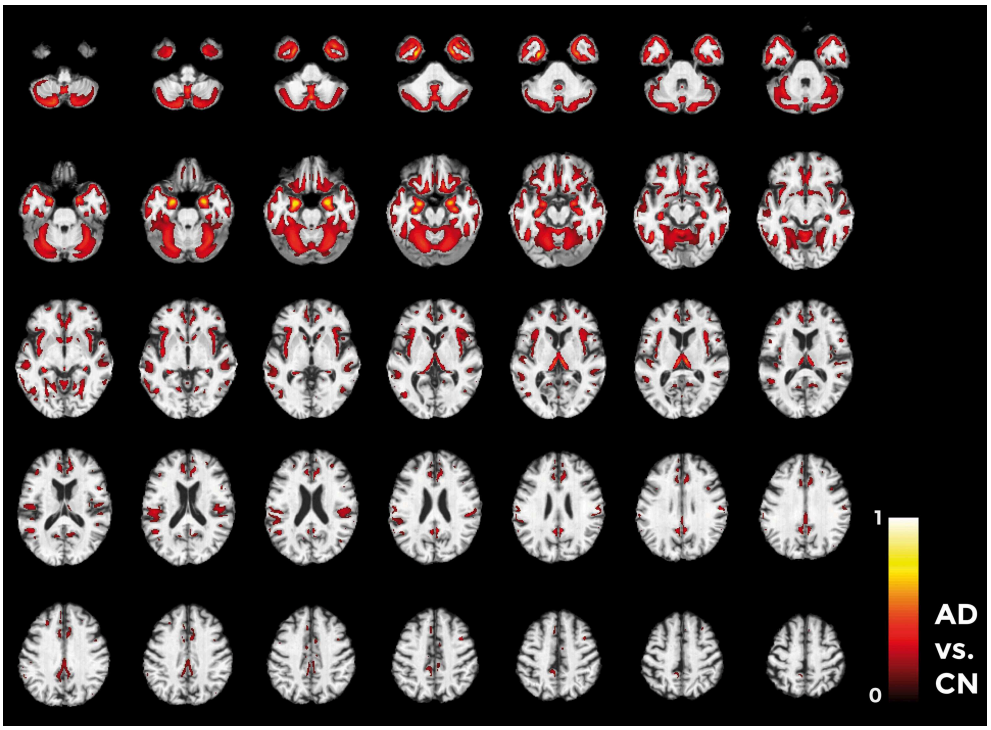


FIGURE 10 | Voxel-based pattern distribution map (axial view) for the classification between AD and CN. Voxel-based pattern distribution (normalized to a range between 0 and 1) is expressed according to the color scale (threshold = 50%) and superimposed on a standard stereotactic brain for spatial localization.

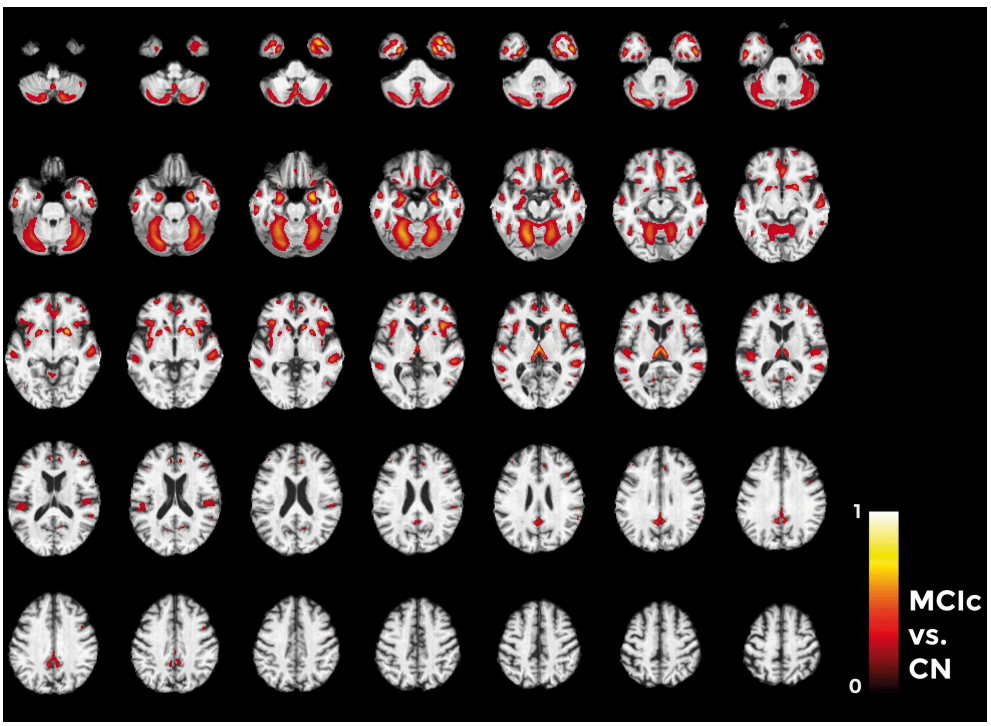


FIGURE 11 | Voxel-based pattern distribution map (axial view) for the classification between MCIc and CN. Voxel-based pattern distribution (normalized to a range between 0 and 1) is expressed according to the color scale (threshold = 45%) and superimposed on a standard stereotactic brain for spatial localization.

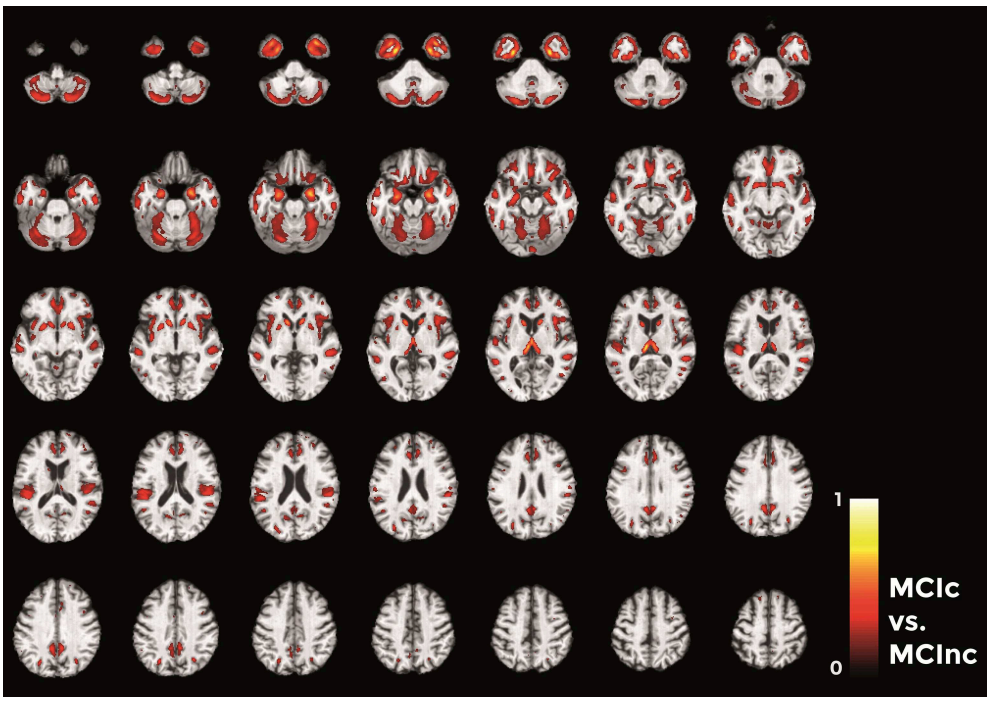


FIGURE 12 | Voxel-based pattern distribution map (axial view) for the classification between MCIc and MCInc. Voxel-based pattern distribution (normalized to a range between 0 and 1) is expressed according to the color scale (threshold = 35%) and superimposed on a standard stereotactic brain for spatial localization.

to a range between 0 and 1) is expressed according to the color scales.

Voxels influencing the classification of AD with respect to CN (**Figure 10**) are localized in the temporal pole, superior temporal cortex, medial temporal cortex including hippocampus and entorhinal cortex, amygdala, thalamus, putamen, caudate, insular cortex, gyrus rectus, lateral orbitofrontal cortex, inferior frontal cortex, superior frontal cortex, anterior cingulate cortex, precuneus, and in the posterior cerebellar lobule.

Considering the comparison between MCIc and CN individuals (**Figure 11**), the major part of voxel-based pattern distribution was similar to the one previously found in AD.

Finally, in the direct comparisons between the two MCI groups (**Figure 12**) we detected only voxels influencing classification of the MCIc with respect to MCInc. In other words, there were no anatomical changes in the MCInc's brain useful to increase the accuracy of classification. Overall, the major part of voxel-based pattern distribution was similar to the one detected in the previous MCIc vs. CN contrast.

Discussion and Conclusions

The localization and spatial extent of the anatomical features identified in our study are in line with previous research showing the precedence of pathologic changes in the temporal and parietal cortex (Braak and Braak, 1991; Schroeter et al., 2009). In fact, a recent neuroimaging meta-analysis (Schroeter et al., 2009) aimed at characterizing the prototypical neural substrates of AD and its prodromal stage amnestic MCI reported the presence of:

(a) Reduction of glucose utilization and perfusion in the inferior parietal lobules and the posterior cingulate cortex and precuneus; hypometabolism was detected in the left anterior superior insula; whereas gray matter atrophy was found in the left temporal pole/anterior superior temporal sulcus, right amygdala, and gyrus rectus when 525 MCI patients were compared with 1097 healthy controls.

(b) Reductions in glucose utilization and perfusion coincided in the inferior parietal lobules, posterior superior temporal sulcus, precuneus, posterior cingulate cortex, anterior medial frontal cortex, anterior cingulate gyrus and right inferior temporal sulcus; hypometabolism was in the right frontal pole, left posterior middle frontal gyrus and left hippocampal head; whereas gray matter atrophy was found in the both amygdalae, both anterior hippocampal formations, entorhinal areas, medial thalamus, posterior insula, and left middle temporal gyrus/superior temporal sulcus when 826 AD patients were compared with 1097 healthy controls.

The only brain region revealed by our pattern recognition analysis not typically related to AD-like atrophy was the cerebellum. The cerebellum is a region generally rather neglected in AD research. Atrophy of this region has been sparsely reported in neuroimaging studies (Thomann et al., 2008a; Nigro et al., 2014), although there is considerable number of histo-pathological studies that demonstrated the presence of

degenerative changes (Li et al., 1999; Wegiel et al., 2000; Wang et al., 2002). These alterations mainly comprise reduced Purkinje cell density, atrophy of the molecular and granular cell layer as well as a large number of amyloid plaques in the cerebellar cortex of AD compared to controls. Moreover the fact that we detected only anatomical changes in the posterior lobule of the cerebellum corroborated our findings, since cognitive performance in AD patients was found to be significantly correlated with volumes of posterior cerebellar lobes (Thomann et al., 2008b).

The development of computer-based automatic methods for the accurate classification of patients in early phase of AD from imaging data has attracted strong interest from the clinical community in the last few years, since its possible critical impact on clinical management and practice (i.e., identification of new biomarkers). Many of these classification methods are based on SVM, a set of algorithms that uses supervised learning of pattern recognition in a training set to build a classifier able to predict the category to which a new example belongs. One of the most important challenging in this field of study is to define automated methods to discriminate MCI patients progressing later to AD from patients who will not (Schroeter et al., 2009). For this reason, this study was aimed at assessing the powerful of machine learning methods in discriminating MCI at a risk state of AD.

In our work we used nested CV to measure the performance of our classifier. Nested CV avoids optimistically biased estimates of performance that may arise from the use of the same CV for parameter estimation and performance evaluation. Specifically, when model parameters are estimated by means of the performance evaluation criterion, then these estimates depend on (1) improvements in generalization performance and (2) statistical features of the particular dataset on which the performance are evaluated. This may result in under-estimates of the CV error. Moreover, in ordinary CV, parameter estimation is performed prior to model building, which could lead to an optimistic evaluation of the performance of the classifier. On the other side, in nested CV parameter estimation is performed simultaneously to performance evaluation (Cawley and Talbot, 2010).

Performances of our classification algorithm evaluated by nested 20-fold CV were 0.76 for AD vs. CN, 0.72 for MCIc vs. CN, and 0.66 for MCIc vs. MCInc. In their published study, Cuingnet et al. (2011) evaluated the performance of ten different machine learning methods (28 algorithm configurations) by using the same group of ADNI subjects employed in our work, splitting datasets in two equal sample groups and using one group to estimate the optimal value of hyperparameters and the other group to evaluate the performance of the classifier. Performances reached by our algorithm for the three classifications (AD vs. CN, MCIc vs. CN, and MCIc vs. MCInc) are better than 27/28 algorithm configurations, since 27 algorithms have a Balanced Accuracy lower than 0.66 for the MCIc vs. MCInc comparison.

The use of our classifier is limited to the early diagnosis of AD. Notwithstanding the vast majority of brain regions identified by our multivariate pattern recognition analysis have been also described to be involved in neurodegenerative processes underlying other dementia disorders (e.g., Fronto-Temporal Lobar Degeneration), machine learning has been also found accurate when applied to MR images for the differential diagnosis of AD (e.g., Klöppel et al., 2008). The clinical use of such a machine learning approach (early and differential diagnosis of AD) should require the training of a multicategory classifier (Beom Choi et al., 2014) on MR images from CN and different dementia patients (e.g., MCIc, MCInc, AD, FTD).

The main innovative result of our work was the extraction of MR-related biomarkers for the early diagnosis of AD by means of machine learning. We assessed the relevance of each brain voxel with respect to the classification analysis, thus allowing regions critically involved in the pathophysiological mechanisms of AD to be identified. Notably, the vast majority of brain regions allowing to perform the best discrimination between AD and CN, as well as between MCIc and CN, were the same regions allowing the discrimination between the two critical forms of MCI, i.e., MCIc and MCInc. In other words, the AD-like atrophy patterns characterized by combined pathological changes within the temporal cortex, hippocampus, entorhinal cortex, thalamus, insular cortex, anterior cingulate cortex, orbitofrontal cortex, and precuneus, allowed distinguishing clinically- and cognitively-matched MCI patients progressing to AD from those who will not.

In conclusion, we demonstrated that an advanced neuroimaging approach based on machine learning is able to accurately classify patients who will or will not develop AD by means of structural MRI data and to extract MR-related biomarkers of AD. Moreover, our advanced neuroimaging study allows us to perform a challenging reflection. Due to the similarity between AD-like atrophy patterns with those detected in MCI who will convert in AD, we can derive that the machine learning approach impacts on the sensitivity of AD-related features rather than specificity. This would suggest that the problem of how to perform diagnosis of AD at a very early stage by MRI seems to be a matter of increasing the MRI detectability of structural biomarkers. For this reason, both current generation MRI systems combined with advanced images processing algorithms and future generation MRI systems with improved sensitivity (e.g., increased MRI resolution and better S/N ratio) will –definitely– move MRI diagnostic role from clinical to pre-clinical stage of AD.

Acknowledgments

This study is part of the CNR Research Project on Aging—PNR.

Data collection and sharing for this study was funded by the Alzheimer's Disease Neuroimaging initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen Idec Inc.; Bristol-Myers Squibb Company; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare;

IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Medpace, Inc.; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Synarc Inc.; and Takeda Pharmaceutical Company. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

Supplementary Material

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fnins.2015.00307

References

Albert, M. S., DeKosky, S. T., Dickson, D., Dubois, B., Feldman, H. H., Fox, N. C., et al. (2011). The diagnosis of mild cognitive impairment due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimers. Dement.* 7, 270–279. doi: 10.1016/j.jalz.2011.03.008

Ashburner, J., and Friston, K. J. (2000). Voxel-based morphometry—the methods. *Neuroimage* 11, 805–821. doi: 10.1006/nimg.2000.0582

Beom Choi, S., Park, J. S., Chung, J. W., Yoo, T. K., and Kim, D. W. (2014). "Multicategory classification of 11 neuromuscular diseases based on microarray data using support vector machine," in *Engineering in Medicine and Biology Society (EMBC), 2014 36th Annual International Conference of the IEEE* (Chicago, IL: IEEE), 3460–3463.

Blennow, K., de Leon, M. J., and Zetterberg, H. (2006). Alzheimer's disease. *Lancet* 368, 387–403. doi: 10.1016/S0140-6736(06)69113-7

Braak, H., and Braak, E. (1991). Neuropathological stageing of Alzheimer-related changes. *Acta Neuropathol.* 82, 239–259. doi: 10.1007/BF00308809

Busatto, G. F., Garrido, G. E., Almeida, O. P., Castro, C. C., Camargo, C. H., Cid, C. G., et al. (2003). A voxel-based morphometry study of temporal lobe gray matter reductions in Alzheimer's disease. *Neurobiol. Aging* 24, 221–231. doi: 10.1016/S0197-4580(02)00084-2

Cawley, G. C., and Talbot, N. L. (2010). On over-fitting in model selection and subsequent selection bias in performance evaluation. *J. Mach. Learn. Res.* 11, 2079–2107.

Cuingnet, R., Gerardin, E., Tessieras, J., Auzias, G., Lehéricy, S., Habert, M. O., et al. (2011). Automatic classification of patients with Alzheimer's disease from structural MRI: a comparison of ten methods using the ADNI database. *Neuroimage* 56, 766–781. doi: 10.1016/j.neuroimage.2010.06.013

Davatzikos, C., Fan, Y., Wu, X., Shen, D., and Resnick, S. M. (2008). Detection of prodromal Alzheimer's disease via pattern classification of magnetic resonance imaging. *Neurobiol. Aging* 29, 514–523. doi: 10.1016/j.neurobiolaging.2006.11.010

Dubois, B., Feldman, H. H., Jacova, C., DeKosky, S. T., Barberger-Gateau, P., Cummings, J., et al. (2007). Research criteria for the diagnosis of Alzheimer's disease: revising the NINCDS–ADRDA criteria. *Lancet Neurol.* 6, 734–746. doi: 10.1016/S1474-4422(07)70178-3

Focke, N. K., Helms, G., Scheewe, S., Pantel, P. M., Bachmann, C. G., Dechent, P., et al. (2011). Individual voxel—based subtype prediction can differentiate progressive supranuclear palsy from idiopathic parkinson syndrome and healthy controls. *Hum. Brain Mapp.* 32, 1905–1915. doi: 10.1002/hbm.21161

Fox, N. C., and Schott, J. M. (2004). Imaging cerebral atrophy: normal ageing to Alzheimer's disease. *Lancet* 363, 392–394. doi: 10.1016/S0140-6736(04)15441-X

Gerardin, E., Chételat, G., Chupin, M., Cuingnet, R., Desgranges, B., Kim, H. S., et al. (2009). Multidimensional classification of hippocampal shape features discriminates Alzheimer's disease and mild cognitive impairment from normal aging. *Neuroimage* 47, 1476–1486. doi: 10.1016/j.neuroimage.2009.05.036

Grabner, G., Janke, A. L., Budge, M. M., Smith, D., Pruessner, J., and Collins, D. L. (2006). "Symmetric atlasing and model based segmentation: an application to the hippocampus in older adults," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2006* (Berlin; Heidelberg: Springer), 58–66.

Habeck, C., Foster, N. L., Perneczky, R., Kurz, A., Alexopoulos, P., Koeppe, R. A., et al. (2008). Multivariate and univariate neuroimaging biomarkers of Alzheimer's disease. *Neuroimage* 40, 1503–1515. doi: 10.1016/j.neuroimage.2008.01.056

Haufe, S., Meinecke, F., Görgen, K., Dähne, S., Haynes, J. D., Blankertz, B., et al. (2014). On the interpretation of weight vectors of linear models in multivariate neuroimaging. *Neuroimage* 87, 96–110. doi: 10.1016/j.neuroimage.2013.10.067

Hidalgo-Muñoz, A. R., Ramírez, J., Górriz, J. M., and Padilla, P. (2014). Regions of interest computed by SVM wrapped method for Alzheimer's disease examination from segmented MRI. *Front. Aging Neurosci.* 6:20. doi: 10.3389/fnagi.2014.00020

Hyman, B. T., and Trojanowski, J. Q. (1997). Editorial on consensus recommendations for the postmortem diagnosis of Alzheimer disease from the National Institute on Aging and the Reagan Institute Working Group on diagnostic criteria for the neuropathological assessment of Alzheimer disease. *J. Neuropathol. Exp. Neurol.* 56, 1095–1097. doi: 10.1097/00005072-199710000-00002

Ishii, K., Kawachi, T., Sasaki, H., Kono, A. K., Fukuda, T., Kojima, Y., et al. (2005). Voxel-based morphometric comparison between early-and late-onset mild Alzheimer's disease and assessment of diagnostic performance of z score images. *Am. J. Neuroradiol.* 26, 333–340.

Jack, C. R., Bernstein, M. A., Fox, N. C., Thompson, P., Alexander, G., Harvey, D., et al. (2008). The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods. *J. Magn. Reson. Imaging* 27, 685–691. doi: 10.1002/jmri.21049

Jagust, W., Gitcho, A., Sun, F., Kuczynski, B., Mungas, D., and Haan, M. (2006). Brain imaging evidence of preclinical Alzheimer's disease in normal aging. *Ann. Neurol.* 59, 673–681. doi: 10.1002/ana.20799

Jovicich, J., Czanner, S., Greve, D., Haley, E., van der Kouwe, A., Gollub, R., et al. (2006). Reliability in multi-site structural MRI studies: effects of gradient non-linearity correction on phantom and human data. *Neuroimage* 30, 436–443. doi: 10.1016/j.neuroimage.2005.09.046

Karas, G. B., Burton, E. J., Rombouts, S. A. R. B., Van Schijndel, R. A., O'Brien, J., Scheltens, P. H., et al. (2003). A comprehensive study of gray matter loss in patients with Alzheimer's disease using optimized voxel-based morphometry. *Neuroimage* 18, 895–907. doi: 10.1016/S1053-8119(03)00041-7

Klöppel, S., Stonnington, C. M., Chu, C., Draganski, B., Scahill, R. I., Rohrer, J. D., et al. (2008). Automatic classification of MR scans in Alzheimer's disease. *Brain* 131, 681–689. doi: 10.1093/brain/awm319

Knopman, D. S., DeKosky, S. T., Cummings, J. L., Chui, H., Corey–Bloom, J., Relkin, N., et al. (2001). Practice parameter: diagnosis of dementia (an evidence-based review) report of the quality standards subcommittee of the American academy of neurology. *Neurology* 56, 1143–1153. doi: 10.1212/wnl.56.9.1143

Kuncheva, L. I. (2004). *Combining Pattern Classifiers: Methods and Algorithms*. Hoboken, NJ: John Wiley & Sons.

Li, Q., Van Antwerp, D., Mercurio, F., Lee, K. F., and Verma, I. M. (1999). Severe liver degeneration in mice lacking the IκB kinase 2 gene. *Science* 284, 321–325. doi: 10.1126/science.284.5412.321

López, M., Ramírez, J., Górriz, J. M., Álvarez, I., Salas-Gonzalez, D., Segovia, F., et al. (2011). Principal component analysis-based techniques and supervised classification schemes for the early detection of Alzheimer's disease. *Neurocomputing* 74, 1260–1271. doi: 10.1016/j.neucom.2010.06.025

Martin, C. R., Preedy, V. R., and Hunter, R. J. (2012). *Nanomedicine and the Nervous System*. Boca Raton, FL: CRC Press.

McKhann, G., Drachman, D., Folstein, M., Katzman, R., Price, D., and Stadlan, E. M. (1984). Clinical diagnosis of Alzheimer's disease report of the NINCDS—ADRDA Work Group* under the auspices of Department of Health and Human Services Task Force on Alzheimer's disease. *Neurology* 34, 939–939. doi: 10.1212/wnl.34.7.939

Mitchell, A. J., and Shiri—Feshki, M. (2009). Rate of progression of mild cognitive impairment to dementia–meta—analysis of 41 robust inception cohort studies. *Acta Psychiatr. Scand.* 119, 252–265. doi: 10.1111/j.1600-0447.2008.01326.x

Moradi, E., Pepe, A., Gaser, C., Huttunen, H., Tohka, J., and Alzheimer's Disease Neuroimaging Initiative. (2015). Machine learning framework for early MRI-based Alzheimer's conversion prediction in MCI subjects. *Neuroimage* 104, 398–412. doi: 10.1016/j.neuroimage.2014.10.002

Morris, J. C. (1993). The Clinical Dementia Rating (CDR): current version and scoring rules. *Neurology* 43, 2412–2414. doi: 10.1212/WNL.43.11.2412-a

Narayana, P. A., Brey, W. W., Kulkarni, M. V., and Sievenpiper, C. L. (1988). Compensation for surface coil sensitivity variation in magnetic resonance imaging. *Magn. Reson. Imaging* 6, 271–274. doi: 10.1016/0730-725X(88)90401-8

Nigro, S., Cerasa, A., Zito, G., Perrotta, P., Chiaravalloti, F., Donzuso, G., et al. (2014). Fully automated segmentation of the Pons and Midbrain using human T1 MR brain images. *PLoS ONE* 9:e85618. doi: 10.1371/journal.pone.0085618

O'Hanlon, E., Newell, F. N., and Mitchell, K. J. (2013). Combined structural and functional imaging reveals cortical deactivations in grapheme-color synaesthesia. *Front. Psychol.* 4:755. doi: 10.3389/fpsyg.2013.00755

Petersen, R. C., Smith, G. E., Waring, S. C., Ivnik, R. J., Tangalos, E. G., and Kokmen, E. (1999). Mild cognitive impairment: clinical characterization and outcome. *Arch. Neurol.* 56, 303–308. doi: 10.1001/archneur.56.3.303

Salvatore, C., Cerasa, A., Castiglioni, I., Gallivanone, F., Augimeri, A., Lopez, M., et al. (2014). Machine learning on brain MRI data for differential diagnosis of Parkinson's disease and Progressive Supranuclear Palsy. *J. Neurosci. Methods.* 222, 230–237. doi: 10.1016/j.jneumeth.2013.11.016

Schölkopf, B., and Smola, A. J. (2002). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Cambridge, MA: MIT Press.

Schroeter, M. L., Stein, T., Maslowski, N., and Neumann, J. (2009). Neural correlates of Alzheimer's disease and mild cognitive impairment: a systematic and quantitative meta-analysis involving 1351 patients. *Neuroimage* 47, 1196–1206. doi: 10.1016/j.neuroimage.2009.05.037

Sperling, R. A., Aisen, P. S., Beckett, L. A., Bennett, D. A., Craft, S., Fagan, A. M., et al. (2011). Toward defining the preclinical stages of Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimers. Dement.* 7, 280–292. doi: 10.1016/j.jalz.2011.03.003

Thomann, P. A., Schläfer, C., Seidl, U., Dos Santos, V., Essig, M., and Schröder, J. (2008a). The cerebellum in mild cognitive impairment and Alzheimer's disease–a structural MRI study. *J. Psychiatr. Res.* 42, 1198–1202. doi: 10.1016/j.jpsychires.2007.12.002

Thomann, P. A., Toro, P., Dos Santos, V., Essig, M., and Schröder, J. (2008b). Clock drawing performance and brain morphology in mild cognitive impairment and Alzheimer's disease. *Brain Cogn.* 67, 88–93. doi: 10.1016/j.bandc.2007.11.008

Tufail, A. B., Abidi, A., Siddiqui, A. M., and Younis, M. S. (2012). Automatic classification of initial categories of Alzheimer's disease from structural MRI phase images: a comparison of PSVM, KNN and ANN methods. *Age* 2012, 1731.

Wang, J., Xu, G., Gonzales, V., Coonfield, M., Fromholt, D., Copeland, N. G., et al. (2002). Fibrillar inclusions and motor neuron degeneration in transgenic mice expressing superoxide dismutase 1 with a disrupted copper-binding site. *Neurobiol. Dis.* 10, 128–138. doi: 10.1006/nbdi.2002.0498

Wechsler, D., (1987). *Manual: Wechsler Memory Scale-Revised*. San Antonio, TX: Psychological Corporation.

Wegiel, J., Wang, K. C., Tarnawski, M., and Lach, B. (2000). Microglial cells are the driving force in fibrillar plaque formation, whereas astrocytes are a leading factor in plaque degradation. *Acta Neuropathol.* 100, 356–364. doi: 10.1007/s004010000199

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

Frontiers for the early diagnosis of AD by means of MRI brain imaging and Support Vector Machines

Christian Salvatore[a*], Petronilla Battista[a] and Isabella Castiglioni[a]

[a] *Institute of Molecular Bioimaging and Physiology, National Research Council (IBFM-CNR), Segrate, MI, Italy*

** Corresponding author*
Institute of Molecular Bioimaging and Physiology, National Research Council (IBFM-CNR)
Via F.lli Cervi, 93
20090 Segrate, MI, Italy
Tel.: +39 02 21717511
Fax: +39 02 21717558
E-mail: christian.salvatore@ibfm.cnr.it

Abstract: The emergence of Alzheimer's Disease (AD) as a consequence of increasing aging population make urgent the availability of methods for the early and accurate diagnosis. Magnetic Resonance Imaging (MRI) could be used as in vivo, non invasive tool to identify sensitive and specific markers of very early AD progression. In recent years, multivariate pattern analysis (MVPA) and machine-learning algorithms have attracted strong interest within the neuroimaging community, as they allow automatic classification of imaging data with higher performance than univariate statistical analysis. An exhaustive search of PubMed, Web of Science and Medline records was performed in this work, in order to retrieve studies focused on the potential role of MRI in aiding the clinician in early diagnosis of AD by using Support Vector Machines (SVMs) as MVPA automated classification method. A total of 30 studies emerged, published from 2008 to date. This review aims to give a state-of-the-art overview about SVM for the early and differential diagnosis of AD-related pathologies by means of MRI data, starting from preliminary steps such as image pre-processing, feature extraction and feature selection, and ending with classification, validation strategies and extraction of MRI-related biomarkers. The main advantages and drawbacks of the different techniques were explored. Results obtained by the reviewed studies were reported in terms of classification performance and biomarker outcomes, in order to shed light on the parameters that accompany normal and pathological aging. Unresolved issues and possible future directions were finally pointed out.

Keywords: Alzheimer's Disease, Automatic classification, Automatic diagnosis, Machine learning, Magnetic Resonance Imaging, Mild Cognitive Impairment, Structural neuroimaging biomarkers, Support Vector Machine

1. INTRODUCTION

Alzheimer's Disease (AD) is the most frequent neurodegenerative disorder in the elderly population, putting forward early and accurate diagnosis as one of the main challenges in medicine to date. Therefore, research increased the focus on earlier stages of disease, and it has become clear the importance of biomarkers of AD since they precede cognitive and behavioral symptoms by years [1]. These evidences have prompted a revision of the established research diagnostic criteria for AD dementia that had served so well since 1984 [2]. These criteria were mainly focused on typical form of AD that arises with an insidious onset of episodic memory impairment bound to the selective atrophy of limbic regions, progressively involving other neocortical regions over time causing multiple cognitive impairments [3-6]

New revised criteria [7] considered not only the typical AD phenotype but also a more complete pattern of cognitive aging, and the inclusion of intermediate stage of mild cognitive impairment (MCI) that precedes dementia [8]. Other atypical phenotypes are considered, such as the visual variant of AD (corresponding to posterior cortical atrophy) or the language variant (corresponding to logopenic progressive aphasia). These clinical profiles are also consistent with MCI due to AD [8] [6], although they may evolve in different form of dementia. Furthermore, a prodromal, even earlier, period of preclinical AD could be identified by the presence of biomarkers detected in vivo in asymptomatic subjects, years before the onset of cognitive decline [9-11].

Among the biomarkers for the early diagnosis of AD at the prodromal disease stage, revised research diagnostic criteria propose biomarkers from in vivo neuroimaging techniques, including Magnetic Resonance Imaging (MRI) biomarkers,

as additional supportive features to clinical data [12] [11]. Structural MRI measurements in patients with AD showed brain atrophy over regions including entorhinal cortex, hippocampus, lateral and inferior temporal structures, anterior and posterior cingulate [13].

Methods to measure MRI-related biomarkers consist into neuroimage processing methods and include a variety of algorithms. Until the last few years, methods used to analyze neuroimaging data were mainly based on Region-of-interest measures, image segmentation and mass univariate statistical approaches, e.g. Voxel-Based Morphometry [14]. Recently, advances in statistical learning have led to a growing interest in multivariate pattern analysis (MVPA). This approach aims at the prediction of a single variable of interest (e.g. AD vs. controls) from the analysis and comparison of distributed patterns of activity.

The main advantage of MVPA-based approaches is that, given their multivariate nature, they are able to detect spatially distributed activations and cerebral patterns over a set of pixels, which results in a relatively higher sensitivity than conventional univariate analysis [15][16][22].

Among MVPA approaches, Support Vector Machines (SVMs) [17] [18] showed promising results in the automatic classification of medical images, e.g. [19], mainly for the classification of non-pathological vs. pathological images or for the classification of images belonging to different subtypes of disease.

When applied to AD patients using structural MRI images, SVM showed high potential in the classification of AD vs. normal subjects as well as in the prediction of conversion to AD in patients with MCI (e.g. [20]). These results make SVM comparable to a well-trained neuroradiologist in classify typical Alzheimer's disease [20] and a potential diagnostic tool in a clinical setting to support the early and differential diagnosis of AD [21] [22].

However, several approaches have been proposed for the automatic early and differential classification of patients with AD-related pathologies using structural MRI and SVM. The main differences among these approaches involve: 1) pre-processing methods of MRI images (e.g., different kernel type or different kernel size for smoothing); 2) feature extraction techniques (e.g. voxel-level approaches, anatomical measurements, data-driven techniques); 3) feature selection techniques (e.g. *a priori* Regions Of Interest (ROIs), univariate feature selection, *a posteriori* approaches and other data-driven techniques); 4) kernels for the SVM classification; 5) validation procedures; 6) the possibility to map the importance of each MRI image voxel for the classification, i.e., to identify cerebral patterns of differences between the two considered subject groups [23].

In this review, we focused on the potential role of MRI in the classification AD-related pathologies by means of SVM. A total of 30 studies emerged from an exhaustive search of PubMed, Web of Science and Medline records, published from 2008 to 2014. This paper aims to provide a systematic overview about the SVM approaches in the automatic classification of AD and in the prediction of conversion from MCI to AD. Both main achievements in terms of classification performance (e.g., accuracy, specificity and sensitivity) and limitations are described, including: 1) the effects of pre-processing on classification performances; 2) the effects of feature extraction and selection methods, 3) the effects of classification and validation procedures, 4) the interpretation of maps showing the importance of each MRI image voxel for the classification.

2. PRE-PROCESSING

Classification of medical images always requires pre-processing procedures, which are needed for image standardization purposes. For example, co-registration of images and spatial normalization to a standard anatomical space are useful since they allow voxel-based comparison among two or more images, and voxel-based spatial localization in a standard reference atlas, respectively.

However, there are some other pre-processing steps which could lead to an improvement in image classification performances. Among these, smoothing is a widely used strategy in medical image analysis that allows inter-subject differences and noisy on images to be reduced. This process consists in averaging the signal intensity of each voxel with the intensities of its neighbors, often using a filter (e.g. Gaussian). Smoothing aids normalizing the error distribution, and this is a mandatory condition for image analysis by inferences based on parametric tests. Smoothing is able to improve the signal-to-noise ratio and to account for local transformation errors which may occur during spatial co-registration [24].

Among the studies based on SVM and considered in this review, 10 applied smoothing (i.e. [6][24-33]). Among these, Vemuri et al. [25] applied an 8 mm FWHM Gaussian smoothing followed by an 8 mm isotropic down-sampling step that was performed by simple averaging. Characteristics of smoothing processing employed in the considered studies are summarized in Table 1 for AD vs. CN and in Table 2 for other comparisons.

3. FEATURE EXTRACTION

The main advantage of multivariate pattern analysis with respect to conventional univariate analysis is that they are able to classify samples by working in multidimensional spaces. This results in a relatively higher sensitivity, because a larger amount of information can be considered during the construction of the classifier. However, MVPA-based algorithms must be able to manage high data dimension.

Only a small number of SVM-based studies for the classification of MRI brain images directly makes use of information at a voxel level, e.g. pixel intensities from the whole brain or probability values from segmented tissues. The great majority of these studies involve a feature extraction step, which is aimed at reducing the dimension of the data to be handled without losing relevant information.

Feature extraction is the operation of generating a new set of features as a function of the original input data. The new set of features should have the following characteristics with respect to the original input data: 1) reduced redundancy, by removing those features carrying no more information than the selected subset; 2) increased relevance (or reduced irrelevance), by removing those features that do not provide useful information for class discrimination (independently

from the selected subset); 3) reduced dimensions, as a consequence of points 1) and 2).

This operation results in two main advantages. First, the reduction of the number of features to be handled by the classification algorithm usually enhances classification performance; this is only possible if the feature extraction technique is able to retain significant information for discrimination while discarding irrelevant and redundant information. Second, the reduction of the dimension of the feature set may result in a decrease of computational costs for the learning process.

In this section, we first describe those approaches that directly make use of information at a voxel level, such as voxel intensities from the whole brain, or voxel probability values from segmented tissues. Then, a survey of the most common feature extraction techniques is included: features as measured by anatomical structures on images, data-driven feature extraction techniques, hybrid feature extraction methods, with a brief description of those techniques that integrate features extracted from structural MR images with other information (e.g. demographic and clinical data). Feature extraction techniques employed by each considered study are summarized in Tables 1-2 for AD vs. CN and for other comparisons, respectively.

3.1. Voxel-level approaches: whole brain and tissue segmentation

In this category, we can include those approaches that directly use information at a voxel level for the classification of MRI brain images.

Methods that use voxel intensities from the whole brain for the classification of MRI images in AD are not widespread in literature. The main advantage of these methods is that no information is excluded *a priori* from the classification process. On the other side, this involves the use of redundant or irrelevant information for classification, which may lead to a decrease of the discrimination accuracy.

Methods that use probability values from segmented tissues are widely used.. These methods involve a processing of images , which aims at segmenting whole brain MR images into tissue probability maps, such as Gray Matter (GM), White Matter (WM), Cerebrospinal Fluid (CSF) or Ventricles (VN). Segmentation into tissue probability maps is made using fully-automated algorithms, such as Statistical Parametric Mapping (SPM—Wellcome Trust Centre for Neuroimaging, Institute of Neurology, UCL, London UK—http://www.fil.ion.ucl.ac.uk/spm) [34-36]. Most studies focused on classification of AD use GM map because it has been proven to be strongly involved in the pathophysiological mechanisms of early and late AD [37] [38]. Indeed, recent studies on AD classification have shown good discrimination performances when using GM (e.g. [39]). Among the considered studies on AD classification by means of SVM and MRI, Kloppel et al. [23], Magnin et al. [40], Abdulkadir et al. [41] and Retico et al. [33] used GM. Plant et al. [27], Dukart et al. [30] and Hidalgo-Munoz et al. [42] used both GM and WM; Fan et al. [24], Vemuri et al. [25] and Davatzikos et al. [24] used GM, WM and CSF; Varol et al. [43] used GM, WM and VN. In order to reduce the number of voxels within segmented maps, other studies included in this review used the above mentioned tissue segmentation methods in combination with feature extraction techniques [28] [31][44-48]. In the following sections a brief description of these feature extraction methods is presented.

3.2. Anatomical measurements

Feature extraction by anatomical measurements on structural MRIs is a widely used approach in the analysis and classification of AD (e.g. [21]). Among all possible anatomical measurements, volume and thickness of several cerebral regions have been shown to be reduced in AD patients as a consequence of brain atrophy. Oliveira et al. [49] measured the volume of cortical and subcortical structures as features for the classification of AD. Raamana et al. [32] used cortical thickness features. Schmitter et al. [50] used volume measurements of predetermined regions (i.e., left and right hippocampi, left and right temporal GM areas, lateral 3 and 4 ventricles, total GM and total CSF) as features for the early and differential diagnosis of AD.

Also cortical curvature represents a measure of interest, because its reduction has been observed as a possible effect of increased sulcal widening [51]. Diciotti et al. [52] measured the cortical thickness and mean curvature of predefined cortical structures together with the volume of predefined cortical and subcortical structures. In their work, each volume measurement was further normalized to the total intracranial volume in order to remove possible intra-subject differences due to dissimilarities in overall head size [53].

On the basis of previous studies about AD-related pathologies (e.g. [54-60]), Costafreda et al. [61] restricted their analysis by extracting shape features of the hippocampal region. Cui et al. [62] measured volume, thickness, curvature, surface area and folding of different regions (i.e., subcortical structures, brain stem, cerebellum and cerebral cortex) for the early classification of AD.

Other studies used anatomical measurements in combination with other feature extraction techniques[29] [31] [44] [46-48]. Among these, Ferrarini et al. [44] applied a local analysis of shape differences between AD and CN populations as feature extraction technique after tissue segmentation. Yang et al. [47] extracted shape features from structural MRIs in order capture the variation of anatomical shape that cannot be described using only volumetric measurements. Ota et al. [31] used mean GM density calculated over predetermined ROIs. In the study by Farhan et al. [48], the volume of GM, WM and CSF together with the size of hippocampi were extracted from MR images and used as features for classification.

3.3. Data-driven feature extraction techniques

This category includes those techniques that perform feature extraction starting from all available data and work after data analytical or geometrical transformation. This approach is particularly useful to identify patterns in data, as it allows to highlight similarities and differences even in high-dimensional sets. However, data-driven approaches may lead to the generation of a predictive model tailored for a particular dataset and lacking of generalization ability.

Among the published works considered in this review, Duchesne et al. [63], Wilson et al. [28] and Yang et al. [47] used Principal Components Analysis (PCA). PCA is a quite common feature extraction technique in neuroimaging studies (e.g. [64-66]) that is based on an orthogonal

transformation of a given dataset from a set of (possibly) correlated variables to a set of orthogonal (uncorrelated) variables, called principal components of the original dataset. Principal components are the eigenvectors of the covariance matrix of the dataset. The main feature of principal components is that they are sorted in descending order according to the proportion of variance explained, with the constraint for them to be orthogonal with each other. The maximum number of principal components which can be extracted from a dataset is always equal to the number of original variables and the sub-space defined by the principal components is called PCA subspace. Final features are extracted by projecting the original dataset onto the PCA subspace and they are called PCA coefficients of the original dataset. PCA is useful to reduce the dimension of the features to be handled without losing relevant information.

Another data-driven feature extraction approach is Independent Component Analysis (ICA) that aims at separating data into maximally independent groups by expressing a set of random variables as a linear combination of non-normal and statistically independent components. In this way, a multivariate signal (observation) can be separated into additive subcomponents (source) characterized by minimum mutual information [67]. In particular, spatial-ICA (sICA) aims at achieving maximal independence in space. sICA was used by Yang et al. [45] as feature extraction technique, as it satisfies the assumption that each structural MR image can be expressed as linear combination of spatially and statistically independent images.

Other studies used different data-driven feature extraction techniques for the classification of AD patients. Gerardin et al. [68] computed spherical harmonics coefficients to model the shape of the hippocampi and then used these coefficients as features for the classification process. Wolz et al. [29] used tensor-based morphometry [69] [70] and manifold-based learning [71] coupled with anatomical measurements to extract the features used to perform classification.

3.4. Hybrid feature extraction methods

Feature extraction approaches described above are not exclusive with one another. Several studies perform feature extraction by means of more than one method in succession, in order to take advantage of the strengths of each method. Among the considered studies, 8 used more than one feature extraction approach: Ferrarini et al [44], Wilson et al. [28], Wolz et al. [29], Yang et al. [45], Zhang et al. [46], Yang et al. [47], Farhan et al. [48], Ota et al. [31].

It is worth noting that many studies used features extracted from structural MR images in combination with data obtained by means of other modalities, e.g. demographic, clinical and Positron Emission Tomography (PET) data or neuropsychological and functional measures. For example, Nho et al. [72] integrated features obtained by anatomical measurements on predetermined ROIs with clinical features for the classification of patients with AD and MCI. Cui et al. [73] combined MRI features with CSF biomarkers and neuropsychological measurements. However, in this review we have not included this category of multi-modality approaches.

4. FEATURE SELECTION

Feature selection has the same purposes of feature extraction, such as dimensionality reduction and removal of redundant and irrelevant data. For this reason, feature extraction and feature selection are often confused. The main difference between feature extraction and selection is that the first generates a new set of features as a function of the original input data, while the second selects a subset of features from the original input data. As a consequence of this, feature selection does not suffer from problems of interpreting results, which is typical of some feature extraction techniques. On the contrary, feature selection can even aid the interpretation of predictive models, because it restricts the dataset to a subset of discriminative features without operating any transformation on them.

Besides reducing computational costs and improving model interpretability, selection of a subset of relevant features from the original dataset before the generation of the predictive model also helps reducing overfitting problems, which brings to a better evaluation of the generalization ability of the classifier.

In this section, we give a description of the most common feature selection methods, by listing them into five distinct categories: 1) *a priori* Regions Of Interest (ROIs); 2) univariate feature selection; 3) *a posteriori* feature selection; 4) other data-driven feature selection techniques; 5) hybrid feature selection methods. We highlight that feature selection can be considered an optional step when input data contain redundant or irrelevant information for the learning process, while feature extraction is a mandatory step for pattern recognition algorithms to classify high dimensional data. Among the studies considered in this paper, seven do not apply feature selection. In addition to these studies, Kloppel et al. [23] assessed the performances of a classification method with and without a feature selection process. For each of the considered studies, Tables 1-2 summarize the adopted feature selection techniques for AD vs. CN and for other comparisons, respectively.

4.1. A priori Regions-Of-Interest

Feature selection can use *a priori* knowledge of some information when involves local restriction to anatomical regions which are known to carry discriminative information. In this sense, *a priori* ROIs-based approaches have the advantage that selected features are relevant for classification. Because of this restriction to predetermined regions, ROI selection can even aid the interpretation of the results obtained by the classification.

Discriminative ROIs can be chosen on the basis of the results of previous studies. In the case of AD, relevant ROIs can be selected according to previous findings from neuroimaging studies, e.g. regions of brain atrophy from MRI, and regions with reduced glucose absorption from *Fluorodeoxyglucose* (^{18}FDG)–PET, respectively.

The main drawback of this approach is that *a priori* anatomical constraints may also exclude discriminative information. This could bring to the generation of an incomplete predictive model.

On the basis of previous literature (e.g. [74]), some studies restricted their analysis to the hippocampal region. Costafreda et al. [61] focused on hippocampi, thus obtaining information about the discriminative power of hippocampal subregions for the diagnosis of AD. In the study by Farhan et

al. [48], size information was extracted over the hippocampal region (together with anatomical measurements of GM, WM and CSF) and subsequently used as feature for classification. Gerardin et al. [68] segmented both the hippocampus and the amygdala ROIs using a previously implemented fully-automated method [75] [76]; this method was based on prior knowledge on the location of these two structures.

Duchesne et al. [63] choose the medial temporal lobe area as *a priori* ROI for the classification of AD vs. CN. Kloppel et al. [23] compared the performance of SVM classification obtained using grey matter from the whole brain with the one obtained using grey matter of antero-medial lobe volume of interest.

In the study by Chincarini et al. [77], ROI extraction was performed by means of rigid registration and matching to a Voxel Of Interest (VOI) template [78]. Selection was based on previous MRI and ^{18}FDG-PET findings relevant to AD-related pathology [79-82]. *A priori* selected ROIs included temporal lobe structures affected in early AD (hippocampus, enthorinal cortex, amygdale, middle and inferior temporal gyrus, insula, superior temporal gyrus), but also a control region which is almost unaffected in early AD (rolandic). It is interesting to note that the authors suggested the application of their algorithm to other ROIs not considered in their work, in order to explore and detect brain regions, which could be affected by early atrophic changes (e.g. thalamus or parietal cortex).

Diciotti et al. [52] constrained their analysis to *a priori* chosen cortical (e.g. temporo-parietal lobes) and subcortical (e.g. hippocampus and amygdala) regions. Selected areas were determined on the basis of well-known structural biomarkers of early AD [60]. In this way, twentyfive cortical and two subcortical ROIs were selected per cerebral hemisphere. Also Cui et al. [73] restricted their analysis to a predetermined number of cortical and subcortical structures that were automatically parcelled by processing each MRI scan with a free-available automatic software tool (FreeSurfer software package, http://surfer.nmr.mgh.harvard.edu/). In the paper by Cui et al. [62], feature extraction was performed over subcortical structures, brain stem, cerebellum and cerebral cortex.

Dukart et al. [30] performed *a priori* ROI selection on the basis of the results of a coordinate-based voxel-wise meta-analysis involving 1351 AD patients and 1097 CN subjects [83]. The aim of this meta-analysis was to identify the prototypical network of AD by seeking MRI-related and ^{18}FDG-PET-related biomarkers. ROIs selected by Dukart et al. represented the maxima of atrophy or reductions in glucose metabolism in AD, and included posterior insula, medial thalamus, hippocampal body/tail, middle temporal gyrus, superior temporal sulcus, amygdala, anterior hippocampal formation, uncus and (trans-) enthorinal area.

Ota et al. [31] extracted the mean GM density of predetermined ROIs for the early diagnosis of AD. These ROIs were chosen on the basis of three different brain atlases, with the aim of comparing the corresponding performances of classification.

Schmitter et al. [50] used total GM and total CSF information as well as features extracted from *a priori* chosen ROIs, i.e. left and right hippocampi, left and right temporal GM areas and lateral 3 and 4 ventricles.

On the other hand, some studies use an exclusive ROI approach, i.e., they select a ROI to be excluded from the classification process. Vemuri et al. [25] removed the cerebellum from the input dataset while retaining all other structures for subsequent analyses. An automatic parcellation process was performed by Magnin et al. [40]; in this case, the ROI of the cerebellum was excluded because it was partially cut off in some of the MR images. Also Zhang et al. [46] removed the cerebellum, and then obtained 93 ROI regions by applying spatial registration to a manually labelled template [84].

4.2. Univariate feature selection

A practical way to perform feature selection on structural MR images is applying univariate statistics to the features in order to rank them from the most to the least discriminative ones on a specific classification problem. In this way, a threshold can be applied to remove the less discriminative features but retaining the most relevant ones on the basis of statistical results.

This approach can be applied to the features extracted by means of one of the methods described in section 3, but it can be also applied directly to image voxels that can be ranked one by one and then selected. An obvious downside of this approach is that the power of multivariate tools such as SVM is likely to be reduced by the application of an univariate feature selection method before the classification process. Moreover, this method may cause lack of generalization because features are selected on a particular dataset.

Two studies among the ones considered in this paper performed feature selection by means of univariate statistics. Geradin et al. [68] computed Student's t-tests to determine which features conveyed relevant information for discrimination of AD and MCI. Their approach involved the use of a bagging strategy [85] [86]. Features with the highest T-values were selected as features to be used for the SVM classification process.

In the study by Varol et al. [43], Welch's t-test was used in order to directly rank image voxels according to their discriminative power with respect to class labels. Voxels ranked by their t-score were used to construct nested feature sets, i.e. an ensemble of sets where each set is contained in the subsequent one. In this way, the first set only contains voxels with the highest ranking, while the last one contains all voxels. These nested feature sets were then used to construct an SVM ensemble model.

4.3. A posteriori feature selection

This category includes all those approaches that perform feature selection *a posteriori* with respect to classification. In this case, features are selected according to their discriminative power on the basis of the results of classification.

A typical algorithm is the following: 1) features are divided into multiple subsets; 2) classification is repeated iteratively by varying the subset of the features; 3) for each iteration (i.e., for each subset of features used), a given parameter (e.g. accuracy of classification) is calculated, in order to quantify the goodness of classification; 4) the subset of most discriminative features is the one for which the calculated parameter is maximized.

This technique of feature selection performs selection on the basis of a particular dataset, which may lead to lack of generalization ability. Being *a posteriori* process, this method is part of the evaluating process of the classification performance. Thus, the choice of the evaluating process (see Section 6) is of great importance to avoid over-fitting, e.g. when the same dataset is used, in a *a posteriori* feature selection process, for both training and testing the classifier

In Fan et al. [24] and Davatzikos et al. [24], the classification error rate was used as parameter to perform *a posteriori* feature selection. Oliveira et al. [49] performed feature selection using the F-score as parameter to quantify the goodness of a subset of features. Cui et al. [73] applied *a posteriori* selection of features after ranking them using a data-driven approach. The optimal feature subset was then identified by maximizing the Area Under the ROC Curve (AUC) for SVM classification. Also in the second considered study by Cui et al. [62], an AUC-based *a posteriori* feature selection was used.

In the study by Vemuri et al. [25], the weight assigned by SVM during classification is used as parameter to perform feature selection. In this way, a weight-based feature ranking was obtained [87], reflecting their importance for classification. By applying a threshold, features with low magnitude weights were discarded from the dataset while features with high magnitude weights were retained. As the authors wrote, this method is particularly interesting when applied directly to voxel-level data, because it is able to consider at the same time all the voxel locations and to rank them according to their importance for classification alone, i.e. independently from their position. In this way, voxels affected by the disease can be identified even if they are located in independent anatomic regions.

Also Hidalgo-Munoz et al. [42] used the weight assigned by SVM to determine the relative importance of each feature. In their work, they employed a Recursive Feature Elimination (RFE) strategy [88], that is a technique in which features are iteratively removed in order to find the optimal subset of data for classification. Specifically, they divided the feature set into different subsets and then applied the RFE strategy to each subset. For each subset, features were ordered according to the weight vector returned by SVM, and the less relevant features were discarded. The use of different subsets helps avoiding biased results and increasing the generalization ability of the predictive model. Also Retico et al. [33] adopted a RFE approach by ranking features according to the weight returned by SVM. In Ota et al. [31], RFE was performed by ranking features *via* an accuracy-based criterion.

4.4. Other data-driven feature selection techniques

In this category, we include all other approaches that perform feature selection starting from a particular dataset and operating data transformation,

Methods used for selecting feature by data-driven approaches are very similar to data-driven feature extraction techniques described in section 3. In some cases, methods used for feature extraction can be useful as feature selection techniques as well. For example, by using PCA, extracted PCA coefficients are ranked according to the percentage of explained variance, and this criterion can be used in order to select the most discriminative features.

As for feature extraction approaches, feature selection by data-driven techniques allows similarities and differences in high-dimensional datasets to be identified, but it may bring to the generation of models with poor generalization ability.

Fan et al. [24] and Davatzikos et al. [24] performed feature selection by computing a measure of voxel discriminative power (using the Pearson correlation coefficient) and spatial consistency.

Plant et al. [27] selected the most relevant features by rating their interestingness for class separation using the Information Gain approach [89] [90]. After that, they also applied a clustering technique [91] in order to identify sets of contiguous voxels with high discriminatory power and to remove noisy information.

In the work by Cui et al. [73], feature ranking was performed by means of the minimum redundancy and maximum relevance (mRMR) method [92] [93]. This ranking step was followed by an *a posteriori* feature selection strategy.

Chincarini et al. [77] used Random Forest (RF) classifier [94] to compute the weight (i.e. relative importance) of each feature for RF classification. In this way, they obtained an Important Features Map (IFM) identifying relevant voxels for the classification of AD and MCI. The application of a threshold to the IFM thus allowed removing irrelevant features.

In the second considered work by Cui et al. [62], feature selection included a step based on Fisher scores, that were calculated for each feature (higher Fisher scores correspond to higher discriminative power between different groups). The optimal subset of features was chosen according to the scores calculated for each feature.

4.5. Hybrid feature selection methods

As for feature extraction, also feature selection can be performed using more than one method in succession, with the aim of exploiting the selective power of each approach. Among the considered studies, eight studies used more than one feature selection approach, i.e, Fan et al. [24], Vemuri et al. [25], Davatzikos et al. [24], Gerardin et al. [68], Chincarini et al. [77], Cui et al. [73] [62], Ota et al. [31].

5. CLASSIFICATION BY SVM

SVM algorithm was introduced by Vapkin and colleagues [4]. SVM is a binary classifier which aims at generating a predictive model for the discrimination of new samples.

Let us suppose to have a set of training data consisting of a vector of N samples belonging to two classes and the corresponding vector of class labels (e.g. -1 and +1 for control and patient class, respectively). SVM generates a hyper-plane able to discriminate between the two classes of training dataset. The main advantage of this classification approach is that SVM minimizes the empirical classification error and maximizes the separation distance between the two training classes, that is why the decision function is often called maximum margin hyper-plane. This hyper-plane is described by the following function:

$$y(x) = \sum_{n=1}^{N} w_n \cdot t_n \cdot k(x, x_n) + b$$

where N is the number of samples in the training dataset; w is a weight assigned by SVM during the training phase to each training sample, reflecting its importance for classes discrimination; t is the class label of the training sample; k(x,xn) is a kernel function; b is a bias parameter. Samples of the two classes which lie on the margin of the hyper-plane are called support vector, and they are the only samples for which the assigned weight w is non-zero. The function above is the predictive model generated by SVM, and it returns the class label y for an unseen sample x.

In the modified version of SVM published by Cortes and Vapkin [95], the idea of soft margin was introduced, which is useful when training classes cannot be sharply discriminated. In this case, the soft margin approach allows to misclassify a fraction of training samples, while preserving the ability of the hyper-plane to maximizing its distance from the nearest samples of the two classes.

The main parameter to be set in a SVM classifier is the kernel function, which is mostly set as linear, polynomial or Gaussian Radial Basis Function (RBF). Linear kernels are defined as

$$k(x_i, x_j) = (x_i \cdot x_j)$$

while (homogeneous and inhomogeneous) polynomial kernels are given by

$$k(x_i, x_j) = (x_i \cdot x_j)^d$$

$$k(x_i, x_j) = (x_i \cdot x_j + 1)^d$$

and Gaussian RBF kernels have the following form

$$k(x_i, x_j) = \exp\left(-\gamma \left\| x_i - x_j \right\|^2\right) \qquad \gamma > 0$$

As pointed out by Orrù et al. [96], non-linear kernels are thought to be more flexible than linear kernels in solving difficult discrimination problems, which reflects in better classification performance. This flexibility can lead, on the other side, to over-fitting problems, as non-linear kernels may not result in a higher generalization power when classifying unseen samples. This drawback is highlighted when the sample size is small. Thus, in neuroimaging classification problems, linear kernels are usually preferred. This is also due to the consideration that, when the sample size is small with respect to the number of features involved, a linear decision function separating the data can easily be found [97] [98].

Other parameters to be optimized are those related to the choice of the kernel, e.g. γ when using Gaussian RBF. Moreover, another parameter to be optimized is the soft margin regularization constant C [95].

Both kernel function and related parameters can be set to default values before the classification process. Another option is to perform parameter optimization, which can be accomplished using a grid search approach. In this approach, a subset of the parameter space is defined, and classification performances are evaluated for each parameter configuration. The configuration corresponding to the best classification performance is chosen as the optimal one. It is important to note that data used for parameter optimization should not be the same used for training and testing the classifier, because this could lead to a biased estimate of the generalization error.

Table 1 and Table 2 summarize the kernel used the considered papers, for AD vs. CN and for other comparisons, respectively. Among the considered studies, 13 used a linear kernel and 13 used RBF. When multiple kernels were used in order to compare the results of classification, only the kernel function corresponding to the best performance was reported.

6. VALIDATION AND COMPARISON OF DIFFERENT TECHNIQUES

One of the most important issues about the generation of a predictive model is its validation, because a good validation process allows to correctly quantify the discriminatory power of a given model, which also gives the possibility to compare classification techniques based on different approaches. For example, if parameter selection, training of the predictive model and validation are performed using the same dataset, the generated classifier will probably show limited generalization ability when classifying unseen samples. In this sense, a minimally biased estimate of the true diagnostic performance of the classifier is needed.

In this section, Cross Validation (CV) techniques are described in contrast with train-and-test (T&T) approaches. A brief description of the most popular metrics in literature for the quantification of the discriminatory power of a model is given.

6.1. Cross validation

This approach is the most popular among validation techniques because it allows to quantify the discriminatory power of a predictive model even if the size of the dataset is small. The use of CV has been suggested for the assessment of the generalization ability of a predictive model in machine learning methods, because it is able to decrease image variability problems [42] [99] [100]. Moreover, the probability of the test error of a classifier estimated using this approach is almost unbiased [101].

CV involves partitioning the original dataset into complementary subsets, the training set and the testing set. The training set is used to train the classifier, while the testing set is used to validate the generated predictive model. By using different partitions of the original dataset, multiple rounds of CV can be performed, which can aid reducing the variability of the partitioned subsets. Results obtained from multiple rounds can then simply be averaged in order to obtain a quantification of the performance of the classifier.

One type of CV approach is the so called k-fold CV, in which the original dataset is randomly partitioned into k subsets of equal size. The training of the classifier is performed using k-1 subsets, while the testing is performed using the remaining subset. The procedure is then repeated k times, until all subsets are used once as testing set. The advantage of this particular type of CV is that each sample of the original dataset is used once for validation, with all samples being used for both training and testing phases.

Leave-One-Out (LOO) CV can be considered a particular form of k-fold CV in which k equals the number of samples in the original dataset. In LOO, the training of the classifier is performed using n-1 samples of the original dataset; while the testing is performed using the remaining samples (n being the total number of samples in the original dataset). The procedure is then repeated n times, until all samples are used once for validation.

Among the studies considered in this paper, 9 used CV ([28], [77], [61], [46], [62], [52], [48], [42], [33]) and 15 used LOO ([63], [24], [44], [23], [68], [49], [27], [41], [29], [6], [30], [47], [31], [50]).

Magnin et al. [40] used a particular bootstrapping approach in which train and test data were randomly drawn without replacement from the original dataset, and a correct classification rate was calculated from classification results. This procedure was repeated 5000 times, so that each sample was classified using different combinations of training data. A mean correct classification rate was calculated at the end of the whole procedure.

6.2. Train-and-test

This kind of procedure is used when the number of samples in the original dataset is high enough to allow its splitting into two subsets including different samples, which can be used to train and test the classifier. Specifically, this approach involves partitioning the original dataset into 2 complementary subsets, the training set and the testing set. The first set is used to train the classifier, the second is used for validation. When using this approach, over-training problems are reduced because the training and the testing sets are completely independent. Anyway, results could be related to the particular choice of the partition into training and testing subsets.

In the case of AD studies, an interesting approach is the one used by Plant et al. [27] and by Cui et al. [73], who generated a predictive model for the classification of MCI-converter (MCIc) vs. MCI-non converter (MCInc) by training their classifier on independent sets of AD patients and controls.

Among the studies considered in this review 13 performed training and testing of the classifier using two independent subsets ([44], [23], [25], [68], [27], [41], [61], [73], [45], [43], [30], [32], [33]).

6.3. Metrics

Among all the metrics used in classification studies to quantify the diagnostic performance of a classifier, the most popular are accuracy, sensitivity, specificity and AUC. These statistical measures will be described in the case of binary classifiers.

Accuracy of classification is a simple measure of the number of correctly classified samples (for both classes) divided by the total number of classified samples. If the error rate is defined as the number of misclassified samples (both classes) divided by the total number of classified samples, it is evident that accuracy and error rate are complementary measures. Accuracy is the most used metric in classification problems.

Two metrics of great importance in medicine are sensitivity and specificity, as they measure the rate of correctly classified samples in the positive and negative class, respectively (e.g. AD class and Controls class). Sensitivity (also known as True Positive Rate or Recall) is given by the number of correctly classified samples belonging to the positive class (true positives) divided by the total number of samples belonging to the positive class (true positives plus false negatives). Specificity (also known as True Negative Rate) is given by the number of correctly classified samples belonging to the negative class (true negatives) divided by the total number of samples belonging to the negative class (true negatives plus false positives). Here, true positive (negative) gives the number of correctly classified samples belonging to the positive (negative) class, while false positive (negative) gives the number of misclassified samples belonging to the negative (positive) class.

Another important metric in classification problems is given by the study of the Receiver Operating Characteristic (ROC) curve [102] [103]. For a binary classifier, A ROC curve is a plot of the true positive rate (sensitivity) against the false positive rate (1 – specificity), which can be obtained at different setting thresholds. AUC gives a quantification of the classifier performance, with a higher statistical consistency than accuracy [104].

Table 1 and Table 2 show classification results of the considered methods in terms of accuracy, sensitivity, specificity and AUC (when available) for AD vs. CN and for other comparisons, respectively.

Figure 1 shows accuracy as a function of the considered study for AD vs. CN classification. A distinction between results obtained using CV or T&T strategy is made according to the legend.

Figure 2 shows specificity and sensitivity rates as a function of the considered study for AD vs. CN classification. A distinction between results obtained using CV or T&T strategy is made according to the legend.

Figure 3 shows accuracy, specificity and sensitivity rates as a function of the considered study for MCIc vs. MCInc classification. A distinction between results obtained using CV or T&T strategy is made according to the legend.

In Figure 4, a boxplot shows the results of AD vs. CN classification in terms of accuracy for the studies considered in this review as a function of pre-processing: with and without smoothing. Figure 5 shows a boxplot reporting the results of AD vs. CN classification in terms of accuracy as a function of the applied SVM kernel function: linear and Gaussian RBF function. In Figure 6, the results in terms of accuracy of CV and T&T classification strategies for the considered studies are reported using a boxplot. The following comparisons were considered: AD vs. CN (left) and MCIc vs. MCInc (right).

7. EXTRACTION OF BIOMARKERS

Once the predictive model is generated and validated, it can be used as a tool to determine which of the input features were the most important for class discrimination. In the case of MR images of AD patients, this results in the extraction of sMRI-related biomarkers for the (early and/or differential) diagnosis of AD.

Extraction of biomarkers from neuroimages is an important characteristic because it also serves as an additional qualitative validation of the generated predictive model if the findings mirror the expected pattern distribution of anatomical changes in AD or if they return biologically sensible results.

Methods to extract biomarkers by means of SVMs can substantially be divided into two categories based on two different approaches: 1) extraction of those biomarkers maximizing feature classification performance, and 2) weigh maps. In the following sections, we considered only studies using these two approaches and reported the resulting findings.

7.1. Extraction by feature classification performance

This technique performs feature extraction according to the performance of the predictive model by varying the subset of features used for classification. This approach is strictly related to *a posteriori* feature selection described in Section 4. In several studies (e.g. [49]) *a posteriori* feature selection and biomarker extraction overlap, as the feature selected according to the classification performance are also presented as possible biomarkers. In this case, results of feature selection and biomarker extraction are the same.

The typical algorithm of extraction by feature classification performance follows the same steps used for *a posteriori* feature selection: 1) features are divided into multiple subsets; 2) classification is repeated iteratively by varying the subset of used features; 3) for each iteration, the subset used for classification is changed and a given parameter is measured, in order to quantify the goodness of classification; 4) features for which classification performance was maximized are extracted as possible biomarkers.

Among the studies described in this paper, Oliveira et al. [49], Cui et al. [73] [62] and Chincarini et al. [77] used AUC as measure of the classification performance. Diciotti et al. [52] used accuracy, sensitivity and specificity rates. Farhan et al. [48] used accuracy rates of classification. Ota et al. [31] performed an accuracy-based RFE to select the most important features for the classification of MCIc vs. MCInc. These features were highlighted as possible biomarkers for early and differential diagnosis of AD.

SVM, as a classification method, has proven to be effective for the identification of functional and structural parameters that can be indicators of biological process underlying the pathology of AD [96] [105]. In particular, neuroimaging MRI values are able to identify atrophied area specifics for AD progression.

Different comparisons arose from the literature and here are taken into consideration: AD vs CN, MCI vs. NC, MCIc vs. MCInc, AD vs. MCI.

When studies compared AD vs. CN, features influencing the classification involved:

1. hippocampi [40] [49] [77], area of left hippocampus [48]
2. parahippocampal gyrus [40]
3. amygdala [77]
4. anterior and posterior corpus callosum, right lateral ventricular horn [49]
5. middle inferior temporal gyrus [77]

When studies compared MCI vs. CN or MCIc vs. CN, features influencing the classification involved:

1. hippocampus, amigdala, left middle inferior temporal gyrus [77]
2. right straight gyrus, right supramarginal gyrus, right short insular gyri, left orbital gyri, left angular gyrus [62]

When studies compared MCIc and MCInc, features influencing the classification involved:

1. left hippocampus [77] [31], right and left hippocampi [73],
2. left middle inferior temporal gyrus [77], right and left middle temporal gyrus [73]
3. amygdala, insula [77]
4. left enthorinal cortex, right inferior parietal cortex, left retrosplenial cortex [73]
5. left inferior occipital gyrus, left parahippocammpal gyrus, right middle frontal gyrus, right middle occipital gyrus, left superior occipital gyrus, right supramarginal gyrus, left angular gyrus, left precentral gyrus, right caudate [31]

Only one study [52] performed extraction of biomarkers for the comparisons between mild AD and CN and between mild AD and MCI. Features influencing the classification for both comparisons involved cortical and subcortical volumes.

7.2. Weight maps

A different approach with respect to extraction by feature classification performance can be followed when using SVM classifiers (e.g. [23] [66]). During the training phase, SVM classifiers assign a specific weight to each sample of the input dataset, according to its importance for the generation of the predictive model and, hence, for class discrimination. By multiplying each sample for its weight and by summing them, a map of the most discriminative features can be obtained. When using medical images as samples and image voxels as features of each sample, by multiplying each image for its weight and by summing them on a voxel basis results in a map reflecting the relevance of each voxel for classification. By overlapping the resulting map onto a standard anatomical template, possible biomarkers can be identified. It is worth noting that the weight calculated by SVM is non-zero only for support vectors, while it has positive or negative sign depending on the class to which the sample belongs.

Although the potential of this technique, the difficulty in the interpretation of SVM weight maps still remains an open issue in neuroimaging research [22]. Several studies tried to apply arbitrary thresholds or statistical tests to the weight maps, these approaches improving model interpretability. However, as stressed by Schrouff et al. [22], the debate about this point is still open, because, given the multivariate nature of SVMs, spatial inference on weight maps by means

of univariate statistics should be avoided. In particular, the SVM model is constructed from the combination of all weights, which are not independent one another as in statistical parametric maps.

Kloppel et al. [23] used this technique to extract possible biomarkers not only for the AD vs. CN but also for the AD vs. FTD comparisons by using whole brain GM images. In this case, classification depended on voxels in frontal as well as parietal areas. Davatzikos et al. [24] extracted biomarkers comparing MCI vs. CN by using a particular approach based on the gradient of SVM decision function. GM clusters influencing the classification involved the lateral and inferior parts of hippocampi, the bilateral superior, the middle and inferior temporal gyri, the bilateral orbitofrontal, left fusiform gyrus, right collateral sulcus, posterior cingulate. They also found voxels of reduced WM volumes in the inferior temporal gyri, middle and superior frontal gyri.

Costafreda et al. [61] performed extraction of biomarkers by means of a method based on the distance of samples from the SVM separating hyper-plane for the classification of AD. More recently, Retico et al. [33] used a weight map approach for biomarkers extraction in AD vs. CN.

Hidalgo-Munoz et al. [42] used the weight assigned by SVM to determine the relative importance of each feature by means of the RFE strategy (Section 4) in combination with a feature classification performance approach using accuracy as measure of the goodness of classification.

Among the studies considered in this review, features influencing the classification between AD and CN involved:

1. parahippocampal gyrus [23] [25] [33] [42]
2. middle temporal gyrus [25] [33]
3. hippocampi [42], bilateral atrophy in lateral and medial aspects of hippocampal head, lesser extent in hippocampal body [61]
4. amygdala [42] [33]
5. insula [25] [42]
6. parietal cortex [23], parietal lobe [25]
7. lenticular nucleus [42], lentiform nucleus (putamen) [33]
8. temporal-parietal association cortex, posterior cingulate/precuneus, temporal lobe [25]
9. entorhinal cortex, fusiform gyrus [42]
10. uncus, inferior frontal gyrus, declive, middle frontal gyrus, superior temporal gyrus, [33]

When studies compared MCI and CN, features influencing the classification involved:

1. lateral and inferior parts of hippocampi, bilateral superior, middle and inferior temporal gyri, bilateral orbito-frontal, left fusiform gyrus, right collateral sulcus, posterior cingulated, middle and superior frontal gyri [24]

Only one study [23] performed extraction of biomarkers for the comparisons between AD and FTD. Features influencing the classification involved frontal and parietal areas.

In Table 3, employed technique for the extraction of biomarkers and corresponding findings in terms of anatomical structures for each of the considered studies and for AD vs. CN comparison are shown. Method and results of biomarkers extraction in these studies for other comparisons, i.e. MCI vs. CN, MCIc vs. CN, MCIc vs. MCInc, mild AD vs. CN, mild AD vs. MCI and AD vs. FTD are shown in Table 4.

In Fig. 7, results of biomarker extraction for the considered studies are visually summarized for the following comparisons: AD vs. CN (triangles), MCI vs. CN (circles) and MCIc vs. MCInc (squares). It's worth noting that we only specified the name of a structure if the number of studies reporting its finding was greater than one (i.e., at least two studies). Marker size is proportional to the number of studies reporting each structure (from 1 to 6). Lateral (top) and medial (bottom) views are shown. Reported structures are not intended to be localized in the left or right hemisphere.

8. DISCUSSION

Over the past decade, there has been a huge interest for the implementation of supervised whole brain automatic classification methods in order to detect and classify, through a computer algorithm, specific patterns related to AD corresponding to a sensitive biomarkers (or a combination of them).

In this paper, we considered 30 studies that used SVM for the classification of AD-related pathologies by means of MRI data. A trend of the 30 publications on this topic (SVM and AD) found from 2008 to date and included in this review is shown in Figure 8, including studies investigating or not on potential MRI-related biomarkers of AD.

As a general results, the SVM method , allowed discriminating between AD and CN, MCI and CN, AD and FTD. However, SVM allowed also predicting conversion from MCI to AD. Only very few studies tried to discriminate between different forms of early-onset of dementia: Wilson et al. [28] considered the utility of SVM and MR images to discriminate three linguistic variants (progressive non-fluent aphasia, semantic dementia and logopenic progressive aphasia) of Primary Progressive Aphasia (PPA). Ridgway et al., [6] investigated how to differentiate between posterior cortical atrophy, typical amnestic AD and logopenic progressive aphasia from MRI studies by SVM. Gerardin et al. [68] and Raamana et al. [32] focused on the SVM classification of single- (amnestic) and multi-domain MCI. These findings could represent a valid support for clinicians given the difficulty to distinguish among these early variants in the clinical practice.

The most popular techniques of pre-processing, feature extraction and feature selection were described, together with different validation and biomarker extraction strategies. The main advantages and limitations were pointed out during the review, and the results obtained by the 30 considered papers were shown in Tables 1-4 and Fig. 1-3.

The effects of smoothing on classification performances were briefly described in section 2. Results obtained by this review process allow us to give a more quantitative description of the effects related to the use of pre-processing steps. As shown in Fig. 4 for the comparison between AD and CN, the difference between studies which make use or not use of smoothing is not significant in terms of accuracy. . Nevertheless, the median value of accuracy for studies which make use of smoothing was found to be slightly higher than in those which do not make use of smoothing (0.95 and 0.88, respectively).

Unlike most SVM parameters that can be optimized iteratively, the choice of a kernel function for SVM classification is usually made *a priori*. Among the considered studies, 13 used a linear function and 13 a Gaussian RBF kernel function (4 studies did not provide kernel information)Linear kernels are could reduce overfitting problems, while RBF should improve model flexibility, this resulting in higher classification accuracy [96]., No significant difference between linear and RBF kernels was found in terms of accuracy of classification of AD vs. CN (Fig. 5) However, the median accuracy using a linear kernel function was found to be slightly lower than the median accuracy for RBF kernel (0.88 and 0.91, respectively).

Another important aspect to be discussed is the use of different validation strategies, i.e., CV or T&T. As described in section 6, the probability of the test error of a classifier estimated using CV is almost unbiased [4] [101]. On the other side, in T&T training and testing sets are completely independent, and this helps reducing over-fitting problems. As shown in Fig. 6, for both AD vs. CN and MCIc vs. MCInc comparisons, there are no significant differences between accuracy obtained *via* both CV and T&T strategies. However, it is interesting to note that median accuracy obtained using CV were always higher than median accuracy obtained using T&T. Specifically, for the comparison between AD and CN, median accuracy were 0.90 and 0.87 when using CV and T&T, respectively; for the comparison between MCIc and MCInc, median accuracy were 0.78 and 0.62 when using CV and T&T, respectively. Moreover, as shown in Figures 1-3 and Tables 1-2, , it can be seen that accuracy obtained *via* T&T were always lower than or (at most) equal to accuracy obtained *via* CV (i.e., studies # 4, 8, 12 and 23 in Figures 1-2 and study # 10 in Fig. 3).

The use of SVM analysis based on structural images also offers the possibility to objectively identify MRI-related biomarkers through a group comparison. Therefore, in this systematic review, we also briefly reported 14 studies that applied SVM method in order to detect the more important patterns of difference for AD classification and that could lead to the identification of possible MRI-related biomarkers.

Why it should be important to detect biomarkers in an early disease phase?

As already commented in the introduction, there is growing a widespread consensus about the fact that brain changes associated with AD could be distinguished years before the onset of the clinical manifestation [106] [107] [9] [74] [11], leaving cognitive symptoms as the last stage of the pathological process [108]. In such scenario, MRI could provide crucial information about the status of disease bringing out the grade of volume atrophy. MRI biomarkers have been shown to be sensitive to the diagnosis of AD, for instance morphological abnormalities in the medial temporal lobes (e.g. the entorhinal cortex), the posterior cingulate gyrus, and the temporal–parietal associative areas have been demonstrated to be useful markers in a clinical setting [109] [110]. However, these MRI-based measures derive from a manual quantification performed by trained neuroradiologists, such visual inspection being susceptible of bias since the human eye is not able to detect little but fundamental changes in the cerebral volume.

Some studies tried to compare manual performance of classification with automatic methods highlighting SVM as comparable to a well-trained neuroradiologist in classifying typical AD patients, and even more sensible then a non-trained one (e.g. [23]).

Tables 3-4 and figure 7 show areas identified by the authors considered in our review. What clearly emerges is the lack across the studies of a specific area for detecting AD, although hippocampus and parahippocampal gyrus are the most consistent findings among the studies comparing AD vs. CN [40] [49] [77] [42] [48] [23] [25] [33]. This result is consistent with those findings present in literature that indicate hippocampus as one of the most affected structures in AD-related pathologies (for a review see Jack et al. [74]) associated to learning and memory (in particular short term topographical memory [111]), that are often involved in typical-AD onset. Specifically, recent studies focus their attention on hippocampal subregions, stating that the whole structure may not be uniformly affected in AD (e.g. [112]). Costafreda et al. [61] investigated the role of different subregions of the hippocampi, finding that bilateral atrophy affects both lateral and medial aspects of hippocampal head and, in minor part, hippocampal body.

Other structures emerging among some studies, when taking into account the comparison between AD and CN, are amygdala, insula and lenticular nucelus [77] [42] [25] [33]. Amygdala atrophy is usually related to the earliest clinical stages of AD [113], with potential relationships to anxiety and irritability, and with a mutual dependence with the hippocampus in encoding memories with emotional connotation [114]. Moreover, AD patients are often affected by autonomic instability, reduced behavioral control (judgments regarding inner well-being) and visceral dysfunction. All these functions are also regulated by insular cortex [115]. The findings of hippocampal and parahippocampal structures occur for AD vs. CN, MCI vs. CN [24] [77] and MCIc vs. MCInc [77] [73] [31].

As the structures above mentioned, middle-temporal gyrus is reported as biomarker for different comparison (AD vs. CN, MCI vs. CN, MCIc vs. MCInc), but it is not possible to consider it as univocal biomarker for one of these comparisons. On the other hand, it is important to note that, in literature, middle-temporal lobe atrophy was considered for a long time as a biomarker able to predict dementia in patients with MCI [116].

Parietal and frontal neocortices are usually affected later by AD, and are associated with other cognitive functions, as well as language, praxic, visuospatial and behavioral impairments [13]. Among the studies considered in our review, when considering AD vs. CN, parietal cortex was found by Kloppel et al. [23] and Vemuri et al. [25], while frontal areas were found by Retico et al. [33]. However,

parietal and frontal areas were found as possible biomarkers also for the comparisons between AD and FTD [23], MCI and CN [24], MCIc and MCInc [73] [31].

Surprisingly, very few studies (i.e., [25] [42] [24] [73] found the structures of the limbic lobe (such as entorhinal cortex and posterior cingulate gyrus), for AD vs. CN, MCI vs. CN and MCIc vs. MCInc) and these structures have been already suggested as diagnostic markers for early AD [109]. From a pathophysiological point of view, the entorhinal cortex is progressively interested by neurofibrillary pathology and cell loss in early phase of AD, disconnecting the hippocampus from neocortical regions [3] [117-120].

Overall, some structures arose more frequently than others do. Notably, none of the above mentioned anatomical regions is specific for just one of the comparisons; on the other side, results are not completely consistent among all the studies and a pool of areas are indicated by single studies. This could be a consequence of the low sensitiveness of the algorithms to detect the most involved areas in AD. On the other hand, the overlapping of some reported regions through different comparisons (see Figure 7) may suggest that early detection of AD is a matter of sensitivity and that MVPA approaches are potentially able to capture sensible features supporting automatic and objective diagnosis.

Finally, we have identified some crucial points that are considerable as possible directions for future studies.

SVM approach is highly affected from:

1) the lack of standardization. This involves MRI data acquisition procedures, image pre-processing, feature extraction and selection, classification and validation approaches, approaches for extraction of biomarkers;

2) the inhomogeneity of the studied samples. This inovolves for example, sample size, demographic and clinical features. In this sense, the study by Cuingnet et al. [21] represents an interesting example of comparison of different SVM-based approaches for the classification of AD on the same population;

3) the exclusion of predetermined ROIs on the basis of previous literature, which could exclude important areas involved in the pathophysiological process of AD. For example, it is interesting to note that, although some studies [25] [46] excluded *a priori* the cerebellum, Retico et al. [33] found part of the cerebellum as important region for the discrimination of AD vs. CN;

4) the way of clinical diagnosis as gold standard for the classification performance. It should note that in all considered studies, the definition of the label for each class was based on a probable diagnosis (not from a post-mortem confirmation). This could introduce a limitation on the potential power of classifier.

9. CONCLUSIONS

In conclusion, a state-of-the-art overview about SVM for the diagnosis of AD by means of MRI studies was provided in this systematic review. The focus of our analysis was the early and differential diagnosis of AD-related pathologies. Since there is a strong need of AD diagnostic methods and biomarkers in clinical practice, in order to properly manage the patients and address them to the optimal care and treatment. We have described the main advantages and drawbacks of some SVM approaches and mapped the SVM-derived biomarkers of early AD, as pointed out from 30 studies published on this matter.

Our findings lead to shed light on parameters that accompany normal and pathological aging.

However, although SVM could represent a breakthrough in the way to perform AD diagnosis, there are still some issues that need to be clarified and future directions that could be investigated before the adoption of SVM-based AD diagnosis in the clinical practice.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

REFERENCES

[1] Weintraub S, Wicklund AH, Salmon DP. The neuropsychological profile of Alzheimer disease. Cold Spring Harbor perspectives in medicine 2(4): a006171.(2012).

[2] McKhann G, Drachman D, Folstein M, Katzman R, Price D, Stadlan EM. Clinical diagnosis of Alzheimer's disease Report of the NINCDS-ADRDA Work Group* under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease. Neurology 34(7): 939-39.(1984).

[3] Braak H, Braak E. Neuropathological stageing of Alzheimer-related changes. Acta neuropathologica 82(4): 239-59.(1991).

[4] Vapnik VN. Statistical learning theory. Wiley New York (1998).

[5] Jack C, Petersen R, Xu Y, O'brien P, Smith G, Ivnik R, *et al.* Rates of hippocampal atrophy correlate with change in clinical status in aging and AD. Neurology 55(4): 484-90.(2000).

[6] Ridgway GR, Lehmann M, Barnes J, Rohrer JD, Warren JD, Crutch SJ, *et al.* Early-onset Alzheimer disease clinical variants Multivariate analyses of cortical thickness. Neurology 79(1): 80-84.(2012).

[7] McKhann GM, Knopman DS, Chertkow H, Hyman BT, Jack Jr CR, Kawas CH, *et al.* The diagnosis of dementia due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimer's & Dementia 7(3): 263-69.(2011).

[8] Albert MS, DeKosky ST, Dickson D, Dubois B, Feldman HH, Fox NC, *et al.* The diagnosis of mild cognitive impairment due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimer's & Dementia 7(3): 270-79.(2011).

[9] Perrin RJ, Fagan AM, Holtzman DM. Multimodal techniques for diagnosis and prognosis of Alzheimer's disease. Nature 461(7266): 916-22.(2009).

[10] Jack Jr CR, Knopman DS, Jagust WJ, Shaw LM, Aisen PS, Weiner MW, *et al.* Hypothetical model of dynamic

biomarkers of the Alzheimer's pathological cascade. The Lancet Neurology 9(1): 119-28.(2010).
[11] Sperling RA, Aisen PS, Beckett LA, Bennett DA, Craft S, Fagan AM, *et al.* Toward defining the preclinical stages of Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimer's & Dementia 7(3): 280-92.(2011).
[12] Dubois B, Feldman HH, Jacova C, DeKosky ST, Barberger-Gateau P, Cummings J, *et al.* Research criteria for the diagnosis of Alzheimer's disease: revising the NINCDS–ADRDA criteria. The Lancet Neurology 6(8): 734-46.(2007).
[13] McDonald C, McEvoy L, Gharapetian L, Fennema-Notestine C, Hagler D, Holland D, *et al.* Regional rates of neocortical atrophy from normal aging to early Alzheimer disease. Neurology 73(6): 457-65.(2009).
[14] Ashburner J, Friston KJ. Voxel-based morphometry—the methods. Neuroimage 11(6): 805-21.(2000).
[15] Mahmoudi A, Takerkart S, Regragui F, Boussaoud D, Brovelli A. Multivoxel Pattern Analysis for fMRI Data: A Review. Computational and mathematical methods in medicine 20122012).
[16] Pereira F, Mitchell T, Botvinick M. Machine learning classifiers and fMRI: a tutorial overview. Neuroimage 45(1): S199-S209.(2009).
[17] Vapnik V. The nature of statistical learning theory. springer (2000).
[18] Shawe-Taylor J, Cristianini N. Kernel methods for pattern analysis. Cambridge university press (2004).
[19] Focke NK, Helms G, Scheewe S, Pantel PM, Bachmann CG, Dechent P, *et al.* Individual voxel-based subtype prediction can differentiate progressive supranuclear palsy from idiopathic parkinson syndrome and healthy controls. Human brain mapping 32(11): 1905-15.(2011).
[20] Klöppel S, Stonnington CM, Barnes J, Chen F, Chu C, Good CD, *et al.* Accuracy of dementia diagnosis—a direct comparison between radiologists and a computerized method. Brain 131(11): 2969-74.(2008).
[21] Cuingnet R, Gerardin E, Tessieras J, Auzias G, Lehéricy S, Habert M-O, *et al.* Automatic classification of patients with Alzheimer's disease from structural MRI: a comparison of ten methods using the ADNI database. Neuroimage 56(2): 766-81.(2011).
[22] Schrouff J, Rosa MJ, Rondina JM, Marquand AF, Chu C, Ashburner J, *et al.* PRoNTo: pattern recognition for neuroimaging toolbox. Neuroinformatics 11(3): 319-37.(2013).
[23] Klöppel S, Stonnington CM, Chu C, Draganski B, Scahill RI, Rohrer JD, *et al.* Automatic classification of MR scans in Alzheimer's disease. Brain 131(3): 681-89.(2008).
[24] Fan Y, Batmanghelich N, Clark CM, Davatzikos C. Spatial patterns of brain atrophy in MCI patients, identified via high-dimensional pattern classification, predict subsequent cognitive decline. Neuroimage 39(4): 1731-43.(2008).
[25] Vemuri P, Gunter JL, Senjem ML, Whitwell JL, Kantarci K, Knopman DS, *et al.* Alzheimer's disease diagnosis in individual subjects using structural MR images: validation studies. Neuroimage 39(3): 1186-97.(2008).
[26] Davatzikos C, Fan Y, Wu X, Shen D, Resnick SM. Detection of prodromal Alzheimer's disease via pattern classification of magnetic resonance imaging. Neurobiol Aging 29(4): 514-23.(2008).
[27] Plant C, Teipel SJ, Oswald A, Böhm C, Meindl T, Mourao-Miranda J, *et al.* Automated detection of brain atrophy patterns based on MRI for the prediction of Alzheimer's disease. Neuroimage 50(1): 162-74.(2010).
[28] Wilson SM, Ogar JM, Laluz V, Growdon M, Jang J, Glenn S, *et al.* Automated MRI-based classification of primary progressive aphasia variants. Neuroimage 47(4): 1558-67.(2009).
[29] Wolz R, Julkunen V, Koikkalainen J, Niskanen E, Zhang DP, Rueckert D, *et al.* Multi-method analysis of MRI images in early diagnostics of Alzheimer's disease. PloS one 6(10): e25446.(2011).
[30] Dukart J, Mueller K, Barthel H, Villringer A, Sabri O, Schroeter ML. Meta-analysis based SVM classification enables accurate detection of Alzheimer's disease across different clinical centers using FDG-PET and MRI. Psychiatry Research: Neuroimaging 212(3): 230-36.(2013).
[31] Ota K, Oishi N, Ito K, Fukuyama H. A comparison of three brain atlases for MCI prediction. Journal of neuroscience methods 221: 139-50.(2014).
[32] Raamana PR, Wen W, Kochan NA, Brodaty H, Sachdev PS, Wang L, *et al.* The sub-classification of amnestic mild cognitive impairment using MRI-based cortical thickness measures. Frontiers in Neurology 52014).
[33] Retico A, Bosco P, Cerello P, Fiorina E, Chincarini A, Fantacci ME. Predictive Models Based on Support Vector Machines: Whole-Brain versus Regional Analysis of Structural MRI in the Alzheimer's Disease. Journal of Neuroimaging2014).
[34] Ashburner J, Csernansk JG, Davatzikos C, Fox NC, Frisoni GB, Thompson PM. Computer-assisted imaging to assess brain structure in healthy and diseased brains. The Lancet Neurology 2(2): 79-88.(2003).
[35] Ashburner J, Friston KJ. Unified segmentation. Neuroimage 26(3): 839-51.(2005).
[36] Ashburner J, Barnes G, Chen C, Daunizeau J, Flandin G, Friston K, *et al.* SPM8 manual. Functional Imaging Laboratory, Institute of Neurology2012).
[37] Frisoni G, Testa C, Zorzan A, Sabattoli F, Beltramello A, Soininen H, *et al.* Detection of grey matter loss in mild Alzheimer's disease with voxel based morphometry. Journal of Neurology, Neurosurgery & Psychiatry 73(6): 657-64.(2002).
[38] Frisoni GB, Pievani M, Testa C, Sabattoli F, Bresciani L, Bonetti M, *et al.* The topography of grey matter involvement in early and late onset Alzheimer's disease. Brain 130(3): 720-30.(2007).
[39] Lilia M, Marie S, Valerie H-B, Bruno D, Patrick G, Serge K. DTI and Structural MRI Classification in Alzheimer's Disease. Advances in Molecular Imaging 20122012).
[40] Magnin B, Mesrob L, Kinkingnéhun S, Pélégrini-Issac M, Colliot O, Sarazin M, *et al.* Support vector machine-

based classification of Alzheimer's disease from whole-brain anatomical MRI. Neuroradiology 51(2): 73-83.(2009).
[41] Abdulkadir A, Mortamet B, Vemuri P, Jack Jr CR, Krueger G, Klöppel S. Effects of hardware heterogeneity on the performance of SVM Alzheimer's disease classifier. Neuroimage 58(3): 785-92.(2011).
[42] Hidalgo-Muñoz AR, Ramírez J, Górriz JM, Padilla P. Regions of interest computed by SVM wrapped method for Alzheimer's disease examination from segmented MRI. Frontiers in aging neuroscience 62014).
[43] Varol E, Gaonkar B, Erus G, Schultz R, Davatzikos C, Eds. Feature ranking based nested support vector machine ensemble for medical image classification. Proceedings of the Biomedical Imaging (ISBI), 2012 9th IEEE International Symposium on; 2012. IEEE (Year).
[44] Ferrarini L, Palm WM, Olofsen H, van der Landen R, van Buchem MA, Reiber JH, *et al.* Ventricular shape biomarkers for Alzheimer's disease in clinical MR images. Magnetic resonance in medicine 59(2): 260-67.(2008).
[45] Yang W, Lui RL, Gao J-H, Chan TF, Yau S-T, Sperling RA, *et al.* Independent component analysis-based classification of Alzheimer's disease MRI data. Journal of Alzheimer's Disease 24(4): 775-83.(2011).
[46] Zhang D, Wang Y, Zhou L, Yuan H, Shen D. Multimodal classification of Alzheimer's disease and mild cognitive impairment. Neuroimage 55(3): 856-67.(2011).
[47] Yang S-T, Lee J-D, Chang T-C, Huang C-H, Wang J-J, Hsu W-C, *et al.* Discrimination between Alzheimer's Disease and Mild Cognitive Impairment Using SOM and PSO-SVM. Computational and mathematical methods in medicine 20132013).
[48] Farhan S, Fahiem MA, Tauseef H. An Ensemble-of-Classifiers Based Approach for Early Diagnosis of Alzheimer's Disease: Classification Using Structural Features of Brain Images. Computational and mathematical methods in medicine 20142014).
[49] Oliveira Jr PPdM, Nitrini R, Busatto G, Buchpiguel C, Sato JR, Amaro Jr E. Use of SVM methods with surface-based cortical and volumetric subcortical measurements to detect Alzheimer's disease. Journal of Alzheimer's Disease 19(4): 1263-72.(2010).
[50] Schmitter D, Roche A, Maréchal B, Ribes D, Abdulkadir A, Bach-Cuadra M, *et al.* An evaluation of volume-based morphometry for prediction of mild cognitive impairment and Alzheimer's disease. NeuroImage: Clinical2014).
[51] Im K, Lee J-M, Won Seo S, Hyung Kim S, Kim SI, Na DL. Sulcal morphology changes and their relationship with cortical thickness and gyral white matter volume in mild cognitive impairment and Alzheimer's disease. Neuroimage 43(1): 103-13.(2008).
[52] Diciotti S, Ginestroni A, Bessi V, Giannelli M, Tessa C, Bracco L, *et al.*, Eds. Identification of mild Alzheimer's disease through automated classification of structural MRI features. Proceedings of the Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE; 2012. IEEE (Year).
[53] Desikan RS, Cabral HJ, Hess CP, Dillon WP, Glastonbury CM, Weiner MW, *et al.* Automated MRI measures identify individuals with mild cognitive impairment and Alzheimer's disease. Brain 132(8): 2048-57.(2009).
[54] Kantarci K, Weigand S, Przybelski S, Shiung M, Whitwell J, Negash S, *et al.* Risk of dementia in MCI Combined effect of cerebrovascular disease, volumetric MRI, and 1H MRS. Neurology 72(17): 1519-25.(2009).
[55] Risacher SL, Saykin AJ, West JD, Shen L, Firpi HA, McDonald BC, *et al.* Baseline MRI predictors of conversion from MCI to probable AD in the ADNI cohort. Current Alzheimer Research 6(4): 347.(2009).
[56] Csernansky J, Wang L, Swank J, Miller J, Gado M, McKeel D, *et al.* Preclinical detection of Alzheimer's disease: hippocampal shape and volume predict dementia onset in the elderly. Neuroimage 25(3): 783-92.(2005).
[57] Apostolova LG, Dutton RA, Dinov ID, Hayashi KM, Toga AW, Cummings JL, *et al.* Conversion of mild cognitive impairment to Alzheimer disease predicted by hippocampal atrophy maps. Archives of Neurology 63(5): 693-99.(2006).
[58] Morra JH, Tu Z, Apostolova LG, Green AE, Avedissian C, Madsen SK, *et al.* Automated 3D mapping of hippocampal atrophy and its clinical correlates in 400 subjects with Alzheimer's disease, mild cognitive impairment, and elderly controls. Human brain mapping 30(9): 2766-88.(2009).
[59] Ferrarini L, Frisoni GB, Pievani M, Reiber JH, Ganzola R, Milles J. Morphological hippocampal markers for automated detection of Alzheimer's disease and mild cognitive impairment converters in magnetic resonance images. Journal of Alzheimer's Disease 17(3): 643-59.(2009).
[60] Frisoni GB, Fox NC, Jack CR, Scheltens P, Thompson PM. The clinical use of structural MRI in Alzheimer disease. Nature Reviews Neurology 6(2): 67-77.(2010).
[61] Costafreda SG, Dinov ID, Tu Z, Shi Y, Liu C-Y, Kloszewska I, *et al.* Automated hippocampal shape analysis predicts the onset of dementia in mild cognitive impairment. Neuroimage 56(1): 212-19.(2011).
[62] Cui Y, Sachdev PS, Lipnicki DM, Jin JS, Luo S, Zhu W, *et al.* Predicting the development of mild cognitive impairment: a new use of pattern recognition. Neuroimage 60(2): 894-901.(2012).
[63] Duchesne S, Caroli A, Geroldi C, Barillot C, Frisoni GB, Collins DL. MRI-based automated computer classification of probable AD versus normal controls. Medical Imaging, IEEE Transactions on 27(4): 509-20.(2008).
[64] Habeck C, Foster NL, Perneczky R, Kurz A, Alexopoulos P, Koeppe RA, *et al.* Multivariate and univariate neuroimaging biomarkers of Alzheimer's disease. Neuroimage 40(4): 1503-15.(2008).
[65] López M, Ramírez J, Górriz JM, Álvarez I, Salas-Gonzalez D, Segovia F, *et al.* Principal component analysis-based techniques and supervised classification schemes for the early detection of Alzheimer's disease. Neurocomputing 74(8): 1260-71.(2011).
[66] Salvatore C, Cerasa A, Castiglioni I, Gallivanone F, Augimeri A, Lopez M, *et al.* Machine learning on brain MRI

data for differential diagnosis of Parkinson's disease and Progressive Supranuclear Palsy. Journal of neuroscience methods 222: 230-37.(2014).
[67] Calhoun V, Adali T, Pearlson G, Pekar J. Spatial and temporal independent component analysis of functional MRI data containing a pair of task-related waveforms. Human brain mapping 13(1): 43-53.(2001).
[68] Gerardin E, Chételat G, Chupin M, Cuingnet R, Desgranges B, Kim H-S, *et al.* Multidimensional classification of hippocampal shape features discriminates Alzheimer's disease and mild cognitive impairment from normal aging. Neuroimage 47(4): 1476-86.(2009).
[69] Brun CC, Leporé N, Pennec X, Lee AD, Barysheva M, Madsen SK, *et al.* Mapping the regional influence of genetics on brain structure variability—a tensor-based morphometry study. Neuroimage 48(1): 37-49.(2009).
[70] Koikkalainen J, Lötjönen J, Thurfjell L, Rueckert D, Waldemar G, Soininen H. Multi-template tensor-based morphometry: application to analysis of Alzheimer's disease. Neuroimage 56(3): 1134-44.(2011).
[71] Belkin M, Niyogi P. Laplacian eigenmaps for dimensionality reduction and data representation. Neural computation 15(6): 1373-96.(2003).
[72] Nho K, Shen L, Kim S, Risacher SL, West JD, Foroud T, *et al.*, Eds. Automatic Prediction of Conversion from Mild Cognitive Impairment to Probable Alzheimer's Disease using Structural Magnetic Resonance Imaging. Proceedings of the AMIA Annual Symposium Proceedings; 2010. American Medical Informatics Association (Year).
[73] Cui Y, Liu B, Luo S, Zhen X, Fan M, Liu T, *et al.* Identification of conversion from mild cognitive impairment to Alzheimer's disease using multivariate predictors. PloS one 6(7): e21896.(2011).
[74] Jack Jr CR, Bernstein MA, Borowski BJ, Gunter JL, Fox NC, Thompson PM, *et al.* Update on the magnetic resonance imaging core of the Alzheimer's disease neuroimaging initiative. Alzheimer's & Dementia 6(3): 212-20.(2010).
[75] Chupin M, Mukuna-Bantumbakulu AR, Hasboun D, Bardinet E, Baillet S, Kinkingnéhun S, *et al.* Anatomically constrained region deformation for the automated segmentation of the hippocampus and the amygdala: Method and validation on controls and patients with Alzheimer's disease. Neuroimage 34(3): 996-1019.(2007).
[76] Colliot O, Chételat G, Chupin M, Desgranges B, Magnin B, Benali H, *et al.* Discrimination between Alzheimer Disease, Mild Cognitive Impairment, and Normal Aging by Using Automated Segmentation of the Hippocampus 1. Radiology 248(1): 194-201.(2008).
[77] Chincarini A, Bosco P, Calvini P, Gemme G, Esposito M, Olivieri C, *et al.* Local MRI analysis approach in the diagnosis of early and prodromal Alzheimer's disease. Neuroimage 58(2): 469-80.(2011).
[78] Calvini P, Chincarini A, Gemme G, Penco MA, Squarcia S, Nobili F, *et al.* Automatic analysis of medial temporal lobe atrophy from structural MRIs for the early assessment of Alzheimer disease. Medical physics 36(8): 3737-47.(2009).
[79] Mosconi L, Brys M, Glodzik-Sobanska L, De Santi S, Rusinek H, de Leon MJ. Early detection of Alzheimer's disease using neuroimaging. Experimental gerontology 42(1): 129-38.(2007).
[80] Fennema-Notestine C, Hagler DJ, McEvoy LK, Fleisher AS, Wu EH, Karow DS, *et al.* Structural MRI biomarkers for preclinical and mild Alzheimer's disease. Human brain mapping 30(10): 3238-53.(2009).
[81] Walhovd K, Fjell A, Dale A, McEvoy L, Brewer J, Karow D, *et al.* Multi-modal imaging predicts memory performance in normal aging and cognitive decline. Neurobiology of aging 31(7): 1107-21.(2010).
[82] Liu Y, Paajanen T, Zhang Y, Westman E, Wahlund L-O, Simmons A, *et al.* Analysis of regional MRI volumes and thicknesses as predictors of conversion from mild cognitive impairment to Alzheimer's disease. Neurobiology of aging 31(8): 1375-85.(2010).
[83] Schroeter ML, Stein T, Maslowski N, Neumann J. Neural correlates of Alzheimer's disease and mild cognitive impairment: a systematic and quantitative meta-analysis involving 1351 patients. Neuroimage 47(4): 1196-206.(2009).
[84] MacDonald D, Kabani N, Avis D, Evans AC. Automated 3-D extraction of inner and outer surfaces of cerebral cortex from MRI. Neuroimage 12(3): 340-56.(2000).
[85] Breiman L. Bagging predictors. Machine learning 24(2): 123-40.(1996).
[86] Fan Y, Shen D, Gur RC, Gur RE, Davatzikos C. COMPARE: classification of morphological patterns using adaptive regional elements. Medical Imaging, IEEE Transactions on 26(1): 93-105.(2007).
[87] Mladenić D, Brank J, Grobelnik M, Milic-Frayling N, Eds. Feature selection using linear classifier weights: interaction with classification models. Proceedings of the Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval; 2004. ACM (Year).
[88] Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. Machine learning 46(1-3): 389-422.(2002).
[89] Quinlan JR. C4. 5: programs for machine learning. Morgan kaufmann (1993).
[90] Hall MA, Holmes G. Benchmarking attribute selection techniques for discrete class data mining. Knowledge and Data Engineering, IEEE Transactions on 15(6): 1437-47.(2003).
[91] Ester M, Kriegel H-P, Sander J, Xu X, Eds. A density-based algorithm for discovering clusters in large spatial databases with noise. Proceedings of the Kdd; 1996. (Year).
[92] Peng H, Long F, Ding C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. Pattern Analysis and Machine Intelligence, IEEE Transactions on 27(8): 1226-38.(2005).
[93] Ding C, Peng H. Minimum redundancy feature selection from microarray gene expression data. Journal of bioinformatics and computational biology 3(02): 185-205.(2005).

[94] Breiman L. Random forests. Machine learning 45(1): 5-32.(2001).
[95] Cortes C, Vapnik V. Support-vector networks. Machine learning 20(3): 273-97.(1995).
[96] Orrù G, Pettersson-Yeo W, Marquand AF, Sartori G, Mechelli A. Using support vector machine to identify imaging biomarkers of neurological and psychiatric disease: a critical review. Neuroscience & Biobehavioral Reviews 36(4): 1140-52.(2012).
[97] Muller K, Mika S, Ratsch G, Tsuda K, Scholkopf B. An introduction to kernel-based learning algorithms. Neural Networks, IEEE Transactions on 12(2): 181-201.(2001).
[98] Schölkopf B, Smola AJ. Learning with kernels: support vector machines, regularization, optimization, and beyond. MIT press (2002).
[99] Kohavi R, Ed. A study of cross-validation and bootstrap for accuracy estimation and model selection. Proceedings of the IJCAI; 1995. (Year).
[100] Ortiz A, Górriz JM, Ramírez J, Martínez-Murcia FJ. LVQ-SVM Based CAD tool applied to structural MRI for the diagnosis of the Alzheimer's disease. Pattern Recognition Letters 34(14): 1725-33.(2013).
[101] Chapelle O, Vapnik V, Eds. Model Selection for Support Vector Machines. Proceedings of the NIPS; 1999. (Year).
[102] Flach PA, Ed. The geometry of ROC space: understanding machine learning metrics through ROC isometrics. Proceedings of the ICML; 2003. (Year).
[103] Fawcett T. An introduction to ROC analysis. Pattern Recognition Letters 27(8): 861-74.(2006).
[104] Huang J, Ling CX. Using AUC and accuracy in evaluating learning algorithms. Knowledge and Data Engineering, IEEE Transactions on 17(3): 299-310.(2005).
[105] Gomez-Ramirez J, Wu J. Network-based biomarkers in Alzheimer's disease: review and future directions. Frontiers in aging neuroscience 62014).
[106] Selkoe DJ. Alzheimer's disease is a synaptic failure. Science 298(5594): 789-91.(2002).
[107] Coleman P, Federoff H, Kurlan R. A focus on the synapse for neuroprotection in Alzheimer disease and other dementias. Neurology 63(7): 1155-62.(2004).
[108] Kehoe EG, McNulty JP, Mullins PG, Bokde AL. Advances in MRI biomarkers for the diagnosis of Alzheimer's disease. Biomarkers in medicine 8(9): 1151-69.(2014).
[109] Jones BF, Barnes J, Uylings HB, Fox NC, Frost C, Witter MP, *et al.* Differential regional atrophy of the cingulate gyrus in Alzheimer disease: a volumetric MRI study. Cerebral Cortex 16(12): 1701-08.(2006).
[110] Choo I, Lee DY, Oh JS, Lee JS, Lee DS, Song IC, *et al.* Posterior cingulate cortex atrophy and regional cingulum disruption in mild cognitive impairment and Alzheimer's disease. Neurobiology of aging 31(5): 772-79.(2010).
[111] Hartley T, Bird CM, Chan D, Cipolotti L, Husain M, Vargha-Khadem F, *et al.* The hippocampus is required for short-term topographical memory in humans. Hippocampus 17(1): 34-48.(2007).
[112] Greene SJ, Killiany RJ. Hippocampal subregions are differentially affected in the progression to Alzheimer's disease. The Anatomical Record 295(1): 132-40.(2012).
[113] Poulin SP, Dautoff R, Morris JC, Barrett LF, Dickerson BC. Amygdala atrophy is prominent in early Alzheimer's disease and relates to symptom severity. Psychiatry Research: Neuroimaging 194(1): 7-13.(2011).
[114] Richardson MP, Strange BA, Dolan RJ. Encoding of emotional memories depends on amygdala and hippocampus and their interactions. Nature neuroscience 7(3): 278-85.(2004).
[115] Bonthius DJ, Solodkin A, Van Hoesen GW. Pathology of the insular cortex in Alzheimer disease depends on cortical architecture. Journal of Neuropathology & Experimental Neurology 64(10): 910-22.(2005).
[116] Korf ES, Wahlund L-O, Visser PJ, Scheltens P. Medial temporal lobe atrophy on MRI predicts dementia in patients with mild cognitive impairment. Neurology 63(1): 94-100.(2004).
[117] Gómez-Isla T, Price JL, McKeel Jr DW, Morris JC, Growdon JH, Hyman BT. Profound loss of layer II entorhinal cortex neurons occurs in very mild Alzheimer's disease. The Journal of neuroscience 16(14): 4491-500.(1996).
[118] Kordower JH, Chu Y, Stebbins GT, DeKosky ST, Cochran EJ, Bennett D, *et al.* Loss and atrophy of layer II entorhinal cortex neurons in elderly people with mild cognitive impairment. Annals of neurology 49(2): 202-13.(2001).
[119] Price JL, Davis P, Morris J, White D. The distribution of tangles, plaques and related immunohistochemical markers in healthy aging and Alzheimer's disease. Neurobiology of aging 12(4): 295-312.(1991).
[120] van Hoesen GW, Augustinack JC, Dierking J, Redman SJ, Thangavel R. The parahippocampal gyrus in Alzheimer's disease: clinical and preclinical neuroanatomical correlates. Annals of the New York Academy of Sciences 911(1): 254-74.(2000).

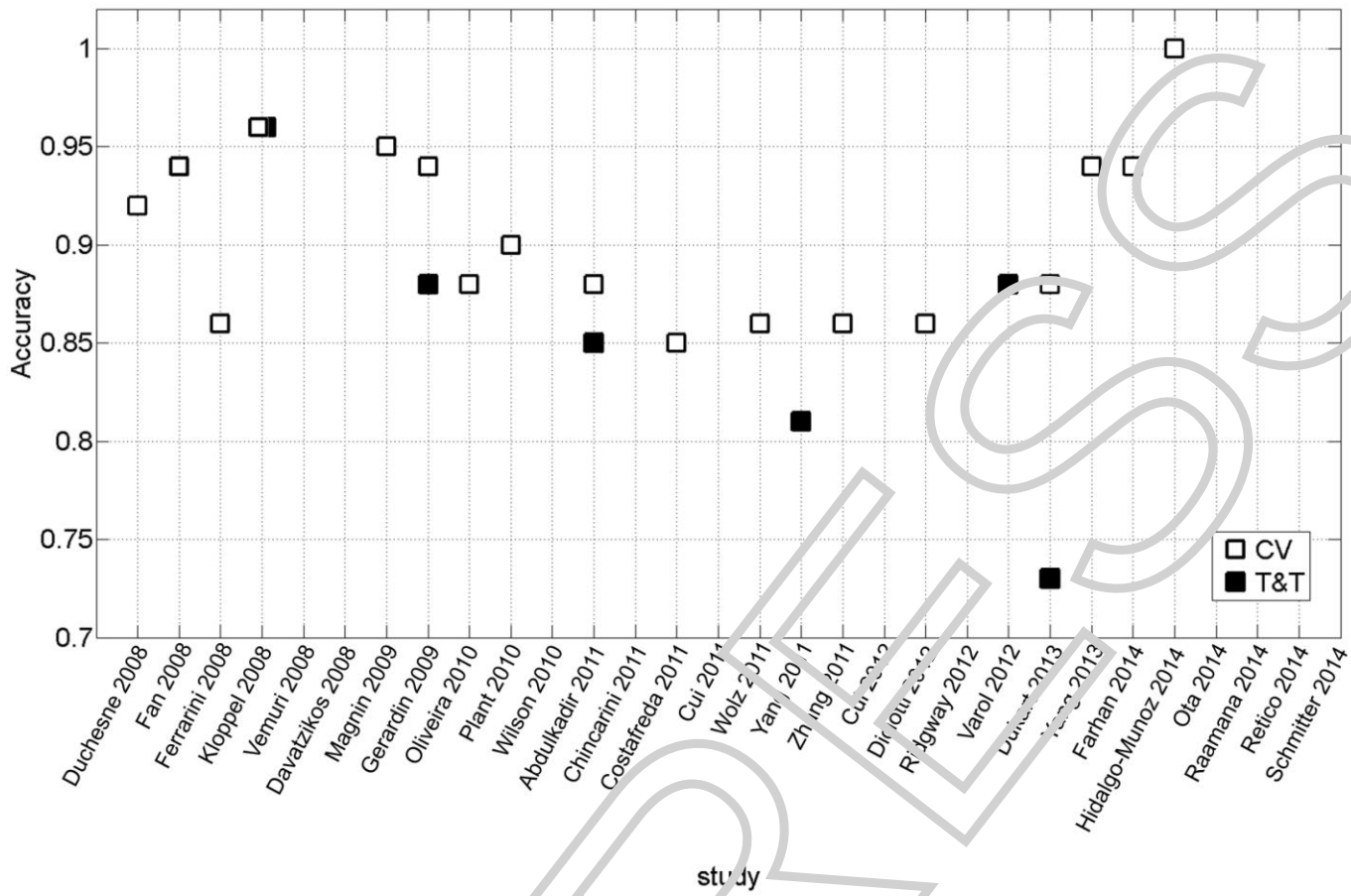


Figure 1 Accuracy rate as a function of the considered study for AD vs. CN classification. Validation strategies: CV (white squares), T&T (black squares).

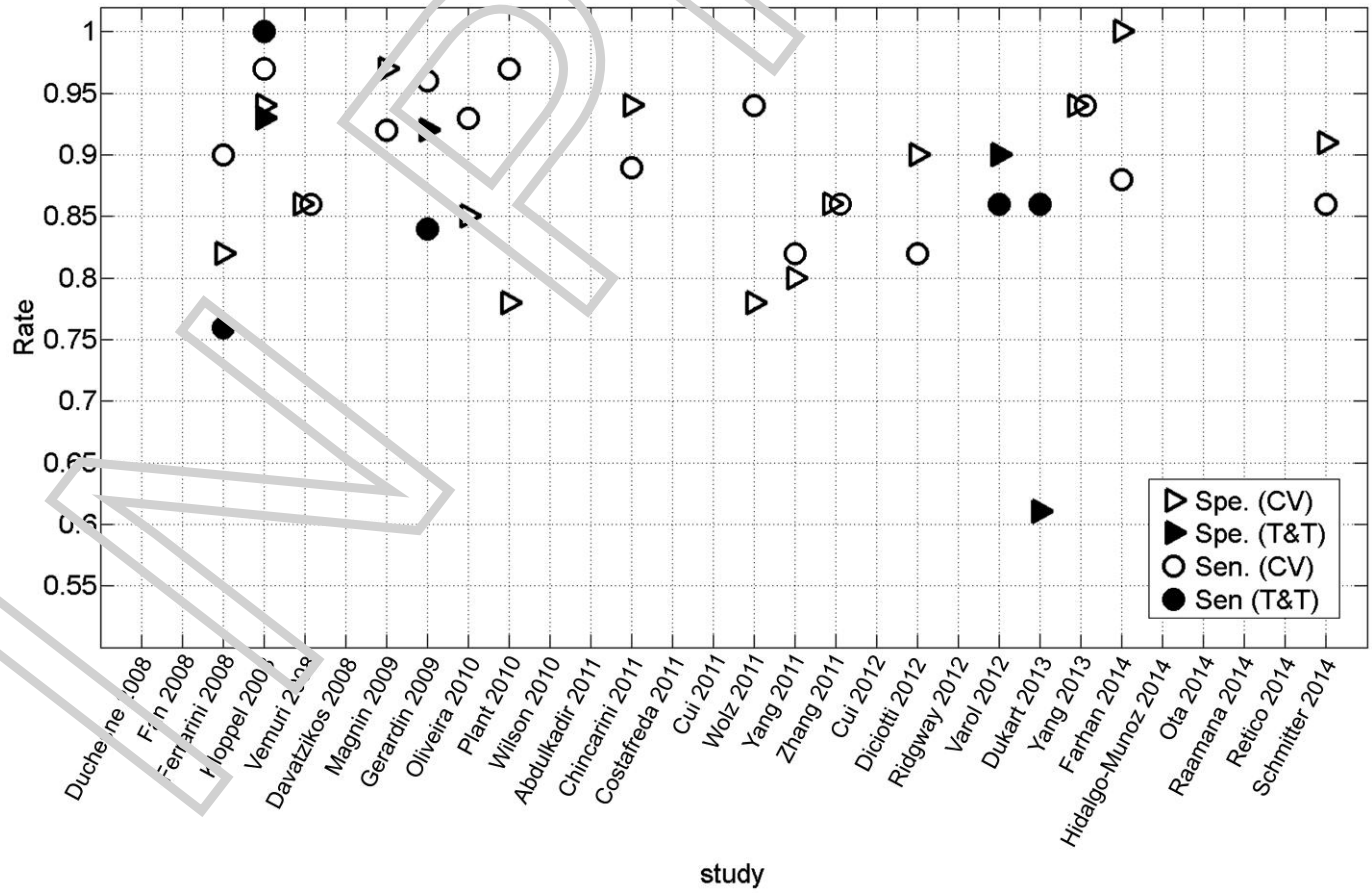


Figure 2 Specificity (Spe.) and sensitivity (Sen.) rates as a function of the considered study (numbered #) for AD vs. CN classification. Validation strategies: CV (squares), T&T (circles). Specificity and sensitivity rates are indicated in red and green, respectively.

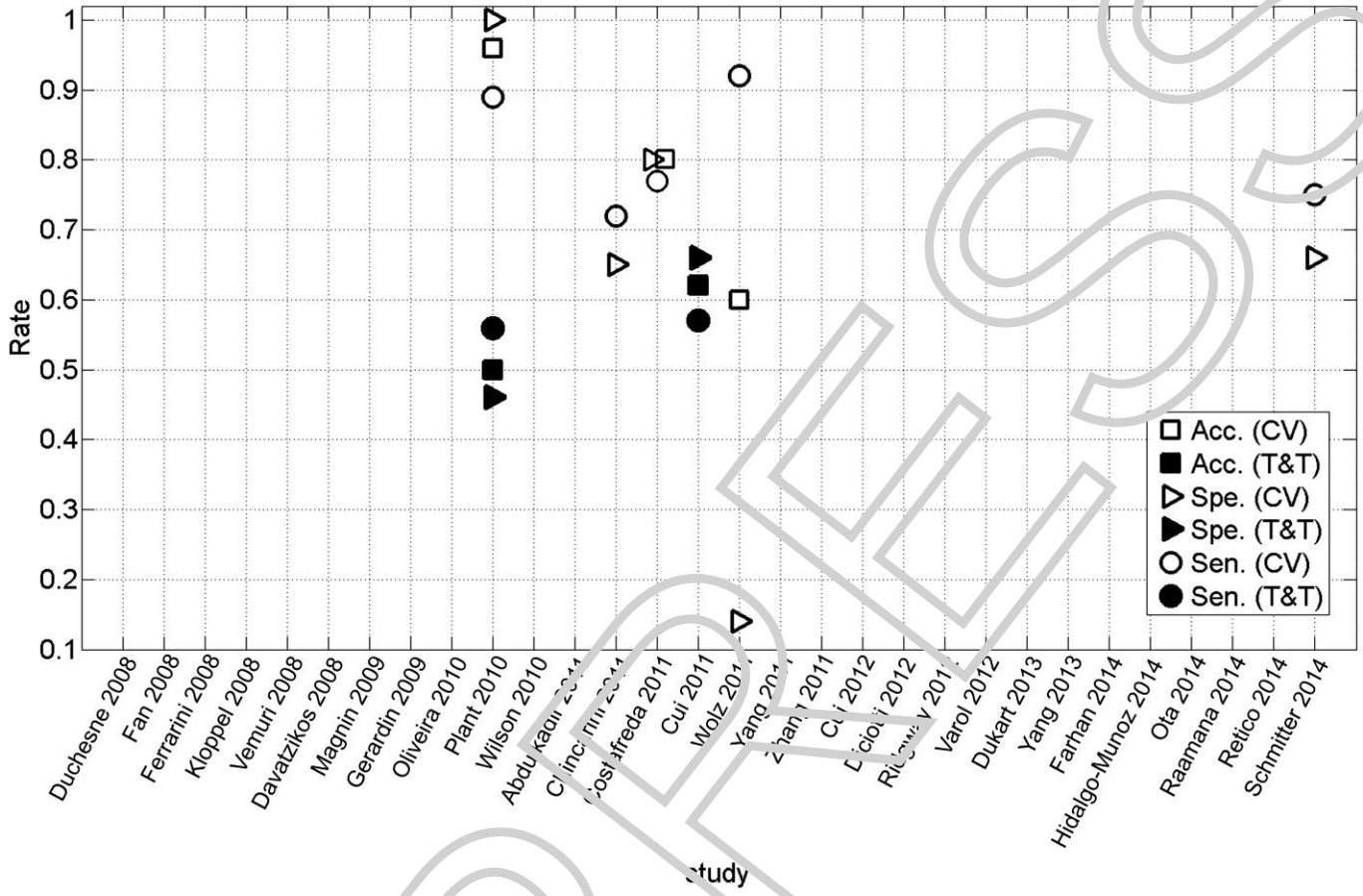


Figure 3 Accuracy (Acc.), specificity (Spe.) and sensitivity (Sen.) rates as a function of the considered study for MCIc vs. MCInc classification. CV and T&T validation strategies are indicated in white and black, respectively. Accuracy, specificity and sensitivity rates are indicated by squares, triangles and circles, respectively.

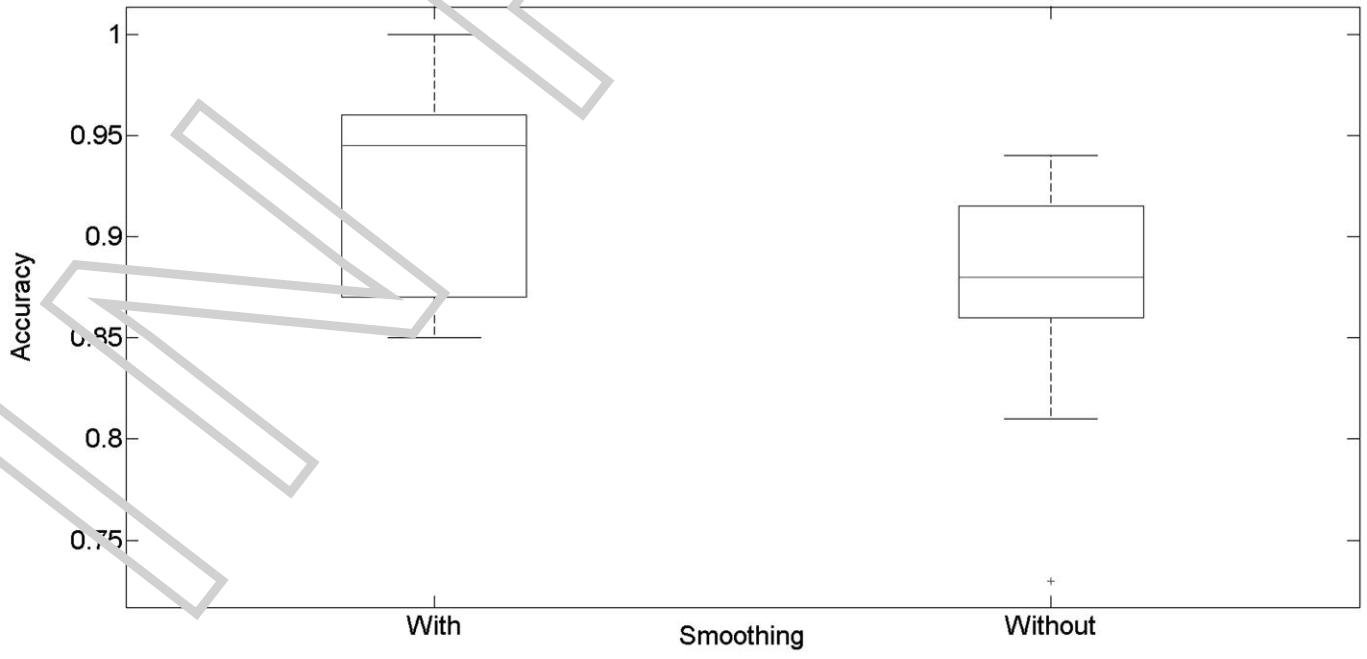


Figure 4 Boxplot showing results for AD vs. CN comparison in terms of accuracy as a function of the applied pre-processing (with or without smoothing).

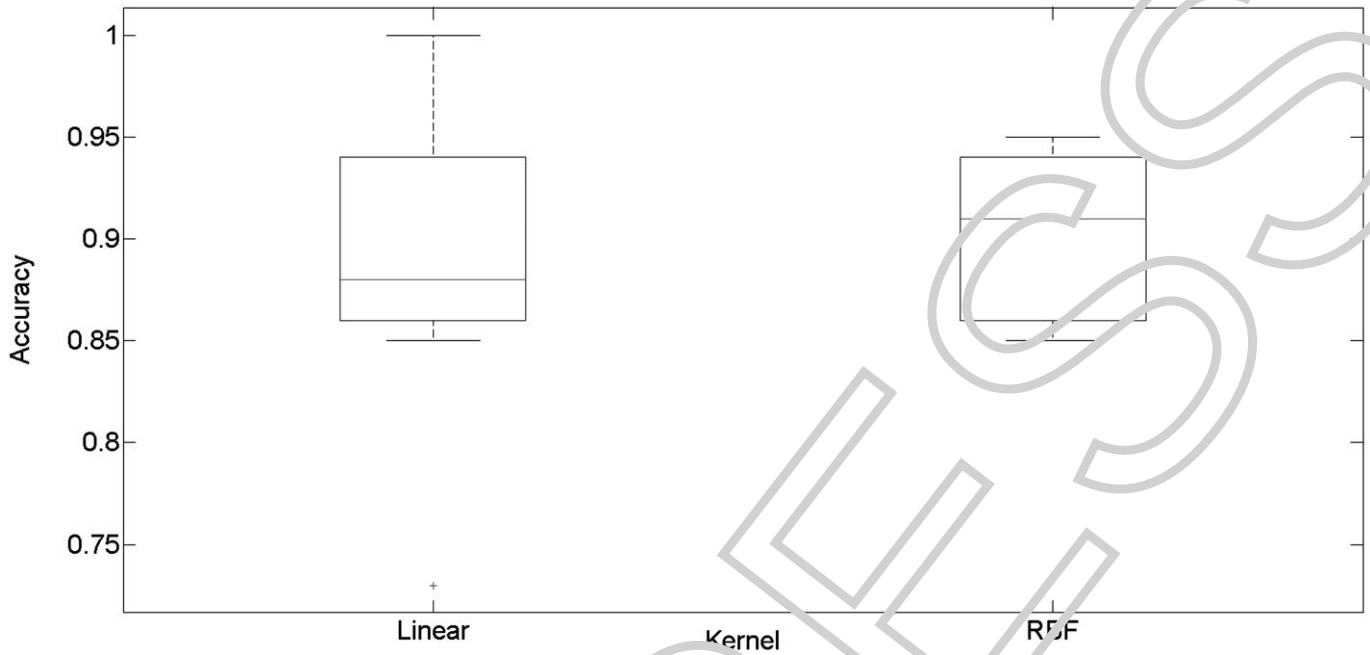


Figure 5 Boxplot showing results for AD vs. CN comparison in terms of accuracy as a function of the SVM kernel function (linear or RBF).

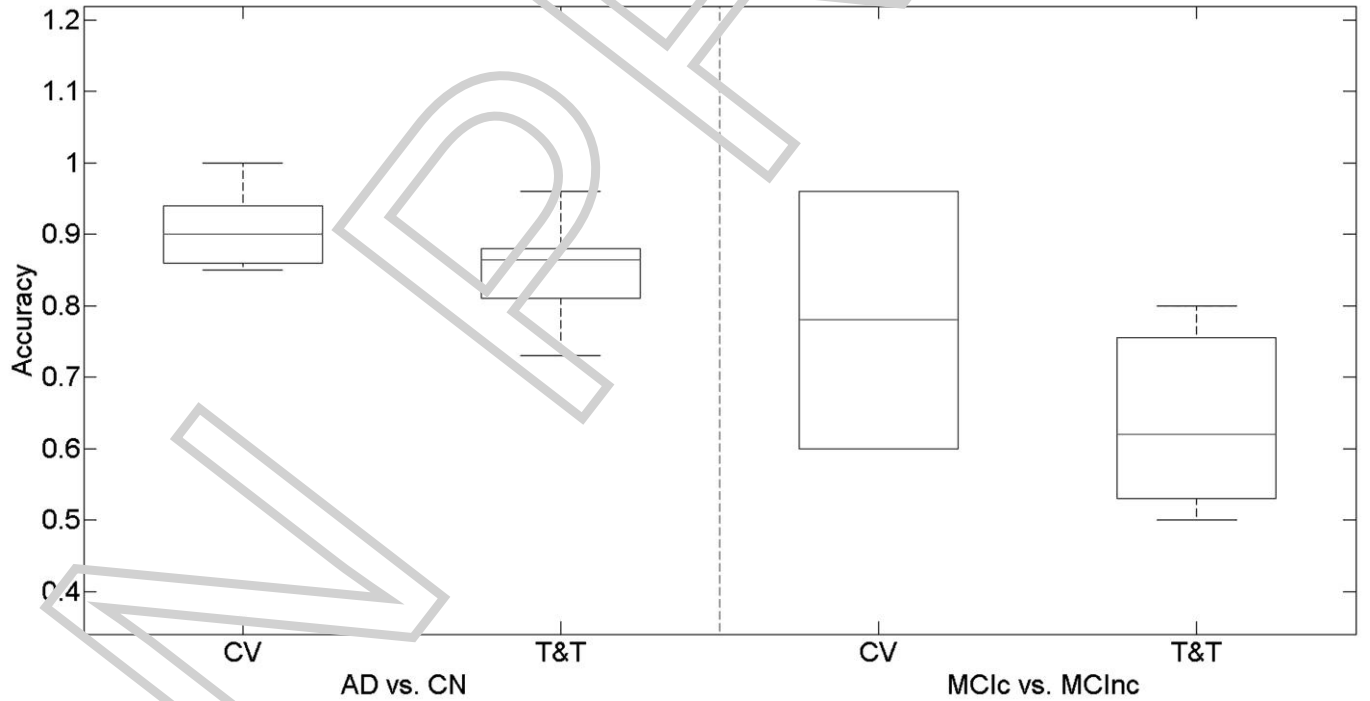


Figure 6 Boxplot showing results of cross validation (CV) and train-and-test (T&T) strategies in terms of accuracy for the following comparisons: AD vs. CN (left), MCIc vs. MCInc (right).

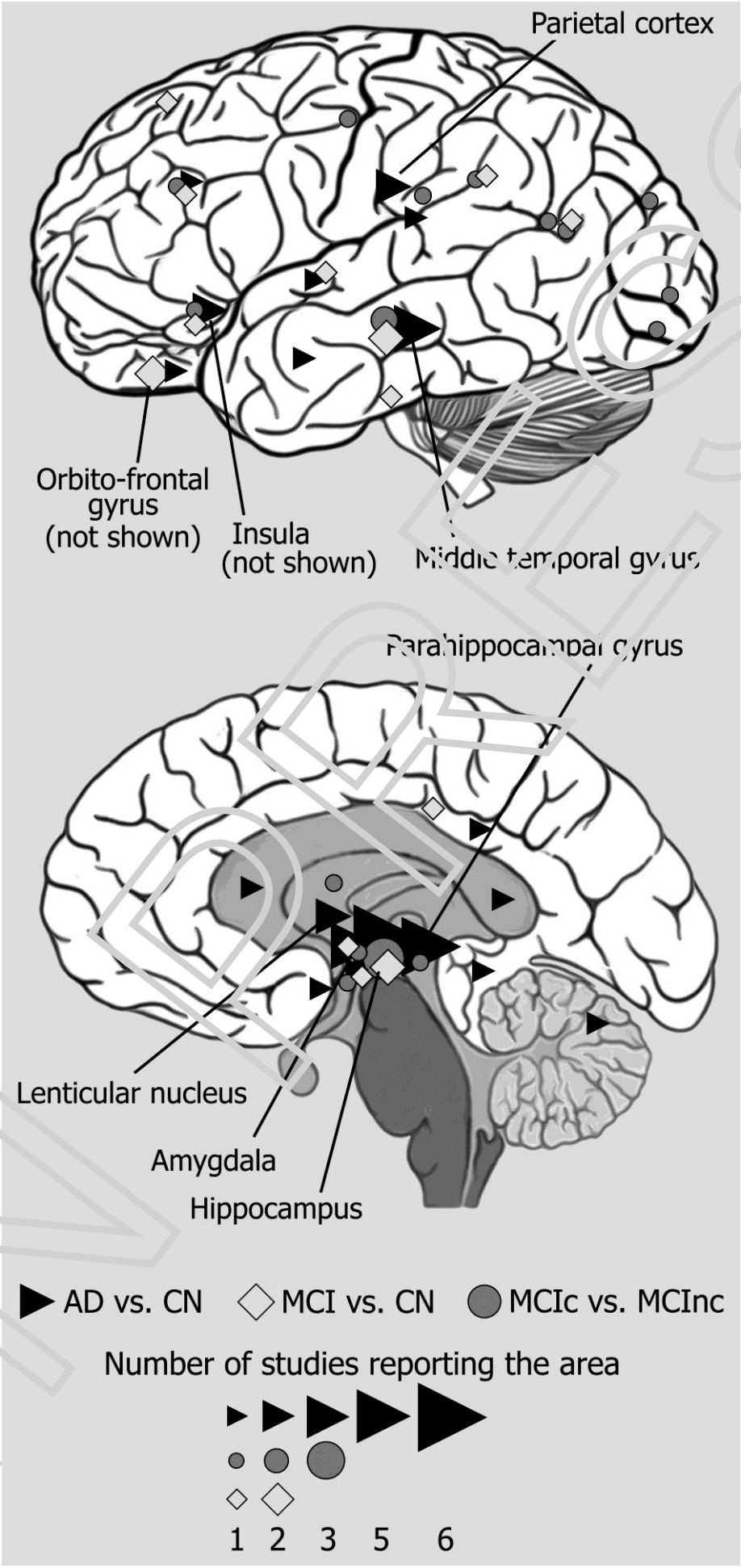

Parietal cortex
Orbito-frontal gyrus (not shown)
Insula (not shown)
Middle temporal gyrus
Parahippocampal gyrus
Lenticular nucleus
Amygdala
Hippocampus
AD vs. CN
MCI vs. CN
MCIc vs. MCInc
Number of studies reporting the area
1 2 3 5 6

Figure 7 An overview of the whole-brain anatomical regions obtained by the considered studies as possible biomarkers for the following comparisons: AD vs. CN (triangles), MCI vs. CN (circles), MCIc vs. MCInc (squares). The name of each structure is specified only if the number of studies reporting its finding was greater than one. Marker size is proportional to the number of studies reporting each structure (from 1 to 6). Lateral (top) and medial (bottom) views are shown.

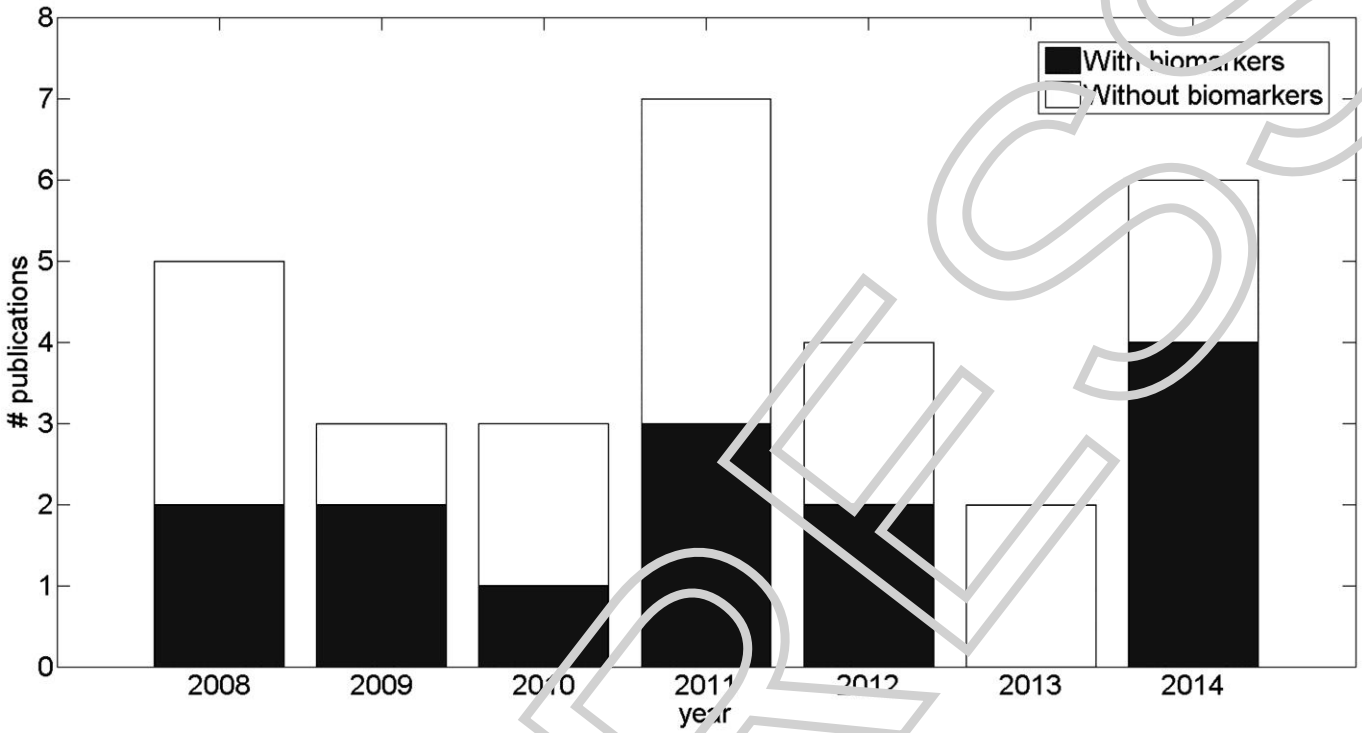


Figure 8 Bar graph showing the number of considered publications with (black) or without (white) extraction of biomarkers as a function of the year.

Table 1 Methods of pre-processing (smoothing), feature extraction, feature selection, classification (kernel function), validation in the reviewed studies (ordered by year) and corresponding results in terms of Accuracy (Acc), Specificity (Spe), Sensitivity (Sen) and AUC for the comparison between AD and CN

Author (year)	Smoothing	Feature extraction	Feature selection	Kernel	Validation	Acc	Spe	Sen	AUC
Duchesne (2008)	-	Data-driven	A priori ROIs	Linear	LOO	.92	-	-	-
Fan (2008)	8 mm^3 FWHM	Tissue segmentation	Pearson correlation and spatial consistency + A posteriori	RBF	LOO	.94	-	-	.97
Ferrarini (2008)	-	Tissue segmentation + Shape modeling	Permutation test	Linear	LOO	.86	.82	.90	-
					T&T	-	-	.76	-
Kloppel (2008)	-	Tissue segmentation	-	Linear	LOO	.96	.94	.97	-
					T&T	.96	.93	1	-
			A priori ROIs	Linear	LOO	.94	.91	.97	-
					T&T	.71	.93	.50	-
Vemuri (2008)	8 mm^3 FWHM + Down-sampling	Tissue Segmentation	A priori ROIs + A posteriori	Linear	T&T	-	.86	.86	-
Magnin (2009)	-	Relative weight of GM w.r.t. WM and CSF	A priori ROIs	RBF	Bootstrap resampling	.95	.97	.92	-
Gerardin (2009)	-	Spherical harmonics	A priori ROIs + Univariate	RBF	LOO	.94	.92	.96	-
					T&T	.88	.92	.84	-
Oliveira (2010)	-	Anatomical measurements	A posteriori	RBF	LOO	.88	.85	.93	-
Plant (2010)	10 mm^3 FWHM	Tissue segmentation	Data-driven	Linear	LOO	.90	.78	.97	-
Abdulkadir (2011)	-	Tissue segmentation	-	Linear	LOO	.88	-	-	-
					T&T	.85	-	-	-
Chincarini (2011)	-	Filtering	A priori ROIs + Data-driven	n.s.	CV	-	.94	.89	.97
Costafreda (2011)	-	Anatomical measurements	A priori ROIs	RBF	CV	.85	-	-	-
Wolz (2011)	20 mm^3 FWHM	Anatomical measurements + Data-driven	-	RBF	LOO	.86	.78	.94	-

Author (year)	Smoothing	Feature extraction	Feature selection	Kernel	Validation	Acc	Spe	Sen	AUC
Yang (2011)	-	Whole brain + Data-driven	-	n.s.	T&T	.77	.80	.74	-
		Tissue segmentation + Data-driven			T&T	.81	.80	.82	-
Zhang (2011)	-	Tissue segmentation + Anatomical measurements	A priori ROIs	Linear	CV	.86	.86	.86	-
Varol (2012)	-	Tissue segmentation	Univariate feature selection	Linear	T&T (ensemble classifier)	.88	.90	.86	.94
Dukart (2013)	12 mm^3 FWHM	Tissue segmentation	A priori ROIs	Linear	LOO	.88	-	-	-
					T&T	.73	.61	.86	-
Yang (2013)	-	Tissue segmentation + Anatomical measurements + Data-driven	-	RBF	LOO	.94	.94	.94	-
Farhan (2014)	-	Tissue segmentation + Anatomical measurements	A priori ROIs	RBF	CV	.94	1	.88	-
Hidalgo-Munoz (2014)	-	Tissue segmentation	RFE	Linear	CV	1	-	-	-
Retico (2014)	Gaussian (n.s. FWHM)	Tissue segmentation	RFE	Linear	CV	-	-	-	.89
Schmitter (2014)	-	Anatomical measurements	A priori ROIs	Linear	LOO	-	.91	.86	-

Acc: accuracy; Spe: specificity; Sen: sensitivity; AUC: Area Under the ROC Curve; FWHM: Full Width at Half Maximum; ROI: Region Of Interest; RBF: Radial Basis Function; CV: Cross Validation; LOO: Leave-One-Out; T&T: train-and-test; n.s.: not specified by the authors.

Table 2 Methods of pre-processing (smoothing), feature extraction, feature selection, classification (kernel function), validation in the reviewed studies (ordered by year) and corresponding results in terms of Accuracy (Acc), Specificity (Spe), Sensitivity (Sen) and AUC for other comparisons than AD vs. CN

Author (year)	Smoothing	Feature extraction	Feature selection	Kernel	Comparisons	Validation	Acc	Spe	Sen	AUC
Fan (2008)	8 mm^3 FWHM	Tissue segmentation	Pearson correlation and spatial consistency + A posteriori	RBF	MCI vs. CN	LOO	.82	-	-	.85
					AD vs. MCI	LOO	.74	-	-	.76
Davatzikos (2008)	-	Tissue Segmentation	Pearson correlation and spatial consistency + A posteriori	RBF	MCI vs CN	LOO	.90	-	-	-
Kloppel (2008)	-	Tissue segmentation	-	Linear	mild AD vs. CN	LOO	.81	.93	.61	-
					AD vs. FTD	T&T	.89	.95	.83	-
			A priori ROIs	Linear	mild AD vs. CN	LOO	.86	.91	.76	-
Gerardin (2009)	-	Spherical harmonics	A priori ROIs + Univariate	RBF	aMCI vs. CN	LOO	.83	.84	.83	-
						T&T	.80	.80	.80	-
Plant (2010)	10 mm^3 FWHM	Tissue segmentation	Data-driven	Linear	MCI vs. CN	LOO	.98	1	.96	-
					MCIc vs. MCInc	LOO	.96	1	.89	-
						T&T (train on AD vs. CN)	.50	.46	.56	-
Wilson (2010)	8 mm^3 FWHM	Tissue segmentation + Data-driven	-	Linear	LPA vs. CN	CV	1	1	1	1
					LPA vs. SD	CV	.94	.94	.94	.98
					LPA vs. PNFA	CV	.81	.81	.81	.88
Chincarini (2011)	-	Filtering (intensity and textural features)	A priori ROIs + Data-driven	n.s.	MCIc vs. CN	CV	-	.80	.89	.92
					MCIc vs. MCInc		-	.65	.72	.74
Costafreda (2011)	-	Anatomical measurements	A priori ROIs	RBF	MCIc vs. MCInc	T&T (train on AD vs. CN)	.80	.80	.77	-
Cui (2011)	-	Anatomical measurements	A priori ROIs + Data-driven + A posteriori	RBF	MCIc vs. MCInc	T&T (train on AD vs. CN)	.62	.66	.57	.65
Wolz (2011)	20 mm^3 FWHM	Anatomical measurements + Data-driven	-	RBF	MCIc vs. CN	LOO	.82	.67	.93	-
					MCIc vs. MCInc	LOO	.60	.14	.92	-
Yang (2011)	-	Whole brain + Data-driven	-	n.s.	MCI vs. CN	T&T	.72	.73	.71	-
		Tissue segmentation + Data-driven			MCI vs. CN	T&T	.71	.69	.73	-

Author (year)	Smoothing	Feature extraction	Feature selection	Kernel	Comparisons	Validation	Acc	Spe	Sen	AUC
Zhang (2011)	-	Tissue segmentation + Anatomical measurements	A priori ROIs	Linear	MCI vs. CN	CV	.72	.60	.79	-
Cui (2012)	-	Anatomical measurements	A priori ROIs + Data-driven + A posteriori	RBF	MCI vs. CN	CV	.64	.64	.64	.73
Diciotti (2012)	-	Anatomical measurements	A priori ROIs	RBF	mild AD vs. CN	CV	.86	.90	.82	-
					mild AD vs. MCI	CV	.74	.77	.72	-
Ridgway (2012)	20 mm^3 FWHM	Anatomical measurements	-	n.s.	PCAt vs. LPA	LOO	-	-	-	.85
					LPA vs. tAD	LOO	-	-	-	.78
					PCA vs. tAD	LOO	-	-	-	.76
Yang (2013)	-	Tissue segmentation + Anatomical measurements + Data-driven	-	RBF	MCI vs. CN	LOO	.89	.89	.94	-
Ota (2014)	8 mm^3 FWHM	Tissue segmentation + Anatomical measurements	A priori ROIs + RFE (accuracy-based)	RBF	MCIc vs. MCInc	LOO	.78	.79	.77	-
Raamana (2014)	10 mm^3 FWHM	Anatomical measurements	-	n.s.	sd-aMCI vs. CN	T&T	.50	.44	.58	.52
					md-aMCI vs. CN	T&T	.61	.60	.62	.66
					sd-aMCI vs. md-aMCI	T&T	.53	.53	.53	.54
Retico (2014)	Gaussian (n.s. FWHM)	Tissue segmentation	RFE (weight-based)	Linear	MCIc vs. MCInc	T&T (train on AD vs. CN)	-	-	-	.71
Schmitter (2014)	-	Anatomical measurements	A priori ROIs	Linear	MCI vs. CN	LOO	-	.83	.69	-
					AD vs. MCI	LOO	-	.67	.69	-
					MCIc vs. MCInc within 2 years	LOO	-	.71	.67	-
					MCIc vs. MCInc within 3 years	LOO	-	.66	.75	-

Acc: accuracy; Spe: specificity; Sen: sensitivity; AUC: Area Under the ROC Curve; FWHM: Full Width at Half Maximum; ROI: Region Of Interest; RBF: Radial Basis Function; MCI: Mild Cognitive Impairment; CN: Controls; AD: Alzheimer's Disease; FTD: Fronto-Temporal lobar Degeneration; aMCI: amnestic MCI; MCIc: MCI-converter; MCInc: MCI-non converter; LPA: Logopenic Progressive Aphasia; SD: Semantic Dementia; PNFA: Progressive Non-Fluent Aphasia; PCAt: Posterior Cortical Atrophy; tAD: typical AD; sd-aMCI: single domain amnestic MCI; md-aMCI: multi domain amnestic MCI; CV: Cross Validation; LOO: Leave-One-Out; T&T: train-and-test; n.s.: not specified by the authors.

Table 3: Method and results of biomarkers extraction for AD vs. CN in the reviewed studies

Author (year)	Extraction of biomarkers	Biomarkers
Kloppel (2008)	Weight map	GM: parahippocampal gyrus, parietal cortex
Vemuri (2008)	Weight map	GM: medial temporal lobe, temporal-parietal association cortex, posterior cingulate/precuneus, insula. WM: temporal lobe, parahippocampal gyrus, parietal lobe
Magnin (2009)	Feature classification performance	GM: hippocampus, parahippocampal gyrus
Oliveira (2010)	Feature classification performance	GM: anterior and posterior corpus callosum, left and right hippocampus, right lateral ventricular horn
Chincarini (2011)	Feature classification performance	GM: ROIs including (ranked by importance) hippocampus (l), amygdala (r), hippocampus (r), middle inferior temporal gyrus (l), amygdala (l), middle inferior temporal gyrus (r)
Costafreda (2011)	Weight map	GM: bilateral atrophy in lateral and medial aspects of hippocampal head, lesser extent in hippocampal body
Farhan (2014)	Feature classification performance	GM: area of left hippocampus
Hidalgo-Munoz (2014)	Weight map + Feature classification performance	GM: hippocampi, entorhinal cortex, parahippocampal region, insular cortex, amygdala, lenticular nucleus, fusiform gyrus
Retico (2014)	Weight map	GM:23 main regions including uncus, inferior frontal gyrus, declive, middle frontal gyrus, parahippocampal gyrus (amygdala), superior temporal gyrus, middle temporal gyrus, lentiform nucleus (putamen)

GM: Gray Matter; WM: White Matter.

Table 4: Method and results of biomarkers extraction in the reviewed studies for MCI vs. CN, MCIc vs. CN, MCIc vs. MCInc, mild AD vs. CN, mild AD vs. MCI and AD vs. FTD

Author (year)	Extraction of biomarkers	Biomarkers
MCI vs. CN		
Davatzikos (2008)	Weight map	GM: Lateral and inferior parts of hippocampi, bilateral superior, middle and inferior temporal gyri, bilateral orbitofrontal gyrus, left fusiform gyrus, right collateral sulcus, posterior cingulate WM: inferior temporal gyri, middle and superior frontal gyri
Cui (2012)	Feature classification performance	GM: right straight gyrus, right supramarginal gyrus, right short insular gyri, left orbital gyri, left angular gyrus
MCIc vs. CN		
Chincarini (2011)	Feature classification performance	GM: ROIs including (ranked by importance) middle inferior temporal gyrus (l), hippocampus (l), amygdala (r)
MCIc vs. MCInc		
Chincarini (2011)	Feature classification performance	GM: ROIs including (ranked by importance) middle inferior temporal gyrus (l), hippocampus (l), amygdala (r), insula (l)
Cui (2011)	Feature classification performance	GM (ranked by selection frequency): left entorhinal cortex, right middle temporal gyrus, right and left hippocampus, right inferior parietal cortex, left retrosplenial cortex, left middle temporal gyrus
Ota (2014)	Feature classification performance	GM: 37 ROIs including inferior occipital gyrus (l), parahippocammpal gyrus (l), middle frontal gyrus (r), middle occipital gyrus (r), superior occipital gyrus (l), supramarginal gyrus (r), angular gyrus (l), precentral gyrus (l), caudate (r), hippocampus (l)
mild AD vs. CN		
Diciotti (2012)	Feature classification performance	GM: subcortical and cortical volumes
mild AD vs. MCI		
Diciotti (2012)	Feature classification performance	GM: subcortical and cortical volumes
AD vs. FTD		
Kloppel (2008)	Weight map	GM: frontal and parietal areas

MCI: Mild Cognitive Impairment; MCIc: MCI-converter; MCInc: MCI-non converter; GM: Gray Matter; WM: White Matter.

PUBLICATION IV

Application of the implemented ML method to the diagnosis of ED (Application III)

Hindawi Publishing Corporation
Behavioural Neurology
Volume 2015, Article ID 924814, 10 pages
http://dx.doi.org/10.1155/2015/924814

Research Article

Biomarkers of Eating Disorders Using Support Vector Machine Analysis of Structural Neuroimaging Data: Preliminary Results

Antonio Cerasa,[1] Isabella Castiglioni,[2] Christian Salvatore,[3] Angela Funaro,[3] Iolanda Martino,[1] Stefania Alfano,[3] Giulia Donzuso,[1] Paolo Perrotta,[1] Maria Cecilia Gioia,[1] Maria Carla Gilardi,[3] and Aldo Quattrone[1,4]

[1]*IBFM-CNR, 88100 Catanzaro, Italy*
[2]*IBFM-CNR, University of Milan-Bicocca, H S. Raffaele, Via Fratelli Cervi 93, 20090 Segrate, Italy*
[3]*Associazione Centro Trauma Ippocampo, Via Rossini 5, 87100 Castrolibero, Italy*
[4]*Unit of Neurology, "Magna Graecia" University, 88100 Catanzaro, Italy*

Correspondence should be addressed to Antonio Cerasa; a.cerasa@unicz.it

Received 15 July 2015; Revised 18 September 2015; Accepted 28 September 2015

Academic Editor: Diego Salas-Gonzalez

Presently, there are no valid biomarkers to identify individuals with eating disorders (ED). The aim of this work was to assess the feasibility of a machine learning method for extracting reliable neuroimaging features allowing individual categorization of patients with ED. Support Vector Machine (SVM) technique, combined with a pattern recognition method, was employed utilizing structural magnetic resonance images. Seventeen females with ED (six with diagnosis of anorexia nervosa and 11 with bulimia nervosa) were compared against 17 body mass index-matched healthy controls (HC). Machine learning allowed individual diagnosis of ED versus HC with an Accuracy ≥ 0.80. Voxel-based pattern recognition analysis demonstrated that voxels influencing the classification Accuracy involved the occipital cortex, the posterior cerebellar lobule, precuneus, sensorimotor/premotor cortices, and the medial prefrontal cortex, all critical regions known to be strongly involved in the pathophysiological mechanisms of ED. Although these findings should be considered preliminary given the small size investigated, SVM analysis highlights the role of well-known brain regions as possible biomarkers to distinguish ED from HC at an individual level, thus encouraging the translational implementation of this new multivariate approach in the clinical practice.

1. Introduction

Eating disorders (ED) are typically adolescent-onset psychiatric conditions that cause serious disturbances to everyday diet, such as eating extremely small amounts of food or severely overeating. Female gender has been demonstrated as a potent risk factor for eating disorders [1], but how much this association can be attributed to biological rather than social factors is uncertain [2]. The most investigated clinical phenotypes of ED are anorexia nervosa (AN) and bulimia nervosa (BN). AN is a serious mental disorder that leads to death in approximately 10% of cases [3]. According to the new DSM-5 criteria, to be diagnosed as having AN a person must display (a) persistent restriction of energy intake relative to requirements leading to a significantly low body weight; (b) intense fear of gaining weight or becoming fat, even though they are underweight; and (c) disturbance in the way in which one's body weight or shape is experienced, undue influence of body weight or shape on self-evaluation, or denial of the seriousness of the current low body weight. Otherwise, BN is characterized by frequent episodes of binge eating followed by inappropriate behaviors such as self-induced vomiting to avoid weight gain. DSM-V criteria reduce the frequency of binge eating and compensatory behaviors that people with BN must exhibit, to once a week from twice weekly as specified in DSM-IV.

To date, individual diagnosis of ED is based only on a clinical interview complemented by physical, psychopathological, and behavioral examinations aimed at assessing the existence of physical, emotional, behavioral and cognitive

disturbances. However, ED diagnosis is unstable, with clinical features changing over time (i.e., weight normalization [11]) and often switching from anorexia to bulimia [4]. For this reason, there is an urgent need to identify biomarkers which may be used for helping and improving early diagnosis, treatment planning, and monitoring of disease progression. In the past 10 years, considerable effort has been expended in developing advanced neuroimaging methods. As a result, a plethora of functional and structural neuroimaging studies have been performed to unravel the pathophysiological mechanisms of ED [5–9]. Whereas the vast majority of these studies reported in AN patients global reductions of total gray and white matter [10], as well as cortical thickness [11], a number of recent studies have emphasized regional group differences. What has been proposed was that AN patients are characterized by widespread brain abnormalities involving (a) the mesolimbic regions (striatum, hippocampus, amygdala, and cerebellum), (b) the dorsolateral prefrontal cortex, (c) the visual cortex, and (d) the cerebellum [12]. Otherwise, neuroimaging literature about BN summarizes the presence of specific involvement of the reward neural system (ventral striatum, nucleus caudate, anterior cingulate cortex (ACC), orbitofrontal cortex (OFC)), hypothesizing that during binge eating a person must consume greater quantities of food to achieve the feeling of satisfaction, like an addict [13–15].

Although significant results have been achieved, the disadvantage of these studies is that they reported neurobiological abnormalities comparing patients and controls at a group level, with consequently limited clinical translation at the individual level. For this reason, attention has recently turned toward alternative kinds of analyses of neuroimaging data. In the last few years, there has been growing interest within the neuroimaging community in classification methods, including machine learning methods. These techniques are based on algorithms able to automatically extract multiple pieces of information from image sets without requiring *a-priori* hypotheses of where they may be found on images. The aim of these methods is to maximize the distance between image groups in order to classify individual structural or functional brain images. Several studies have assessed the clinical relevance of these techniques showing very promising findings mainly in the neurological realm. For instance, machine learning techniques are able to identify very reliable imaging biomarkers allowing individual diagnoses of Alzheimer's disease [16, 17], Mild Cognitive Impairment [18], and Parkinson's disease [19, 20] with an Accuracy of above 90%. In the psychiatric realm, this kind of advanced neuroimaging method is in its relative infancy. Although some interesting applications have been made in patients with posttraumatic stress disorders [21], depression disorders [22], and first-episode psychosis [23], there are no studies investigating the potential role of these methods in ED.

For this reason, this study was aimed at employing a validated supervised machine learning method to define reliable neuroimaging biomarkers useful to distinguish individual with diagnosis of ED patients from healthy controls (HC) by means of structural T1-weighted magnetic resonance images (MRIs). This method makes uses of Principal Components Analysis (PCA) in order to extract the most informative features from MR images [24], while the Support Vector Machine (SVM) approach was used to perform classification [25]. Maps of voxel-based pattern distribution of structural brain differences were generated. These maps show the significance of each image voxel for SVM group discrimination.

2. Methods

2.1. Participants. From 2011 to 2012, a total of 103 patients presenting a first diagnosis of ED were enrolled in this study. All patients were diagnosed by two psychiatrists specialized in ED using the Structured Clinical Interview for Diagnosis (SCID) for DSM-IV-TR. After reviewing the diagnostic information, the psychiatrists made a final diagnosis of ED subtype and proposed the patient's participation in this research project. Inclusion criteria were as follows: (1) age range from 18 to 40 years, (2) being female, and (3) right-handedness. Exclusion criteria were as follows: (1) neurological illness (such as Epilepsy or mental retardation); (2) Axis II disorders (using the SCID-II for DSM-IV-TR) to exclude comorbidity with personality disorders; (3) presence of brain lesions as well as history of cerebrovascular disease, head trauma, or hypertension; (4) psychotropic medication; (5) drug or alcohol abuse; (6) claustrophobia; and (7) past recovery from ED symptoms or psychiatric disorders. After a careful evaluation of these criteria, 17 females with ED were eligible for this study. This group included 11 patients fulfilling DSM-IV criteria for BN and six patients fulfilled DSM-IV criteria for AN restrictive-type. Duration of illness was rather short for all patients (mean duration: 16 ± 5 months).

ED patients were compared with a group of HC. Eighty-one healthy volunteers were recruited by local advertisements. Inclusion criteria for the HC recruitment were as follows: (1) no previous histories of neurological or psychiatric diseases or abnormal brain MRIs and (2) being within the normal range on the Italian version of Minnesota Multiphasic Personality Inventory-2 (MMPI-2) [26]. From this large group, we only enrolled subjects having similar demographical characteristics of those detected in ED patients. Particular attention was paid to potential confounding factors, such as BMI, previously demonstrated to influence brain anatomy [27]. Thus, ED and HC individuals were individually pair-matched by a computer-generated program, according to their age, educational level, and BMI (±2) (for further information, see Supplementary Materials available online at http://dx.doi.org/10.1155/2015/924814). A total sample of seventeen female HC was then enrolled in this study.

All participants gave written informed consent to participate in the present study, approved by the Local Ethical Committee according to the Declaration of Helsinki.

2.1.1. Psychiatric Assessment. Before entering the study, participants completed a battery of self-evaluation questionnaires that included the following.

Eating Disorders Inventory-2 (EDI-2). It is a worldwide validated questionnaire that provides a multidimensional

evaluation of the psychological characteristics of AN and BN [28].

Traumatic Experiences Checklist (TEC). It is a self-report measure addressing potentially traumatizing events [29]. Different scores can be calculated including a cumulative score and scores for emotional neglect, emotional abuse, physical abuse, sexual harassment, sexual abuse, and bodily threat from a person.

Dissociative Experiences Scale v. II (DES-II). It is a lifetime 28-item, self-rating questionnaire developed specifically as a screening instrument to identify subjects that are likely to have dissociative symptoms [30].

Somatoform Dissociation Questionnaire-20 (SDQ-20). It is a self-rating scale developed to the investigated somatic component of dissociation. The SDQ-20 discriminates between dissociative and affective disorders (mood and anxiety disorders) and psychotic symptoms, but a cut-off score is not available [31].

Parental Bonding Instrument (PBI). Perceived parental rearing styles were assessed using the Italian version of PBI. PBI is a self-reporting scale with 25 items to rate paternal or maternal attitude during the first 16 years and has four items comprising care and overprotection factors [32].

Eating Attitude Test-26 (EAT-26). It is a 26-item self-rated questionnaire for evaluating ED-related symptoms [33]. The results are presented as a total score (range, 0–78).

Body Image Dimensional Assessment (BIDA). The BIDA is a silhouette-based scale that starts from neutral figural stimuli and attributes a direct quantitative value to the subject's own current and ideal body image, the most sexually attractive figure, and the most common figure of same-gender-and-age fellows [34].

Finally, for assessing anxiety symptoms, we employed the Hamilton rating scale for anxiety (HAM-A), whereas for defining depression status we employed the Beck Depression Inventory (BDI).

2.1.2. MRI Acquisition. Brain MRI was performed according to our routine protocol by a 3 T scanner with an 8-channel head coil (Discovery MR-750, GE, Milwaukee, WI, USA). Structural MRI data were acquired using a 3D T1-weighted spoiled gradient echo sequence with the following parameters: TR: 9.2 ms, TE: 3.7 ms, flip angle 12°, and voxel-size $1 \times 1 \times 1\,\text{mm}^3$. Subjects were positioned to lie comfortably in the scanner with a forehead-restraining strap and various foam pads to ensure head fixation. All acquired images were visually inspected by expert physicians and neuroradiologists to ensure that none showed signal artifacts.

2.2. Classification of MRI Studies: The Machine Learning Method. We employed a validated supervised machine learning method [20] for the individual differential diagnosis of ED. PCA was applied to whole-brain structural T1-weighted MRIs in order to extract the most informative features for class discrimination, while a SVM algorithm [25] was used to perform classification.

2.2.1. Image Preprocessing. Using the "Tools For NIfTI And ANALYZE Image" toolbox (http://www.mathworks.com/matlabcentral/fileexchange/8797), original images were imported into the Matlab platform (Matlab version R2011b, The MathWorks, Natick, MA).

Image preprocessing was achieved by means of the VBM8 toolbox [35] implemented in the SPM8 software package [36]. This step involved (1) reorientation; (2) cropping; (3) skull-stripping; (4) spatial nonlinear normalization to the MNI152 reference space; (5) smoothing using a Gaussian kernel with full-width at half maximum of $8 \times 8 \times 8$ mm. Resulting nonmodulated whole-brain images were used as input to the feature extraction procedure. Final volume size was of $121 \times 145 \times 121$ voxels. VBM8 was also employed to automatically calculate the total gray matter (GM) and white matter (WM), as well as cerebrospinal fluid (CSF) volumes.

It is worth noting that all images were visually controlled after each step of the preprocessing flow in order to identify possible problems occurring as a consequence of the applied operations.

2.2.2. Feature Extraction. After preprocessing, PCA was applied to structural T1-weighted MRIs considering whole brain, in order to select the most informative features for class discrimination [24, 37]. PCA mainly consists of two steps: the first step is the application of an orthogonal transformation to the dataset, which results in a set of values of linearly uncorrelated variables, or eigenvectors, called "principal components"; extracted principal components are ordered by their variance. The maximum number of eigenvectors that can be extracted with a nonzero associated eigenvalue is related to the lower sample dimension of the dataset. In this case, the number of extracted eigenvectors with a nonzero associated eigenvalue can at most be equal to $N - 1$, N being the number of subjects involved.

The second step is the projection of the dataset itself into the PCA subspace, which heavily decreases the number of features to be handled. Features resulting from this analysis are called PCA coefficients, and they are the ones used for classification in place of the original dataset [38]. For group comparison, we also studied the percentage of retained variance as a function of the number of considered principal components.

Obtained PCA coefficients were finally ordered according to their Fisher Discriminant Ratio (FDR), with the aim of identifying the most discriminative PCA coefficients. Indeed, FDR provides information about the class discriminatory power of a given component, that is, the ability of each component to separate the samples belonging to the two classes. FDR was calculated as follows:

$$\text{FDR} = \frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2}, \quad (1)$$

μ_i and σ_i^2 being the mean and the variance of the ith class, respectively.

2.2.3. Classification Algorithm. SVM algorithm was used to perform classification [20, 25, 39]. Given a set of training data, each piece consisting of an input vector $x_i R^N$ (where i runs from 1 to the number N of samples) and the corresponding label t_i {±1}, the task of the SVM is then to compute the optimal separating hyperplane between the two training classes that will be able to classify unseen examples (x, t) in a correct way. This is done in terms of distance between classes; that is, the optimal separating hyperplane is computed so that its distance from the two training classes to be divided is maximized. The optimal separating hyperplane will then be used as a decision function to classify unseen data as belonging to one of the two training classes. Mathematically, the decision function is defined as follows:

$$y(x) = \sum_{n=1}^{N} a_n \cdot t_n \cdot f(x, x_n) + b. \quad (2)$$

N is the number of samples belonging to the training set; a_n is a weight constant; $f(x, x_n)$ is the kernel function; b is a threshold parameter. Using this decision function, class $y(x)$ for unseen data (x, t) can be predicted.

The implementation of the SVM classification algorithm was carried out using the biolearning toolbox in Matlab. Original datasets were divided into two discrimination groups: ED versus HC. The extracted PCA coefficients and the corresponding labels were used as features to train the classification algorithm [20]. A number k of PCA coefficients was used, where k runs from 1 to the total number N_{PC} of extracted PCA coefficients. A linear kernel was chosen for two reasons: (1) it is able to improve generalization ability; (2) it is the only kernel function that allows the computation of weights and, thus, the generation of voxel-based pattern distribution maps of brain structural differences.

2.2.4. Performances of the Classifier. In order to evaluate the performance of the supervised machine learning method, subjects were randomly divided into 20 subsets, each one containing the same proportion of class labels. Evaluation was performed *via* 20-fold Cross-Validation (CV), by which in turn the training of the classifier was performed using 19 subsets and the testing was performed using the remaining one. This procedure was then repeated 20 times, until all subsets were used once as testing set. In addition, classification performance was also evaluated by 10-fold CV.

Accuracy, Specificity, and Sensitivity were computed over the first k PCA coefficients, where k runs from 1 to the total number N_{PC} of extracted PCA coefficients, as follows:

$$\begin{aligned} \text{Accuracy}_i &= \frac{N_{CC}}{N}, \\ \text{Specificity}_i &= \frac{A_{CC}}{A_{CC} + B_{IC}}, \\ \text{Sensitivity}_i &= \frac{B_{CC}}{B_{CC} + A_{IC}}, \end{aligned} \quad (3)$$

where N is the total number of images which underwent classification; N_{CC} is the total number of Correctly Classified (CC) images; A_{CC} is the number of CC images belonging to the first group; A_{IC} is the number of Incorrectly Classified (IC) images belonging to the first group; B_{CC} is the number of CC images belonging to the second group; B_{IC} is the number of IC images belonging to the second group. It is worth noting that for each round of CV, image preprocessing and feature extraction were performed separately on the training and the testing sets. Accuracy was evaluated as a function of the number of employed PCA coefficients.

2.2.5. Voxel-Based Pattern Distribution. For each discrimination group, maps of voxel-based pattern distribution of brain structural differences were generated. These maps show how significant each image voxel is for SVM group discrimination [20]. In the training phase, in fact, SVM assigns a specific weight to each sample (i.e., the vector of extracted PCA coefficients of each subject) in the training set, this weight reflecting the importance of that sample for group discrimination. In our case, deriving discriminative voxels basing on the SVM weights cannot be done in a direct way, because we use PCA coefficients as input to the SVM instead of image voxels. In order to do this, an intermediate step was needed, that is, back-projection of each sample (i.e., PCA coefficients) from the PCA space to the voxel space. Through this operation, we obtained a back-projected image of the brain of each subject in the voxel space. Finally, maps of values showing the importance of each voxel for group discrimination based on the SVM weights were then obtained by multiplying each back-projected brain of the training set with the corresponding weight assigned by SVM and by summing the results on a voxel basis [16, 20].

However, the multivariate SVM algorithm was not designed to provide single features and their importance. As a consequence, the method to derive discriminative features from the SVM model is a tweak that should be used with caution, because the interpretation of weights assigned by SVM during the training phase could lead to incorrect conclusions. In order to avoid this, we applied the method proposed by Haufe and colleagues to compute activation patterns for backward models [40]. This method ensures the correct interpretation of weights assigned by SVM. Accordingly, in addition to the weight map, we obtained a map of voxel-based pattern distribution of MR image differences between ED and HC.

Both the weight map and the voxel-based pattern distribution obtained using the method proposed by Haufe and colleagues [40] were normalized to a range between 0 and 1, expressed by a proper color scale and superimposed on a standard stereotactic brain for spatial localization. This approach allowed the identification of new MR-related biomarkers for the diagnosis of ED patients (see Supplementary Materials for further information).

2.2.6. Statistical Analysis. Statistical analysis was performed with STATISTICA Version 6.0 (http://www.statsoft.com/). Assumptions for normality were tested for all continuous

variables by using the Kolmogorov-Smirnov test. All variables were normally distributed, except for educational level. Then, Unpaired t-test and Mann-Whitney U test were applied appropriately to assess potential differences between groups for all demographic clinical and MRI variables. All statistical analyses had a 2-tailed alpha level of <0.05 for defining significance.

3. Results

3.1. Clinical Data. Compared with age-/sex-/BMI-matched controls, ED patients did not show global anatomical atrophies in white or gray matter brain volumetry. At a behavioral level, ED group displayed a well-known psychopathological profile (Table 1 and Supplementary Materials). In particular, EDI-2 demonstrated that ED patients had higher scores for (a) drive for thinness scale ($t = 4.45$; p-level < 0.00001); (b) bulimia scale ($t = 2.69$; p-level = 0.01); (c) interoceptive awareness scale ($t = 3.81$; p-level = 0.0006); (d) asceticism scale ($t = 3.81$; p-level = 0.0006); (e) body dissatisfaction ($t = 3.5$; p-level = 0.001); (f) interpersonal distrust scale ($t = 2.07$; p-level = 0.04); and (g) impulse regulation scale ($t = 2.46$; p-level = 0.02). Otherwise, no significant differences were detected for Perfectionism, Ineffectiveness, Maturity Fears, and Social Insecurity scales, in agreement with previous studies [13].

3.2. The Machine Learning Method. Among MR images acquired for this study, no images were excluded from the subsequent analysis due to problems with image quality or problems occurred during preprocessing. As a representative example, 1st and 2nd extracted PCA coefficients that showed the highest FDR are plotted in Figure 1(a) for the ED versus HC group discrimination (data from a single round of CV). In this case, the number of subject involved was equal to 31 (16 ED, 15 HC). The total number of extracted PCA coefficients was equal to 30. The analysis of variance for the ED versus HC group discrimination showed that the percentage of variance retained by the first principal component was equal to 27.0%, while the number of extracted principal components accounting for 50% and 95% of the whole variance was 6 and 27, respectively.

Table 2 shows FDR values of the 30 features (PCA coefficients) used for the ED versus HC group discrimination. Data from a single round of CV are shown as a representative example. As it can be seen, in this case the 8th PCA coefficient showed the highest FDR value, thus resulting the most important feature for group discrimination.

In Figure 1(b), 1st and 2nd extracted PCA coefficients that showed the highest FDR (i.e., after FDR raking) are plotted jointly with 1st and 2nd extracted PCA coefficients (before FDR ranking). As it is shown in this plot, FDR allows finding those features for which discrimination between groups is maximized.

3.3. Classification Algorithm. Figure 2 shows the decision function resulting from the SVM training phase for the ED versus HC group discrimination (1st and 2nd components with highest FDR).

Table 1: Demographic characteristics.

Variables	ED ($n = 17$)	HC ($n = 17$)	P-level
Demographical data			
Age (years)	30.2 ± 5.6	30.1 ± 5.5	0.95
Educational level (years)	17 (13–21)	17 (13–21)	0.88
BMI	23.6 ± 8.2	24.1 ± 4.8	0.79
MRI data			
Total GM Volume	587.3 ± 37.5	608.88 ± 42.1	0.11
Total WM Volume	486.5 ± 63.1	489.6 ± 41.6	0.86
Total CSF Volume	188.3 ± 28.7	187 ± 23.2	0.88
Clinical data			
HAMA	14.6 ± 13	4 ± 2.2	0.04*
BDI	16.8 ± 10.1	6.3 ± 4.7	0.0004*
DES	14.32 ± 12.4	5.12 ± 4	0.007*
EAT-26	23.3 ± 14.4	6.35 ± 3.2	0.00004*
SDQ-20	28.64 ± 14.8	20.6 ± 1.1	0.03*
BIDA	29.9 ± 19.4	19.9 ± 11	0.24
Clinical data EDI-2 scale			
Drive for thinness	9.4 ± 6.3	1.2 ± 1.3	0.0001*
Bulimia	3.47 ± 4.5	0.1 ± 0.5	0.01*
Interoceptive awareness	7.9 ± 6.2	0.7 ± 1.2	0.0006*
Asceticism	5.6 ± 3.8	2 ± 1.1	0.0006*
Body dissatisfaction	12.9 ± 7.2	6.1 ± 2.9	0.001*
Perfectionism	4.3 ± 3.9	3.3 ± 3.1	0.41
Interpersonal distrust	3.6 ± 3.1	1.4 ± 1.2	0.04*
Impulse regulation	3.67 ± 4.9	0.6 ± 1.4	0.02*
Ineffectiveness	3.5 ± 5.2	1.2 ± 2.6	0.12
Maturity fears	5.2 ± 3	3.94 ± 2.6	0.13
Social insecurity	3.53 ± 3.2	2.1 ± 2	0.22

Data are given as mean values (SD) or median values (range) when appropriate.
BMI: Body Mass Index; GM: gray matter; WM: white matter; CSF: cerebrospinal fluid; PBI: parental bonding instrument; STAI: State-Trait Anxiety Inventory; HAMA: Hamilton rating scale for anxiety; BDI: Beck Depression Inventory; DES: Dissociative Experiences Scale; EAT-26: eating attitude test-26; SDQ-20: Somatoform Dissociation Questionnaire-2; BIDA: Body Image Dimensional Assessment; EDI-2: Eating Disorder Inventory-2. Total brain MRI parameters have been calculated using VBM8 tool. *Significant difference.

3.4. Performances of the Classifier. When considering 20-fold CV approach, Accuracy, Specificity and Sensitivity of the classifier for ED versus HC group discrimination were calculated over a number of PCA coefficients ranging from 1 to 32. When using 31 PCA coefficients, Accuracy, Specificity and Sensitivity reached their best values of 0.85, 0.73 and 0.93, respectively.

Figure 3 shows Accuracy, Specificity and Sensitivity as a function of the number of employed PCA coefficients for the ED versus HC group discrimination. As expected, the performance of the classification algorithm increases with the number of employed PCA coefficients.

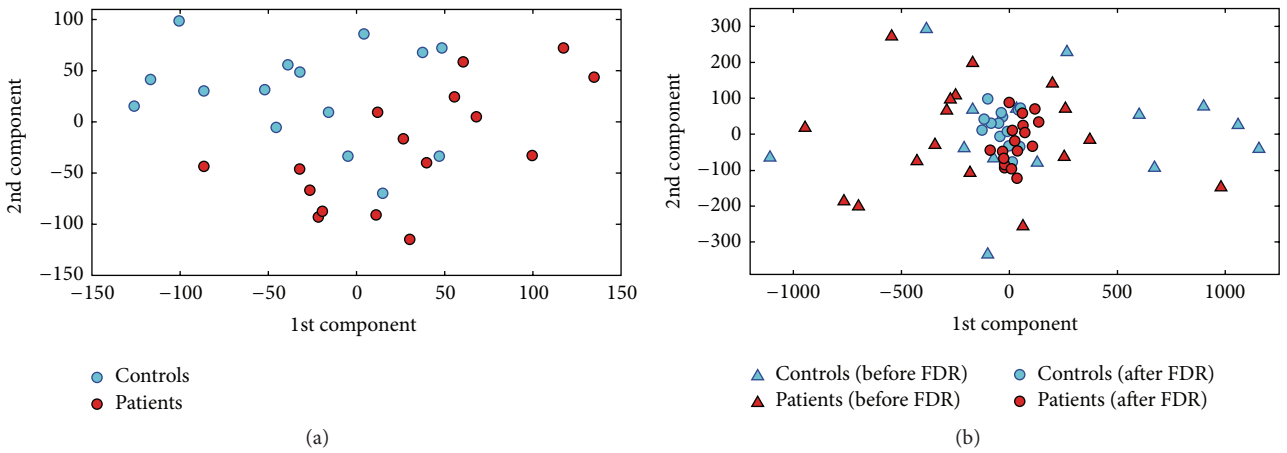


Figure 1: Plot of the PCA coefficients that showed the highest FDR (a) and joint plot of the PCA coefficients before (triangles) and after (circles) FDR ranking (b) for the ED versus HC group discrimination (1st and 2nd components). Data from a single round of CV are shown as a representative example.

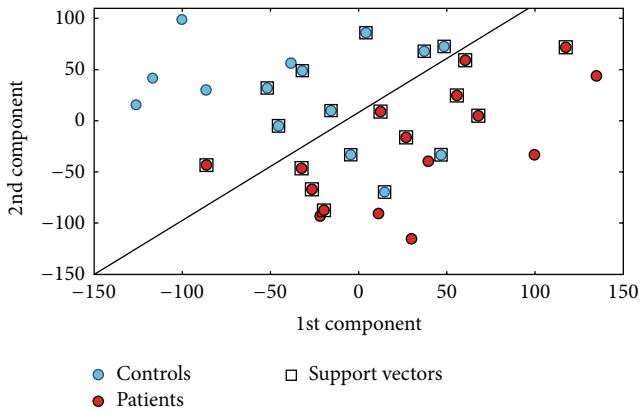


Figure 2: Decision function for the ED versus HC group discrimination (1st and 2nd components with highest FDR). Data from a single round of CV are shown as a representative example.

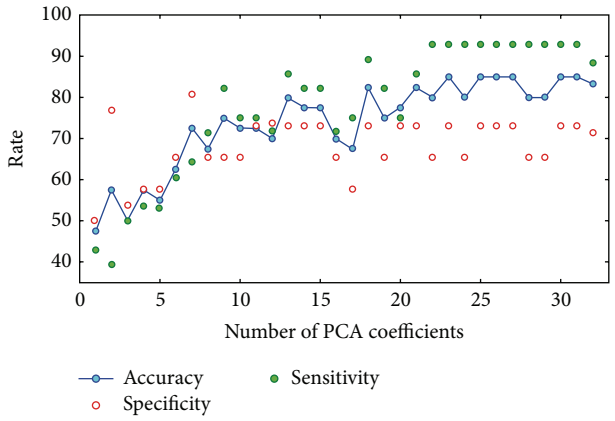


Figure 3: Accuracy, Specificity, and Sensitivity of classification as a function of the number of employed PCA coefficients for the ED versus HC group discrimination (20-fold CV).

When considering 10-fold CV approach, Accuracy, Specificity and Sensitivity of the classifier for ED versus HC group discrimination were calculated over a number of PCA coefficients ranging from 1 to 30. In this case, when using 21 PCA coefficients, Accuracy, Specificity and Sensitivity reached their best values of 0.80, 0.72 and 0.96, respectively.

3.5. Voxel-Based Pattern Distribution. Figure 4 shows the voxel-based pattern distribution map of brain structural differences between ED patients and HC. The pattern of differences emerged mainly in the occipital cortex and the posterior cerebellar lobule. Moreover, other brain regions involved in regulation of emotional processing known to be damaged in ED patients were detected: precuneus, sensorimotor and premotor cortices as well as the ACC and OFC.

4. Discussion

The pathophysiological mechanisms underlying ED remain a matter of debate. In the last few years, several meta-analyses have tried to summarize the large amount of evidence coming from behavioral and neuroimaging realms, providing different key of lectures. At a behavioral level, taking into account the clinical heterogeneity of ED subtypes, a large amount of literature highlights the AN-related psychopathology characterized by excessive Perfectionism, cognitive-behavioral rigidity, asceticism, ruminations, obsessions about food and excessive concerns about weight and shape, whereas BN patients would seem to be characterized by an impulsivity trait with a combination of heightened sensitivity to reward and impaired inhibitory control [15, 41, 42]. As concerns neuroimaging findings, although important pathological markers have been found describing neurobiological differences between AN and BN subtypes, the majority of these findings has never been translated into clinical practice. For this reason, the implementation

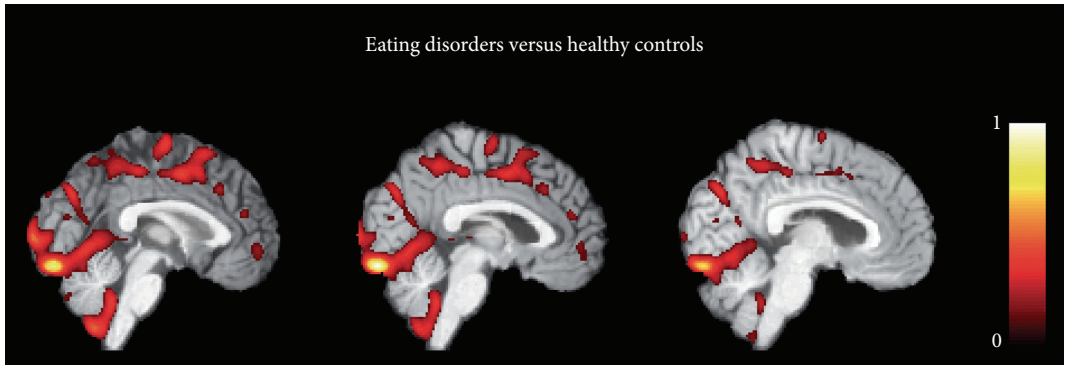


Figure 4: Voxel-based pattern distribution map of brain structural differences between ED patients and healthy controls (sagittal view, threshold = 50%). Voxel-based pattern distribution (normalized to a range between 0 and 1) is expressed according to the color scale and superimposed on a standard stereotactic brain for spatial localization.

Table 2: FDR values of the 30 features (PCA coefficients) used for the ED versus HC discrimination.

PCA coefficient (#)	FDR
1	0.2052
2	0.0172
3	0.0021
4	0.1286
5	0.0005
6	0.0786
7	0.1484
8	0.3923
9	0.0354
10	0.0137
11	0.0919
12	0.3376
13	0.1057
14	0.0002
15	0.0128
16	0.0176
17	0.0279
18	0.0188
19	0.0206
20	0.0511
21	0.0369
22	0.0001
23	0.0200
24	0.0052
25	0.1839
26	0.0431
27	0.0015
28	0.0250
29	0.0321
30	0.0171

Data from a single round of CV are shown as a representative example.

of supervised whole-brain automatic classification methods may become an essential step for improving clinical management of psychiatric patients in longitudinal and prospective studies [43]. SVM has been proposed as a new approach for identifying sensitive biomarkers (or combinations of them) that allow for automatic discrimination of individual subjects. In this work we proposed, for the first time, a SVM algorithm that, working on structural neuroimaging data at a whole-brain level, reached an optimal individual classification in the comparisons between ED patients with controls. The strengths of this work were: (a) the detected pattern of neural abnormalities that allowed the SVM approach to reach this great Accuracy involved well-known brain regions strongly involved in the pathophysiological mechanisms of ED [5–9, 12]; (b) the classification Accuracy in the discrimination of all individual ED patients with respect to controls was equal or higher than those detected in previous studies employing machine learning to classify other psychiatric disorders: ~80–85% in schizophrenic patients [21], 81% in depression disorders [22] and ~75% in first-episode psychosis [23]; (c) the employment of HC matched for BMI, a critical variable known to influence brain anatomy [27] and sparsely controlled in other neuroimaging studies investigating ED patients.

Pattern recognition analysis used to classify ED patients from HC depicted mainly the involvement of the: (a) cerebellum, (b) reward-related cortical regions, (c) occipital cortex and (d) sensorimotor cortex. (a) The cerebellum is a multidimensional brain region involved in a plethora of motor, cognitive and emotional functions. Recent studies have also highlighted the role of the cerebellum in visceral and autonomic regulation, specifically the cerebellar vermis, which has a role in feeding behavior and appetite regulation [44, 45]. This region is extensively connected with limbic brain structures, such as the hippocampus, parahippocampal gyrus, amygdala, thalamus, cingulate and prefrontal cortices [46]. The involvement of the cerebellum (mainly the vermis subregion) in ED has been consistently demonstrated in several structural neuroimaging studies describing the presence of GM volume loss mainly in AN [47–49]. Moreover a recent resting state fMRI study [13], demonstrated the presence of altered intrinsic connectivity of the cerebellar vermis in both AN and BN patients. These authors hypothesized that this dysfunctional neural pattern might be related to some

psychopathological aspects of ED (i.e., the drive thinness) that is pathologically altered in all ED patients. (b) The ACC, together with the OFC, are two regions taking part in the ventral limbic circuit, together with the amygdala, insula and ventral striatum, which are important for identifying the emotional significance of appetizing stimuli for inhibiting impulsive behaviors [14] and regulating reward systems [50]. The current neuroimaging literature mainly highlights the role of this neural network in pathophysiological mechanisms of BN, in which the alterations of mesolimbic reward response mechanisms could explain the lack of control and the impulsivity that are often present in BN patients and that are neurophysiologically expressed through dysfunctional activities in the ACC and OFC regions [15, 41]. However, fronto-striatal neural circuit dysfunctions related to altered reward processing have also described in AN patients [51], thus raising a different perspective in which stimuli that are otherwise aversive for healthy controls (e.g., self-starvation, emaciated body image) are considered rewarding and activate relevant reward linked brain regions in AN patients. (c) The involvement of the visual cortex is another key site associated with ED. Although altered functional activity of the occipital lobe has been reported in both AN and BN individuals [52], body image disturbance is fundamentally considered one of the core characteristics of AN. Several neuroimaging studies have described the neurobiological correlates of this symptom, defining the presence of a specific neural network involved in body processing: the fusiform area, the inferior temporal sulcus and the primary visual cortex. Recent evidence [53] demonstrated altered effective connectivity between these regions in AN patients during the viewing of bodies. (d) Finally, abnormal neural changes in the precuneus and sensorimotor/premotor cortices have been already described in both AN and BN patients [6, 13, 53]. Friederich et al. [41] showed that, using body images of slim fashion models to induce a self-other body shape comparison, AN patients had a higher activation of the premotor cortex. Again, Amianto et al., [13] found altered gray matter volume in the paracentral lobule, precuneus and somatosensory regions when comparing AN and BN patients, as well as the whole ED group, with respect to controls. Altered neural changes in brain areas involved in sensorimotor functions and visuo-proprioceptive information processing may either represent the physiological consequence of physical hyperactivity typical of ED patients [13] or as a dysfunction related to the body awareness. Body awareness is a complex cognition underpinned by aspects of visual perception, proprioception, and touch [54]. The processing of the body image concept requires integration of the different types of body-related perceptual experience and processing of information related to peripersonal space. The presence of altered anatomical changes in these regions together with visual cortex, has been interpreted as a dysfunctional processing of somatosensory information about the perceived body size [6, 55].

One important limitation of this study needs to be considered in discussion of our data: the clinical subtypes of the enrolled ED and the size of these groups. Considering a hypothetical structure of ED as a spectrum (in line with the trans-diagnostic approach of the DSM-V), in this study we enrolled the two extremes of the model. ED is not a uniform disorder characterized by a high heterogeneity in clinical phenotypes. For instance, the 60% of those who exhibit pathological ED behaviors but who do not meet the full criteria for AN or BN, are instead diagnosed as "eating Disorder Not Otherwise Specified" [56]. Again, the diagnosis is further complicated by the presence of other major psychiatric conditions [57], by disease duration [47] and severity of illness [58]. All this evidence highlights that our findings cannot be generalized to all ED populations. Moreover, the small sample size of AN patients as well as the fact that we only included outpatients with a lower disease duration and with mild severity of illness (BMI ~ 17) might have affected the magnitude of our classification Accuracy. Therefore, to sustain the usefulness of SVM application in clinical practice of ED, further studies are warranted employing a larger and heterogeneous sample. Despite this methodological limitation, it is important to highlight that the severe inclusion criteria employed in this study, albeit with a restricted sample selection, eliminated potential confounders (i.e., BMI), thus helping with the interpretation of the results.

In conclusion, our study demonstrates for the first time that using standard morphological brain images, SVM is able to extract neuroimaging biomarkers, which allow to accurately classify individuals with ED. Although we used this method in a diagnostic perspective, the rationale for applying machine learning methods in this psychiatric realm is to allow inferences to be made at the level of the individual for monitoring disease progression as well as improving prevention and treatment decisions. We believe that our preliminary findings offer new avenues for encouraging the application of these multivariate neuroimaging approaches in clinical practice, mainly to differentiate different ED phenotypes.

Abbreviations

ED: Eating disorders
AN: Anorexia nervosa
BN: Bulimia nervosa
ACC: Anterior cingulate cortex
OFC: Orbitofrontal cortex
SVM: Support vector machine.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

References

[1] H. W. Hoek and D. van Hoeken, "Review of the prevalence and incidence of eating disorders," *International Journal of Eating Disorders*, vol. 34, no. 4, pp. 383–396, 2003.

[2] J. Treasure, A. M. Claudino, and N. Zucker, "Eating disorders," *The Lancet*, vol. 375, no. 9714, pp. 583–593, 2010.

[3] S. Nielsen, "Epidemiology and mortality of eating disorders," *Psychiatric Clinics of North America*, vol. 24, no. 2, pp. 201–214, 2001.

[4] C. G. Fairburn and P. J. Harrison, "Eating disorders," *The Lancet*, vol. 361, no. 9355, pp. 407–416, 2003.

[5] O. E. Titova, O. C. Hjorth, H. B. Schiöth, and S. J. Brooks, "Anorexia nervosa is linked to reduced brain structure in reward and somatosensory regions: a meta-analysis of VBM studies," *BMC Psychiatry*, vol. 13, article 110, 2013.

[6] S. Gaudio and C. C. Quattrocchi, "Neural basis of a multidimensional model of body image distortion in anorexia nervosa," *Neuroscience & Biobehavioral Reviews*, vol. 36, no. 8, pp. 1839–1847, 2012.

[7] F. Van den Eynde and J. Treasure, "Neuroimaging in eating disorders and obesity: implications for research," *Child & Adolescent Psychiatric Clinics of North America*, vol. 18, no. 1, pp. 95–115, 2009.

[8] W. H. Kaye, J. L. Fudge, and M. Paulus, "New insights into symptoms and neurocircuit function of anorexia nervosa," *Nature Reviews Neuroscience*, vol. 10, no. 8, pp. 573–584, 2009.

[9] K. van Kuyck, N. Gérard, K. V. Laere et al., "Towards a neurocircuitry in anorexia nervosa: evidence from functional neuroimaging studies," *Journal of Psychiatric Research*, vol. 43, no. 14, pp. 1133–1145, 2009.

[10] J. Seitz, K. Bühren, G. G. Von Polier, N. Heussen, B. Herpertz-Dahlmann, and K. Konrad, "Morphological changes in the brain of acutely ill and weight-recovered patients with anorexia nervosa. A meta-analysis and qualitative review," *Zeitschrift für Kinder—und Jugendpsychiatrie und Psychotherapie*, vol. 42, no. 1, pp. 7–18, 2014.

[11] J. A. King, D. Geisler, F. Ritschel et al., "Global cortical thinning in acute anorexia nervosa normalizes following long-term weight restoration," *Biological Psychiatry*, vol. 77, no. 7, pp. 624–632, 2015.

[12] F. Van den Eynde, M. Suda, H. Broadbent et al., "Structural magnetic resonance imaging in eating disorders: a systematic review of voxel-based morphometry studies," *European Eating Disorders Review*, vol. 20, no. 2, pp. 94–105, 2012.

[13] F. Amianto, F. D'Agata, L. Lavagnino et al., "Intrinsic connectivity networks within cerebellum and beyond in eating disorders," *Cerebellum*, vol. 12, no. 5, pp. 623–631, 2013.

[14] R. Marsh, M. Stefan, R. Bansal, X. Hao, B. T. Walsh, and B. S. Peterson, "Anatomical characteristics of the cerebral surface in bulimia nervosa," *Biological Psychiatry*, vol. 77, no. 7, pp. 616–623, 2015.

[15] S. J. Brooks, M. Rask-Andersen, C. Benedict, and H. B. Schiöth, "A debate on current eating disorder diagnoses in light of neurobiological findings: is it time for a spectrum model?" *BMC Psychiatry*, vol. 12, article 76, 2012.

[16] S. Klöppel, C. M. Stonnington, C. Chu et al., "Automatic classification of MR scans in Alzheimer's disease," *Brain*, vol. 131, no. 3, pp. 681–689, 2008.

[17] M. Dyrba, M. Ewers, M. Wegrzyn et al., "Robust automated detection of microstructural white matter degeneration in Alzheimer's disease using machine learning classification of multicenter DTI data," *PLoS ONE*, vol. 8, no. 5, Article ID e64925, 2013.

[18] C. Plant, S. J. Teipel, A. Oswald et al., "Automated detection of brain atrophy patterns based on MRI for the prediction of Alzheimer's disease," *NeuroImage*, vol. 50, no. 1, pp. 162–174, 2010.

[19] A. Cherubini, M. Morelli, R. Nisticó et al., "Magnetic resonance support vector machine discriminates between Parkinson disease and progressive supranuclear palsy," *Movement Disorders*, vol. 29, no. 2, pp. 266–269, 2014.

[20] C. Salvatore, A. Cerasa, I. Castiglioni et al., "Machine learning on brain MRI data for differential diagnosis of Parkinson's disease and progressive supranuclear palsy," *Journal of Neuroscience Methods*, vol. 222, pp. 230–237, 2014.

[21] U. Castellani, E. Rossato, V. Murino et al., "Classification of schizophrenia using feature-based morphometry," *Journal of Neural Transmission*, vol. 119, no. 3, pp. 395–404, 2012.

[22] J. R. C. Almeida, J. Mourao-Miranda, H. J. Aizenstein et al., "Pattern recognition analysis of anterior cingulated cortex blood flow to classify depression polarity," *British Journal of Psychiatry*, vol. 203, no. 4, pp. 310–311, 2013.

[23] W. Pettersson-Yeo, S. Benetti, A. F. Marquand et al., "Using genetic, cognitive and multi-modal neuroimaging data to identify ultra-high-risk and first-episode psychosis at the individual level," *Psychological Medicine*, vol. 43, no. 12, pp. 2547–2562, 2013.

[24] C. Habeck, N. L. Foster, R. Perneczky et al., "Multivariate and univariate neuroimaging biomarkers of Alzheimer's disease," *NeuroImage*, vol. 40, no. 4, pp. 1503–1515, 2008.

[25] B. Scholkopf and A. J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*, MIT Press, Cambridge, Mass, USA, 1st edition, 2001.

[26] S. R. Hathaway and J. C. McKinley, *The Minnesota Multiphasic Personality Inventory*, University of Minnesota Press, Minneapolis, Minn, USA, 1943.

[27] Y. Taki, S. Kinomura, K. Sato et al., "Relationship between body mass index and gray matter volume in 1,428 healthy individuals," *Obesity*, vol. 16, no. 1, pp. 119–124, 2008.

[28] D. M. Garner, *Eating Disorder Inventory-2 Professional Manual*, Psychological Assessment Resources, Odessa, Fla, USA, 1991.

[29] E. R. S. Nijenhuis, O. Van der Hart, and K. Kruger, "The psychometric characteristics of the traumatic experiences checklist (TEC): first findings among psychiatric outpatients," *Clinical Psychology & Psychotherapy*, vol. 9, no. 3, pp. 200–210, 2002.

[30] E. M. Bernstein and F. W. Putnam, "Development, reliability, and validity of a dissociation scale," *The Journal of Nervous and Mental Disease*, vol. 174, no. 12, pp. 727–735, 1986.

[31] E. R. S. Nijenhuis, P. Spinhoven, R. Van Dyck, O. Van der Hart, and J. Vanderlinden, "The development and psychometric characteristics of the Somatoform Dissociation Questionnaire (SDQ-20)," *The Journal of Nervous and Mental Disease*, vol. 184, no. 11, pp. 688–694, 1996.

[32] G. Parker, H. Tupling, and L. B. Brown, "A parental bonding instrument," *British Journal of Medical Psychology*, vol. 52, no. 1, pp. 1–10, 1979.

[33] D. M. Garner and P. E. Garfinkel, "The eating attitudes test: an index of the symptoms of anorexia nervosa," *Psychological Medicine*, vol. 9, no. 2, pp. 273–279, 1979.

[34] C. Segura-García, M. C. Papaianni, P. Rizza, S. Flora, and P. De Fazio, "The development and validation of the Body Image Dimensional Assessment (BIDA)," *Eating and Weight Disorders*, vol. 17, no. 3, pp. e219–e225, 2012.

[35] C. Gaser, H.-P. Volz, S. Kiebel, S. Riehemann, and H. Sauer, "Detecting structural changes in whole brain based on nonlinear deformations—application to schizophrenia research," *NeuroImage*, vol. 10, no. 2, pp. 107–113, 1999.

[36] J. Ashburner and K. J. Friston, "Voxel-based morphometry—the methods," *NeuroImage*, vol. 11, no. 6 I, pp. 805–821, 2000.

[37] D. Salas-Gonzalez, J. M. Górriz, J. Ramírez et al., "Feature selection using factor analysis for Alzheimer's diagnosis using ^{18}F-FDG PET images," *Medical Physics*, vol. 37, no. 11, pp. 6084–6095, 2010.

[38] I. Alvarez, J. M. Gorriz, J. Ramirez et al., "Alzheimer's diagnosis using eigenbrains and support vector machines," *Electronic Letters*, vol. 45, pp. 342–343, 2009.

[39] V. N. Vapnik, "An overview of statistical learning theory," *IEEE Transactions on Neural Networks*, vol. 10, no. 5, pp. 988–999, 1999.

[40] S. Haufe, F. Meinecke, K. Görgen et al., "On the interpretation of weight vectors of linear models in multivariate neuroimaging," *NeuroImage*, vol. 87, pp. 96–110, 2014.

[41] H.-C. Friederich, M. Wu, J. J. Simon, and W. Herzog, "Neurocircuit function in eating disorders," *International Journal of Eating Disorders*, vol. 46, no. 5, pp. 425–432, 2013.

[42] H. Steiger, J. Richardson, N. Schmitz, M. Israel, K. R. Bruce, and L. Gauvin, "Trait-defined eating-disorder subtypes and history of childhood abuse," *The International Journal of Eating Disorders*, vol. 43, no. 5, pp. 428–432, 2010.

[43] S. Borgwardt and P. Fusar-Poli, "Third-generation neuroimaging in early schizophrenia: translating research evidence into clinical utility," *The British Journal of Psychiatry*, vol. 200, no. 4, pp. 270–272, 2012.

[44] P. Mahler, J.-M. Guastavino, G. Jacquart, and C. Strazielle, "An unexpected role of the cerebellum: involvement in nutritional organization," *Physiology & Behavior*, vol. 54, no. 6, pp. 1063–1067, 1993.

[45] J.-N. Zhu and J.-J. Wang, "The cerebellum in feeding control: possible function and mechanism," *Cellular and Molecular Neurobiology*, vol. 28, no. 4, pp. 469–478, 2008.

[46] F. A. Middleton and P. L. Strick, "Cerebellar projections to the prefrontal cortex of the primate," *The Journal of Neuroscience*, vol. 21, no. 2, pp. 700–712, 2001.

[47] A. Boghi, S. Sterpone, S. Sales et al., "In vivo evidence of global and focal brain alterations in anorexia nervosa," *Psychiatry Research*, vol. 192, no. 3, pp. 154–159, 2011.

[48] B. Suchan, M. Busch, D. Schulte, D. Grönermeyer, S. Herpertz, and S. Vocks, "Reduction of gray matter density in the extrastriate body area in women with anorexia nervosa," *Behavioural Brain Research*, vol. 206, no. 1, pp. 63–67, 2010.

[49] M. M. Husain, K. J. Black, P. M. Doraiswamy et al., "Subcortical brain anatomy in anorexia and bulimia," *Biological Psychiatry*, vol. 31, no. 7, pp. 735–738, 1992.

[50] N. M. Avena and M. E. Bocarsly, "Dysregulation of brain reward systems in eating disorders: neurochemical information from animal models of binge eating, bulimia nervosa, and anorexia nervosa," *Neuropharmacology*, vol. 63, no. 1, pp. 87–96, 2012.

[51] C. Keating, A. J. Tilbrook, S. L. Rossell, P. G. Enticott, and P. B. Fitzgerald, "Reward processing in anorexia nervosa," *Neuropsychologia*, vol. 50, no. 5, pp. 567–575, 2012.

[52] S. J. Brooks, O. G. O'Daly, R. Uher et al., "Differential neural responses to food images in women with bulimia versus anorexia nervosa," *PLoS ONE*, vol. 6, no. 7, Article ID e22259, 2011.

[53] B. Suchan, D. S. Bauser, M. Busch et al., "Reduced connectivity between the left fusiform body area and the extrastriate body area in anorexia nervosa is associated with body image distortion," *Behavioural Brain Research*, vol. 241, no. 1, pp. 80–85, 2013.

[54] G. Berlucchi and S. M. Aglioti, "The body in the brain revisited," *Experimental Brain Research*, vol. 200, no. 1, pp. 25–35, 2010.

[55] A. Favaro, P. Santonastaso, R. Manara et al., "Disruption of visuospatial and somatosensory functional connectivity in anorexia nervosa," *Biological Psychiatry*, vol. 72, no. 10, pp. 864–870, 2012.

[56] C. G. Fairburn and K. Bohn, "Eating disorder NOS (EDNOS): an example of the troublesome 'not otherwise specified' (NOS) category in DSM-IV," *Behaviour Research and Therapy*, vol. 43, no. 6, pp. 691–701, 2005.

[57] T. Hildebrandt, T. Bacow, M. Markella, and K. L. Loeb, "Anxiety in anorexia nervosa and its management using family-based treatment," *European Eating Disorders Review*, vol. 20, no. 1, pp. e1–e16, 2012.

[58] B. C. Vande Berg, J. Malghem, O. Devuyst, B. E. Maldague, and M. J. Lambert, "Anorexia nervosa: correlation between MR appearance of bone marrow and severity of disease," *Radiology*, vol. 193, no. 3, pp. 859–864, 1994.

PUBLICATION V

Application of the implemented ML method to the diagnosis of ASD (Other applications)

J Autism Dev Disord (2015) 45:2146–2156
DOI 10.1007/s10803-015-2379-8

ORIGINAL PAPER

Use of Machine Learning to Identify Children with Autism and Their Motor Abnormalities

Alessandro Crippa · Christian Salvatore · Paolo Perego · Sara Forti · Maria Nobile · Massimo Molteni · Isabella Castiglioni

Published online: 5 February 2015

Abstract In the present work, we have undertaken a proof-of-concept study to determine whether a simple upper-limb movement could be useful to accurately classify low-functioning children with autism spectrum disorder (ASD) aged 2–4. To answer this question, we developed a supervised machine-learning method to correctly discriminate 15 preschool children with ASD from 15 typically developing children by means of kinematic analysis of a simple reach-to-drop task. Our method reached a maximum classification accuracy of 96.7 % with seven features related to the goal-oriented part of the movement. These preliminary findings offer insight into a possible motor signature of ASD that may be potentially useful in identifying a well-defined subset of patients, reducing the clinical heterogeneity within the broad behavioral phenotype.

A. Crippa (✉) · S. Forti · M. Nobile · M. Molteni
Child Psychopathology Unit, Scientific Institute, IRCCS Eugenio Medea, Via Don Luigi Monza 20, 23842 Bosisio Parini, Lecco, Italy
e-mail: alessandro.crippa@bp.lnf.it

A. Crippa · C. Salvatore · I. Castiglioni
Institute of Molecular Imaging and Physiology, National Research Council, Via F.lli Cervi 93, 20090 Segrate, Milan, Italy

P. Perego
Bioengineering Lab, Scientific Institute, IRCCS Eugenio Medea, Via Don Luigi Monza 20, 23842 Bosisio Parini, Lecco, Italy

M. Nobile
Department of Clinical Neurosciences, Hermanas Hospitalarias, FoRiPsi, Albese con Cassano, Italy

Keywords Autism spectrum disorder · Kinematics · Classification · Machine learning · Support vector machines

Introduction

Autism spectrum disorder (ASD) is a highly heterogeneous neurodevelopmental disorder with multiple causes, courses, and a wide range in symptom severity (Amaral et al. 2008). Although the core features of ASD are persistent deficits in social communication and interaction and the presence of restricted, repetitive patterns of behavior, interests, or activities (DSM V, American Psychiatric Association 2013), it is of great importance not to ignore the motor impairments associated with ASD as they are highly prevalent, at 79 %, and can have a significant impact on quality of life and social development (Lai et al. 2014). Motor abnormalities in ASD may occur very early in development (Teitelbaum et al. 1998, Brian et al. 2008) and be apparent over time (Fournier et al. 2010; Van Waelvelde et al. 2010) being a pervasive feature of the disorder. Recent studies have also provided evidence for the specificity of motor impairments identified in high-functioning children with ASD compared to children with attention deficit/hyperactivity (ADHD) (Izawa et al. 2012; Ament et al. 2014) and to typically developing children matched by nonverbal IQ and receptive language (Whyatt and Craig 2013). Overall, these findings suggest that motor abnormalities could be a consistent marker of ASD (Dowd et al. 2012). A number of different motor deficits have been reported in ASD, including anomalies in walking patterns (e.g., Rinehart and McGinley 2010; Nobile et al. 2011), hand movements such as reaching (e.g., Mari et al. 2003; Glazebrook et al. 2006; Forti et al. 2011), and eye-hand

coordination (e.g., Glazebrook et al. 2009; Crippa et al. 2013). The severity of motor deficits correlates with the degree of social withdrawal and the severity of symptoms (Freitag et al. 2007). Motor control has even been speculated to be crucial for communication and social interaction (Leary and Hill 1996). Indeed, Minshew et al. (2004) proposed that studies on motor function could have significant potential in elucidating the neurobiological basis and even improving the diagnostic definition of ASD.

Currently, the gold standard for the diagnosis of ASD has been formalized with the clinical judgment of symptoms and with semistructured, play-based behavioral observations (Lord et al. 2000) and standardized interviews or questionnaires (e.g., Lord et al. 1994). However, recent studies have started to explore the predictive value of neurobiological as well as behavioral measures in ASD in order to identify a well-defined phenotype of individuals and—possibly—to enable a computer-aided diagnosis perspective. These studies typically implement pattern classification methods that are based on machine-learning algorithms to predict or classify individuals of different groups by maximizing the distance between groups of datasets. Machine learning commonly refers to all procedures that train a computer algorithm to identify a complex pattern of data (i.e., "features") that can then be used to predict group membership of new subjects (e.g., patients vs. controls). Machine-learning techniques based, for example, on support vector machines (SVMs; Vapnik 1995) require a well-characterized dataset in the training phase in order to extract the classification algorithm that best separates the groups (i.e., the "hyperplane" or "decision function"). In the testing phase, the classification algorithm can be used to predict the class membership of a participant not involved in the training procedure (e.g., whether a new child has ASD). Pattern classification methods can also identify complex patterns of anomalies not efficiently recognized by other univariate statistical methods. Thus, the use of pattern recognition methods to predict group membership should not be considered merely in a potentially "diagnostic" perspective but also as a useful tool used to develop objective measures for each individual from a set of sample data. Most of the studies have applied pattern classification methods to neuroanatomical data measured by structural magnetic resonance (MRI; Ecker et al. 2010a, b) or by diffusion tensor imaging (Lange et al. 2010; Ingalhalikar et al. 2011; Deshpande et al. 2013), although Oller et al. (2010) analysis of data regarding automated vocal analysis produced promising results.

In the present work, we have undertaken a proof-of-concept study to determine whether a simple upper-limb movement could be useful to accurately classify low-functioning children with ASD who are between the ages of 2 and 4. In order to answer this question, we developed a supervised machine-learning method to identify preschool children with ASD and correctly discriminate them from typically developing children by means of kinematic analysis of a simple reach, grasp and drop task. We decided to analyze this simple motor task because the motor system can be more easily probed in low-functioning autistic children than systems that underlie complex cognitive functions. In addition to the potential predictive value of our machine-learning method in exploring the clinical relevance of simple upper-limb movement measures in ASD, we could identify a limited set of kinematic characteristics that even suggests the hypothesis of a motor signature of autism.

Methods

Participants

Fifteen preschool-aged children with autism (ASD) were compared to fifteen typically developing (TD) children who were matched by mental age. IQ and mental age were assessed in our institute by using the Griffiths Mental Development Scales (Griffiths 1970) as a part of the routine clinical practice with low-functioning children. A poor score on the Griffiths scales at 1 and/or 2 years has been demonstrated to be a good predictor of impairment at school age (Barnett et al. 2004). All participants had normal or corrected-to-normal vision and were drug-naïve.

The participants in the ASD group were recruited at our institute over an 18-month period. All participants in the clinical group had been previously diagnosed according with the criteria described in the Diagnostic and Statistical Manual of Mental Disorders-IV TR (American Psychiatric Association 2000) by a medical doctor specialized in child neuropsychiatry with expertise in autism. The diagnoses were then confirmed independently by a child psychologist through direct observation and discussion with each child's parents. Seven children had been administered the Autism Diagnostic Observation Schedule (ADOS; Lord et al. 2000). The participants in the control group were recruited by local pediatricians and from kindergartens to be mentally age-matched to the clinical sample from the normally developing population. We decided to include, as a comparison group, typically developing children matched by mental age, following the assumption that mental age usually predict ability to understand task instructions, use appropriate strategies and inhibit inappropriate responses (Jarrold and Brock 2004). The TD children had no previous history of social/communicative disorders, developmental abnormalities, or medical disorders with central nervous system implications. All of the participants' legal guardians

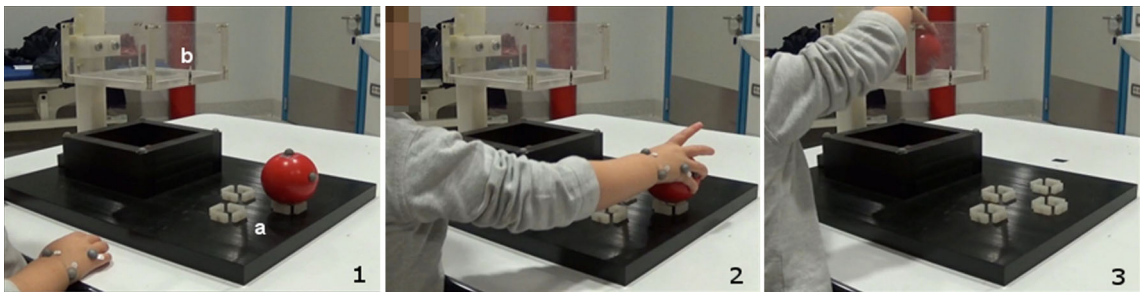


Fig. 1 The experimental task consisted of grasping a rubber ball (**2**) that was placed over a support (see **1**, a); that is, a reach-to-grasp movement before they dropped it in a hole (**3**). The hole (**1**, c) was located inside a see-through *square box* (21 cm high, 20 cm wide) and was large enough not to require fine movements. The goal area is transparent to allow seeing through. 4 markers are placed on the basket under the goal area, 2 on the ball and 3 on each hand (attached to the ulnar and radial surfaces of the participant's wrist and to the hand dorsum on the 4th and 5th metacarpals)

gave their informed written consent prior to the children's participation. The research was approved by the ethics board of our institute in accordance with the Declaration of Helsinki.

Procedure

The participants sat in front of a table of variable height, which was adjusted to the base of the children's trunk. The experimenter sat at the opposite side of the table, and one parent was present in the room. All trials started with the children's hands resting at a set position 20 cm away from the ball support. The experimental task consisted of grasping a rubber ball (6-cm diameter) that was placed over a support (see Fig. 1a); that is, a reach-to-grasp movement before they dropped it in a hole (7-cm diameter). The hole (see Fig. 1b) was located inside a see-through square box (21 cm high, 20 cm wide; see Fig. 1) and was large enough not to require fine movements. Ten trials per participant were conducted: five consecutive trials on the left side (and left hand) and five consecutive trials on the right side (and right hand). The order of trial blocks was counterbalanced between participants. The experimenter performed the task first in order to overtly illustrate the task demand (i.e., reach for the ball, grasp it and drop it in the hole) without any verbal cue. Practice trials, the number of which varied individually, were given to participants before recording in order to verify the children's understanding of the task. The participants were allowed to interrupt the experiment at will in order to rest. The experimental task was simple and interesting enough to ensure the full motivation and compliance of all participants across groups.

Apparatus

An optoelectronic system (The SMART D from BTS Bioengineering® Garbagnate Milanese, Italy) was used to acquire the kinematics data. Three-dimensional kinematic data were collected by eight infrared-motion analysis cameras at 60 Hz (spatial accuracy <0.2 mm), located four per side at 2.5 m from the participants. Passive markers (1 cm) were attached to the ulnar and radial surfaces of the participants' wrists and to the hand dorsum on the fourth and fifth metacarpals (see Fig. 1). Moreover, two markers were placed on the ball and four on the box edges under the goal area. All raw data were first preprocessed with Matlab (Mathworks® Natick, MA, USA); a fifth-order Butterworth, 8-Hz low-pass filter was applied, and movement segmentation and parameters estimation were computed with self-written software.

The overall movement was divided into two sub-movements: *Sub-movement 1*—the movement necessary to reach the ball and place it on its support; *Sub-movement 2*—the movement to transport the ball from its support to the target box hole where the ball was to be dropped. For each of these sub-movements, statistics pertaining to a set of dependent measures was collected: (a) total movement duration (TD), (b) number of movement units[1] (MU), (c) peak velocity (PV), (d) time of PV from sub-movement onset (tPV), (e) peak acceleration (PA), (f) time of PA (tPA), (g) peak deceleration (PD), and (h) time of peak deceleration (tPD). Moreover, final movement accuracy was evaluated by the wrist inclination at the time of the ball drop (delta_WA), calculated as the angle between the palm and the vertical axis of the coordinate system (more precisely, the difference between the WA at the end of the transport phase and the WA at the time of peak deceleration). These 17 kinematic measures were used as input features for the pattern classification procedure.

[1] A movement unit is defined as an acceleration phase followed by a deceleration phase higher than 10 mm/s, starting from the moment at which the increase or decrease in cumulative velocity is over 20 mm/s (Von Hofsten 1991; Thelen et al. 1996).

Data Analysis

After checking that the assumptions were not violated, an analysis of covariance (ANCOVA) was carried out to compare the two groups of children on all kinematic measures with Group (ASD vs. TD) as a between-participant factor, and with IQ and chronological age as between-participant covariates. The alpha level was set to .05 for all data analyses. Effect sizes for ANCOVA are reported using partial eta squared (η_p^2).

The Machine-Learning Method

A pattern classification method based on a machine-learning algorithm was used to classify ASD versus TD by maximizing the distance between the two groups of datasets. A validated supervised machine-learning method (Salvatore et al. 2013) was used. The method involves two different steps: (1) feature selection, the process of selecting a subset of relevant features to be used for classification, and (2) classification, the process of using the selected features to separate the two considered groups of subjects (ASD vs. TD).

Feature Selection

In order to understand which of the collected kinematic features were more discriminative for the ASD versus TD comparison, feature selection was implemented by using a Fisher discriminant ratio (FDR)-based technique (Padilla et al. 2012).

By this technique, for each subject, the collected features and the "label" associated to that subject on a clinical diagnosis basis (i.e., ASD or TD) were considered to calculate a score (FDR score) for each feature.

Specifically, for the feature i, the FDR score was calculated using the following formula:

$$FDR_i = \frac{(\mu_{i-ASD} - \mu_{i-TD})^2}{\sigma^2_{i-ASD} + \sigma^2_{i-TD}}$$

where μ_{i-ASD} and μ_{i-TD} are the mean value of the feature i calculated across the whole ASD and TD datasets, respectively. σ^2_{i-ASD} and σ^2_{i-TD} are the variance of the feature i calculated across the whole ASD and TD datasets, respectively.

Ranked features were then sorted in a decreasing order, from the most to the least discriminative, according to their FDR score.

Classification Algorithm

Classification of ASD and TD subjects was performed using a Support Vector Machine (SVM) approach (Schölkopf et al. 2000; Vapnik 1995, 1998; Vapnik and Chapelle 1999, López et al. 2011), already optimized and validated in a clinical setting (Salvatore et al. 2013).

The aim of the considered SVM is to generate a model able to (1) learn from the selected features of labeled subjects how to discriminate subjects of different groups (binary labeled training datasets), and (2) correctly classify, by means of the same selected features, new unlabeled subjects as belonging to one of the two groups (ASD or TD).

The learning process of the classifier consists of a training phase in which the selected features of the ASD and TD subjects are two training datasets associated to the ASD and TD labels, respectively.

Mathematically, if we have training data consisting of a vector $x_i \in R^N, i = 1, \ldots, N$ and the associated binary label $y_i \in \{\pm 1\}$ (e.g., $+1$ for ASD, -1 for TD), then SVM uses the principle of structural risk minimization to design an optimal hyperplane (OH) that maximizes the distance between the two training groups and that separates them. The lower the distance of a training subject from the OH, the more important that training subject to define the OH. Thus, the distance identifies the "weight" of that training subject in the definition of OH.

The OH can then be used as model to classify new subjects, i.e., subjects for which the label is unknown.

Mathematically, the model used for the identification of the binary label y' of a new subject x, as a result of the classification of that new subject, is given by the following function:

$$y'(x) = \sum_{i=1}^{N} a_i \cdot y_i \cdot k(x, x_i) + b$$

a_i being the weight of the training subject x_i, y_i being the binary label of the training subject i, $k(x, x_i)$ being a linear kernel function, b being a threshold parameter called bias, and N being the number of training subjects. We chose to employ a linear kernel because it represents the more general form of a decision function and because it ensures better computational efficiency.

In this study, the whole machine-learning method was implemented on the Matlab platform (Matlab version R2013b, The MathWorks, Natick, MA). In particular, we used functions of the biolearning toolbox of Matlab to implement the classification algorithm.

Performance of the Classification Algorithm

Performance of the classification algorithm was assessed by using a cross-validation strategy. In general, cross validation involves splitting the original dataset into two complementary subsets: a training set and a testing set. The training set is a set of data associated to a label and used to perform the

training of the classifier (as already described in the previous section); the testing set is a set of data not associated to a label and used to perform the validation of the classifier. By considering different partitions of the data, multiple rounds of cross-validation can then be performed.

In a particular case of cross-validation, called leave-one-out (LOO) cross-validation, the testing set is solely composed of one sample of the original dataset and the training set is made up of the remaining samples of the original dataset ($N - 1$). Therefore, if we want to test all N samples in the original dataset, then it is sufficient that the number of rounds to be performed equals the number N of samples in the original dataset. LOO is a widely used validation approach in literature because it has been proven able to return an almost unbiased estimate of the probability of error (e.g., Vapnik 1998; Chapelle et al. 1999).

In this study, validation of the classifier for the ASD versus TD comparison was performed by using an LOO cross-validation strategy for a number i of selected features running from one to the whole number of features (i.e., 17). A schematic description of the whole procedure is shown in Fig. 2.

In order to quantify the performance of the proposed classification algorithm, the accuracy, specificity, and sensitivity rates were computed. Accuracy of classification measures the rate of correctly classified samples in both positive (ASD) and negative (TD) classes. Specificity and sensitivity measure the rate of correctly classified samples in the positive (ASD) and in the negative (TD) class, respectively.

Mathematically, the accuracy, specificity and sensitivity of the classifier when the first i selected features are used, were computed as follows:

$$Accuracy_i = \frac{N^{CC}}{N}$$

$$Specificity_i = \frac{N_{TD}^{CC}}{N_{TD}^{CC} + N_{TD}^{IC}}$$

$$Sensitivity_i = \frac{N_{ASD}^{CC}}{N_{ASD}^{CC} + N_{ASD}^{IC}}$$

where N is the total number of classified subjects; N^{CC} is the total number of correctly classified (CC) subjects, N_{TD}^{CC} is the number of TD samples that were CC as belonging to the TD gr (true negatives), N_{TD}^{IC} is the number of TD samples that were incorrectly classified (IC) as belonging to the ASD class (false positives); N_{ASD}^{CC} is the number of ASD samples that were CC as belonging to the ASD class (true positives), N_{ASD}^{IC} is the number of ASD samples that were IC as belonging to the TD class (false negatives).

We then studied the dependency of accuracy, specificity, and sensitivity on the number i of selected features.

The maximum values reached for accuracy, specificity, and sensitivity, referred to as maximum accuracy, specificity, and sensitivity, allowed the definition of the most discriminative features.

Overall mean accuracy, specificity, and sensitivity rates were calculated as mean values of accuracy, specificity, and sensitivity as follows:

$$Overall\ mean\ accuracy = \frac{1}{F} \cdot \sum_{i=1}^{F} Accuracy_i$$

$$Overall\ mean\ specificity = \frac{1}{F} \cdot \sum_{i=1}^{F} Specificity_i$$

$$Overall\ mean\ sensitivity = \frac{1}{F} \cdot \sum_{i=1}^{F} Sensitivity_i$$

where F is the whole number of features (17).

Results

Data on the demographic, cognitive, and clinical characteristics of the participants are summarized in Table 1.

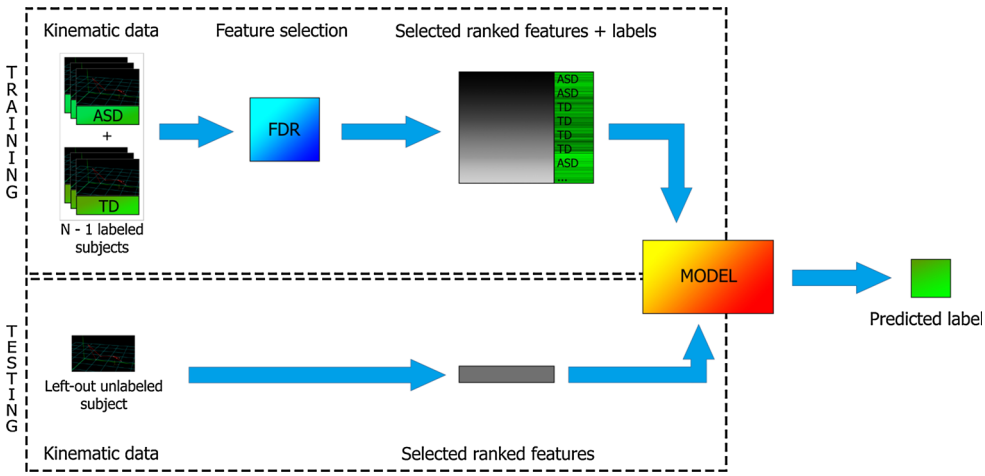


Fig. 2 Flowchart of preprocessing, support vector regression and leave-one-subject-out procedures

Table 1 Demographics of the participants

	ASD	TD	t (1, 28)	*p*
N	15	15		
Females:males	3:12	2:13		
Chronological age[a]	3.5 ± 7.7 (2.8–4.6)	2.6 ± 5.2 (1.7–2.9)	−4.55	<.001
Mental age[a]	2.6 ± 5.7 (1.7–3.4)	2.7 ± 5.9 (1.6–3.2)	.513	n.s.
IQ[b]	75 ± 13.4 (51–96)	105 ± 12.7 (81–119)	6.52	<.001
ADOS[c]				
Social	11 ± 2.2	–		
Communication	7 ± 1.5	–		
SBRI[d]	2 ± 1.6	–		

ASD autism group, *TD* typically developing group; IQ and mental age were assessed using the Griffiths Mental Development Scales (Griffiths 1970)

[a] Mean years; months ± standard deviation (range)

[b] Mean ± standard deviation (range)

[c] ADOS autism diagnostic observation schedule, Lord et al. (2000)

[d] Stereotyped Behavior and Restricted Interests scale

The validity of mental age matching was confirmed ($p > 0.05$). Gender was also balanced between groups, as there were 3 girls in the ASD group and 2 girls in the healthy control group ($\chi^2(1) = .240$; $p > 0.05$). As expected, IQ and chronological age were not balanced across groups (both $p < 0.001$). Table 2 shows kinematic feature values of the two groups of children included in the study (ASD vs. TD) and the results of ANCOVA calculated on all kinematic measures. We found several significant group differences based on the kinematic variables even after having controlled for between-participant differences in IQ and chronological age.

The Machine-Learning Method

Classification Algorithm

In Fig. 3, the optimal hyper-plane separating ASD from TD participants is shown as a representative example of the training phase of the machine-learning method.

Performance of the Classification Algorithm

In Table 3, the accuracy, specificity, and sensitivity of the machine-learning method for the comparison of ASD versus TD are reported.

The machine-learning method was able to successfully classify participants by diagnosis. The classification accuracy reached a maximum accuracy of 96.7 % (specificity 93.8 % and sensitivity 100 %) by using seven features selected by the Fisher discriminant ratio-based technique. Overall mean accuracy, specificity, and sensitivity rates were also calculated over a number of selected features ranging from one to 17 (the whole number of features). The overall mean classification accuracy (specificity/sensitivity) was 84.9 % (mean specificity 89.1 % and mean sensitivity 82.2 %).

Table 2 Kinematic data were initially analyzed through an ANCOVA with Group (ASD vs. TD) as a between-participant factor, and with IQ and chronological age as covariates

		ASD	TD	F(1, 26)	Sig.	η_p^2
Submovement 1						
Movement units	M (SD)	1.91 (0.62)	1.70 (0.37)	<1.0	n.s.	.012
Total movement duration	M (SD)	0.69 (0.14)	0.66 (0.12)	<1.0	n.s.	.010
Peak velocity	M (SD)	0.46 (0.12)	0.59 (0.17)	5,626	**<0.05**	.178
Time of peak velocity	M (SD)	0.34 (0.07)	0.31 (0.04)	<1.0	n.s.	.036
Peak acceleration	M (SD)	3.18 (0.93)	4.26 (1.52)	7,884	**<0.01**	.233
Time of peak acceleration	M (SD)	0.21 (0.07)	0.16 (0.05)	<1.0	n.s.	.031
Peak deceleration	M (SD)	−3.59 (1.28)	−3.93 (1.44)	<1.0	n.s.	.067
Time of peak deceleration	M (SD)	0.47 (0.08)	0.44 (0.06)	<1.0	n.s.	.017
Submovement 2						
Movement units	M (SD)	3.45 (1.78)	1.76 (0.39)	4,408	**<0.05**	.145
Total movement duration	M (SD)	1.35 (0.44)	0.79 (0.15)	13,832	**=0.001**	.347
Peak velocity	M (SD)	0.61 (0.15)	0.76 (0.16)	13,475	**=0.001**	.341
Time of peak velocity	M (SD)	0.41 (0.14)	0.31 (0.05)	18.501	**<0.001**	.416
Peak acceleration	M (SD)	3.85 (1.13)	5.58 (1.94)	12,416	**<0.01**	.323
Time of peak acceleration	M (SD)	0.23 (0.20)	0.13 (0.04)	6,303	**<0.05**	.195
Pick deceleration	M (SD)	−3.29 (1.15)	−4.27 (1.88)	2,632	n.s.	.092
Time of peak deceleration	M (SD)	0.75 (0.24)	0.51 (0.11)	26,652	**<0.001**	.506
Wrist angle	M (SD)	−4.25 (16.34)	−25 (12.40)	6,604	**<0.05**	.203

Bold value indicates significant contrasts

The alpha level was set to .05 for all data analyses. Table depicts group means and standard deviations for kinematic variables, values of *F* test, *p* values and effect sizes reported using partial eta squared (η_p^2)

ASD autism group, *TD* typically developing group

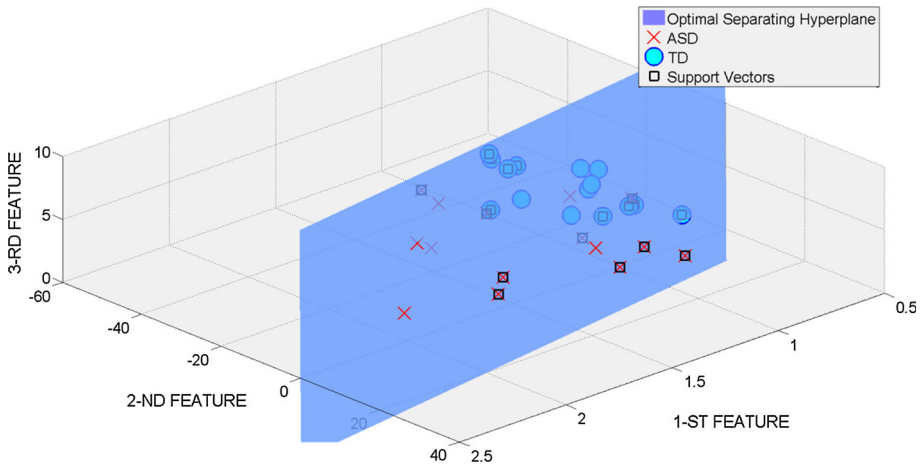


Fig. 3 Optimal separating hyper-plane for the autism group (ASD) versus typically developing groups (TD) (1st, 2nd and 3rd components) is shown as a representative example of the training phase of the machine-learning method

Table 3 Accuracy, specificity and densitivity rates of SVM using LOO validation

	Maximum accuracy (%) (# selected features) Overall mean accuracy (%)	Maximum specificity (%) (# selected features) Overall mean specificity (%)	Maximum sensitivity (%) (# selected features) Overall mean sensitivity (%)
ASD versus TD	96.7 (7) 84.9	93.8 (7) 89.1	100.0 (7) 82.2

ASD autism group, *TD* typically developing group. The maximum values reached by accuracy, specificity and sensitivity were referred to as maximum accuracy, specificity and sensitivity rates. Accuracy, specificity and sensitivity reached their maximum values using 7 features, all related to the second part of the movement—Sub movement 2: (1) Total Duration; (2) delta Wrist Angle; (3) number of Movement Units; (4) time of Peak Deceleration; (5) Peak Acceleration; (6) time of Peak Velocity; (7) Peak Velocity

In Fig. 4, the dependence of the metrics on the number of considered features is shown. The resulting data are shown for a number of features ranging from one to 17. As expected, accuracy, specificity, and sensitivity rates increase with the number of selected features, reaching their maximum values when considering seven selected features.

Besides calculating the accuracy of the SVM method, we were particularly interested in identifying which kinematic features contributed toward the classification. Our analysis showed that seven of 17 features were sufficient to classify autism with a 96.7 % accuracy rate. All of these seven kinematic features are related to the second part of the movement, *sub-movement 2* (i.e., the movement to transport the ball from a support to the target hole in which the ball was to be dropped): (1) total duration; (2) delta wrist angle; (3) number of movement units; (4) time of peak deceleration; (5) peak acceleration; (6) time of peak velocity; and (7) peak velocity. Finally, the most discriminative features between the two groups when considering all of the N rounds (30) of the LOO cross-validation strategy are reported here in descending order: Total Duration sub movement 2, Delta Wrist Angle, Movement Units sub movement 2, time of Peak Deceleration sub movement 2, Peak Acceleration sub movement 2, time of Peak Velocity sub movement 2, Peak Velocity sub movement 2, Peak Velocity sub movement 1, time of Peak Acceleration sub movement 1, Peak Acceleration sub movement 1, time of Peak Acceleration sub movement 2, Peak Deceleration sub movement 2, time of Peak Velocity sub movement 1, Movement Units sub movement 1, time of Peak Deceleration sub movement 1, Peak Deceleration sub movement 1, Total Duration sub movement 1.

Discussion

Autism spectrum disorder is currently diagnosed on the basis of symptoms as qualitatively judged by clinicians and by means of semistructured observations (ADOS) and standardized interviews or questionnaires (ADI-R). Given this gold standard for the diagnosis of ASD, the use of pattern recognition methods to predict group membership has recently attracted strong attention, not only from a computer-aided diagnosis perspective, but also as suitable tool to define objective, quantitative measures of the disorder. Previous works have investigated the predictive value of neurobiological and behavioral measures in patients with ASD. The purpose of the present study was to explore the ability of the kinematic analysis of a simple upper-limb movement to correctly discriminate young low-functioning children with ASD from typically developing

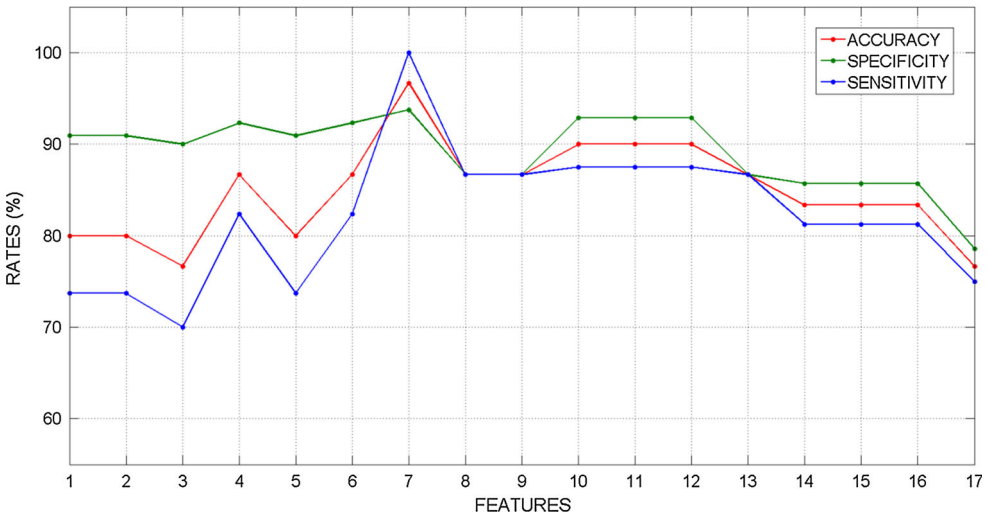


Fig. 4 Graph showing classification accuracy, specificity and sensitivity rates (%) of SVM (Y-axis) in relation of the number of considered features (X-axis). As expected, accuracy, specificity and sensitivity rates increased with the number of selected features. The classification accuracy reached a maximum accuracy of 96.7 % (specificity 93.8 %, and sensitivity 100 %) utilizing seven features. All of these seven kinematic features are related to the second part of the movement—sub-movement 2—the movement to transport the ball from a support to the target hole in which the ball was to be dropped. Such suggests that goal-oriented movements may be critical in separating children with ASD from typically developing children

children. To achieve this goal, we applied our validated supervised machine-learning procedure (Salvatore et al. 2013) to the kinematic analysis of a simple reach, grasp, and drop task performed by preschool children with ASD in comparison to their mental-age-matched, typically developing peers.

The SVM algorithm reached a good mean individual classification in the comparisons between children with ASD and healthy controls (overall mean accuracy = 84.9 %, with overall mean specificity = 89.1 % and overall mean sensitivity = 82.2 %), with a maximum accuracy of 96.7 % (with maximum specificity of 93.8 % and maximum sensitivity of 100 %). The classification accuracy that was achieved in this study is consistent with previous SVM applications to MRI data (Ecker et al. 2010a, b) and to diffusion tensor imaging (DTI) data (Ingalhalikar et al. 2011; Deshpande et al. 2013) or with quadratic discriminant function application on diffusion tensor asymmetries (Lange et al. 2010). Our results are also consistent with the findings of Oller et al. (2010), who derived algorithms that were based on linear discriminant analysis by using an automated analysis of the acoustic characteristics of babble and early language to discriminate typical from language disordered development, such as autism or language delay. Thus, the present findings clearly show the feasibility and the applicability of our SVM method in correctly classifying preschool children with ASD on the basis of a motor task. Indeed, an autism diagnosis is particularly difficult in young, low-functioning children with autism, even using the gold standard diagnostic procedure. Our motor measure might have potential clinical application in such cases, thus providing useful information for clinicians to support a diagnostic decision. A point of relevance of our work, in fact, is that we decided to study the predictive value of a simple reach, grasp, and drop task, because the motor system can be more easily evaluated (i.e., even in young low-functioning children with ASD) than other more complex systems (e.g., cognitive functions). Indeed, because of the easiness and self-explanatory nature of the task, all participants were able to fully understand the experimental demand and to complete the movement successfully. Furthermore, kinematics analysis provides a constraint-free, non-intrusive environment for a challenging clinical population such as ASD in comparison with a magnetic resonance examination that is mostly used in previous pattern-recognition applications. Lastly, kinematic analysis is also a more convenient and less expensive technology than MRI to implement in a clinical setting equipped with an opto-electronic system to acquire kinematic data. Indeed, the task can be easily administered by any professional who works with children. Testing sessions last 15 min, and data analysis can be performed by a trained bioengineer in approximately 30 min for each subject.

Using feature selection, we also found the best classification accuracy of 96.7 % with seven features which had the highest discriminative ability between the groups. All of these seven kinematic features are related to the second part of the movement—*sub-movement 2*—in which the child transported the ball from a support to the target hole

where the ball was to be dropped. This suggests that goal-oriented movements may be critical in separating children with ASD from typically developing children. More specifically, the top three features within the seven kinematic characteristics of *sub-movement 2*—time duration, movement units, and wrist angle—indicate respectively slower and more fragmented movements in children with ASD with inappropriate hand inclination for ball-drops during the final phases of hand transport. Thus, our results extend previous investigations in ASD that report the difficulty of translating intention into a motor chain leading to the action goal (Cattaneo et al. 2007; Fabbri-Destro et al. 2009; Forti et al. 2011). These findings demonstrate that a limited set of kinematic characteristics could reliably identify children with ASD in order to describe a well-defined phenotype of individuals within a complex and highly heterogeneous disorder, even suggesting a possible motor signature of autism related to disrupted planning movement sequences.

Despite our promising results, some methodological limitations of the present exploratory study should be considered. The main limitation is related to the small sample sizes of participant groups; the present findings, therefore, need to be replicated in a larger sample in order to validate the present SVM method by using a data set upon which it has not trained. Another potential limitation of this study is that our SVM classification is highly specific to the sample employed in training the classifier (i.e., preschool children with ASD). Future studies involving females with ASD, children with high-functioning autism, and adult patients are needed to generalize our findings to the heterogeneous spectrum of the disorder. Although we found that our significant between-groups differences were not dependent on IQ and chronological age, it could be worthwhile in future studies to train the computer algorithm with data from age-matched typically developing participants as well. Unfortunately, we did not collect ADOS scores from the entire clinical sample; thus, we could not perform a correlation analysis between our significant findings and the clinical characteristics of children with ASD. Future extensions of this work should also include other neurodevelopmental conditions (e.g., intellectual disability, developmental delays without intellectual disability, or developmental coordination disorders) in order to verify the classifier specificity to ASD, rather than a neurodevelopmental disorder in general. Indeed, some studies have recently indicated the specificity of motor difficulties in older high-functioning children with ASD compared to children with ADHD (Izawa et al. 2012; Ament et al. 2014) and to healthy children matched by nonverbal IQ and receptive language (Whyatt and Craig 2013). Finally, it should be noted that the predictive values of classification methods are restrained by the base rate of neurodevelopmental disorder in the population (Bishop 2010; Heneghan 2010; Yerys and Pennington 2011). Therefore, caution is needed when comparing classification-based accuracy values to the conventional diagnostic measures.

Nevertheless, although the present results should be considered preliminary, this study represents a "proof-of-concept" that kinematic analysis of simple upper-limb movement can reliably identify preschool-aged, low-functioning children with ASD. The significant predictive value of our SVM classification approach might be valuable to support the clinical practice of diagnosing ASD, thus encouraging a computer-aided diagnosis perspective. Moreover, our findings offer insight on a possible motor signature of autism that is potentially useful to identify a well-defined subset of patients, thus reducing the clinical heterogeneity within the broad behavioral phenotype. This may guide further exploration of neuropathology of the disorder with neuroimaging techniques or genetic analysis.

Acknowledgments This research has been partially funded by the FP6-NEST Adventure activities Specific Targeted Research Project: ''TACT'' (Thought in ACTion) to Dr. Molteni and by the Fund for Research of the Italian Ministry of University and Research, within a framework agreement between Lombardy Region and National Research Council of Italy (No. 17125, 27/09/2012). The funding sources had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. We acknowledge the work of Silvia Borini, Cristina Motta, Elisa Mani, and Laura Villa in the diagnostic evaluation of participants with autism and Giuseppe Aceti, Maura Mariani, Claudio Marcolini, Mariangela Perego, Barbara Urbani, and Angela Valli for their help in recruiting participants. We also thank Silvia Colonna and Maddalena Mauri for helping editing the last version of manuscript and the anonymous reviewers for their comments. Lastly, we are especially grateful to all the families of the children who took part in this study.

References

Amaral, D. G., Schumann, C. M., & Nordahl, C. W. (2008). Neuroanatomy of autism. *Trends of Neurosciences, 31*(3), 137–145.

Ament, K., Mejia, A., Buhlman, R., Erklin, S., Caffo, B., Mostofsky, S., et al. (2014). Evidence for specificity of motor impairments in catching and balance in children with autism. *Journal of Autism and Developmental Disorders*. doi:10.1007/s10803-014-2229-0.

American Psychiatric Association. (2000). *Diagnostic and statisticalmanual of mental disorders* (4th ed., text rev.). Washington, DC: American Psychiatric Association.

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Arlington, VA: American Psychiatric Association.

Barnett, A. L., Guzzetta, A., Mercuri, E., Henderson, S. E., Haataja, L., Cowan, F., & Dubowitz, L. (2004). Can the Griffiths scales predict neuromotor and perceptual-motor impairment in term infants with neonatal encephalopathy? *Archives of Disease in Childhood, 89*, 637–643.

Bishop, D. V. M. (2010). The difference between Po0.05 and a screening test. http://deevybee.blogspot.com/2010/07/difference-between-p-05-and-screening.html. Accessed June 30, 2014.

Brian, J., Bryson, S. E., Garon, N., Roberts, W., Smith, I. M., Szatmari, P., & Zwaigenbaum, L. (2008). Clinical assessment of autism in high-risk 18-month-olds. *Autism, 12*(5), 433–456.

Cattaneo, L., Fabbri-Destro, M., Boria, S., Pieraccini, C., Monti, A., Cossu, G., & Rizzolatti, G. (2007). Impairment of actions chains in autism and its possible role in intention understanding. *Proceedings of National Academy of Science of United States of America, 104*(45), 17825–17830.

Chapelle, O., Haffner, P., & Vapnik, V. N. (1999). Support vector machines for histogram-based image classification. *IEEE Transactions on Neural Networks, 10*(5), 1055–1064.

Crippa, A., Forti, S., Perego, P., & Molteni, M. (2013). Eye-hand coordination in children with high functioning autism and Asperger's disorder using a gap-overlap paradigm. *Journal of Autism and Developmental Disorders, 43*(4), 841–850.

Deshpande, G., Libero, L. E., Sreenivasan, K. R., Deshpande, H. D., & Kana, R. K. (2013). Identification of neural connectivity signatures of autism using machine learning. *Frontiers in Human Neuroscience, 7*, 670.

Dowd, A. M., McGinley, J. L., Taffe, J. R., & Rinehart, N. J. (2012). Do planning and visual integration difficulties underpin motor dysfunction in autism? A kinematic study of young children with autism. *Journal of Autism and Developmental Disorders, 42*(8), 1539–1548.

Ecker, C., Marquand, A., Mourão-Miranda, J., Johnston, P., Daly, E. M., Brammer, M. J., et al. (2010a). Describing the brain in autism in five dimensions—Magnetic resonance imaging-assisted diagnosis of autism spectrum disorder using a multiparameter classification approach. *Journal of Neuroscience, 30*(32), 10612–10623.

Ecker, C., Rocha-Rego, V., Johnston, P., Mourao-Miranda, J., Marquand, A., Daly, E. M., et al. (2010b). Investigating the predictive value of whole-brain structural MR scans in autism: A pattern classification approach. *Neuroimage, 49*(1), 44–56.

Fabbri-Destro, M., Cattaneo, L., Boria, S., & Rizzolatti, G. (2009). Planning actions in autism. *Experimental Brain Research, 192*(3), 521–525.

Forti, S., Valli, A., Perego, P., Nobile, M., Crippa, A., & Molteni, M. (2011). Motor planning and control in autism. A kinematic analysis of preschool children. *Research in Autism Spectrum Disorders, 5*(2), 834–842.

Fournier, K. A., Hass, C. J., Naik, S. K., Lodha, N., & Cauraugh, J. H. (2010). Motor coordination in autism spectrum disorders: A synthesis and meta-analysis. *Journal of Autism and Developmental Disorder, 40*, 1227–1240.

Freitag, C. M., Kleser, C., Schneider, M., & von Gontard, A. (2007). Quantitative assessment of neuromotor function in adolescents with high functioning autism and Asperger syndrome. *Journal of Autism and Developmental Disorders, 37*(5), 948–959.

Glazebrook, C. M., Elliott, D., & Lyons, J. (2006). A kinematic analysis of how young adults with and without autism plan and control goal-directed movements. *Motor Control, 10*(3), 244–264.

Glazebrook, C., Gonzalez, D., Hansen, S., & Elliott, D. (2009). The role of vision for online control of manual aiming movements in persons with autism spectrum disorders. *Autism, 13*(4), 411–433.

Griffiths, R. (1970). *The ability of young children. A study in mental measurement*. London: University of London Press.

Heneghan, C. (2010). Why autism can't be diagnosed with brain scans: Using brain scans to detect autism would be a huge waste of money, says Carl Heneghan. http://www.guardian.co.uk/science/blog/2010/aug/12/autism-brainscan-statistic. Accessed June 30, 2014.

Ingalhalikar, M., Parker, D., Bloy, L., Roberts, T. P., & Verma, R. (2011). Diffusion based abnormality markers of pathology: Toward learned diagnostic prediction of ASD. *Neuroimage, 57*(3), 918–927.

Izawa, J., Pekny, S. E., Marko, M. K., Haswell, C. C., Shadmehr, R., & Mostofsky, S. H. (2012). Motor learning relies on integrated sensory inputs in ADHD, but over-selectively on proprioception in autism spectrum conditions. *Autism Research, 5*(2), 124–136.

Jarrold, C., & Brock, J. (2004). To match or not to match? Methodological issues in autism-related research. *Journal of Autism and Developmental Disorders, 34*(1), 81–86.

Lai, M. C., Lombardo, M. V., & Baron-Cohen, S. (2014). Autism. *The Lancet, 383*(9920), 896–910.

Lange, N., Dubray, M. B., Lee, J. E., Froimowitz, M. P., Froehlich, A., Adluru, N., et al. (2010). Atypical diffusion tensor hemispheric asymmetry in autism. *Autism Research, 3*(6), 350–358.

Leary, M. R., & Hill, D. A. (1996). Moving on: Autism and movement disturbance. *Mental Retardation, 34*(1), 39–53.

López, M., Ramírez, J., Górriz, J. M., Álvarez, I., Salas-Gonzalez, D., Segovia, F., et al. (2011). Principal component analysis-based techniques and supervised classification schemes for the early detection of Alzheimer's disease. *Neurocomputing, 74*, 1260–1271.

Lord, C., Risi, S., Lambrecht, L., Cook, E. H, Jr, Leventhal, B. L., DiLavore, P. C., et al. (2000). The autism diagnostic observation schedule-generic: A standard measure of social and communication deficits associated with the spectrum of autism. *Journal of Autism and Developmental Disorders, 30*(3), 205–223.

Lord, C., Rutter, M., & Le Couteur, A. (1994). Autism diagnostic interview-revised: A revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. *Journal of Autism and Developmental Disorders, 24*(5), 659–685.

Mari, M., Castiello, U., Marks, D., Marraffa, C., & Prior, M. (2003). The reach-to-grasp movement in children with autism spectrum disorder. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 358*(1430), 393–403.

Minshew, N. J., Sung, K., Jones, B. L., & Furman, J. M. (2004). Underdevelopment of the postural control system in autism. *Neurology, 63*(11), 2056–2061.

Nobile, M., Perego, P., Piccinini, L., Mani, E., Rossi, A., Bellina, M., & Molteni, M. (2011). Further evidence of complex motor dysfunction in drug naive children with autism using automatic motion analysis of gait. *Autism, 15*(3), 263–283.

Oller, D. K., Niyogi, P., Gray, S., Richards, J. A., Gilkerson, J., Xu, D., et al. (2010). Automated vocal analysis of naturalistic recordings from children with autism, language delay, and typical development. *Proceedings of National Academy of Science of United States of America, 107*(30), 13354–13359.

Padilla, P., Lopez, M., Gorriz, J. M., Ramirez, J., Salas-Gonzalez, D., & Alvarez, I. (2012). NMF–SVM based CAD tool applied to functional brain images for the diagnosis of Alzheimer's disease. *IEEE Transactions on Medical Imaging, 31*(2), 207–216.

Rinehart, N., & McGinley, J. (2010). Is motor dysfunction core to autism spectrum disorder? *Developmental Medicine & Child Neurology, 52*(8), 697.

Salvatore, C., Cerasa, A., Castiglioni, I., Gallivanone, F., Augimeri, A., Lopez, M., et al. (2013). Machine learning on brain MRI data for differential diagnosis of Parkinson's disease and progressive supranuclear palsy. *Journal of Neuroscience Methods, 222*, 230–237.

Schölkopf, B., Smola, A. J., Williamson, R. C., & Bartlett, P. L. (2000). New support vector algorithms. *Neural Computation, 12*(5), 1207–1245.

Teitelbaum, P., Teitelbaum, O., Nye, J., Fryman, J., & Maurer, R. G. (1998). Movement analysis in infancy may be useful for early

diagnosis of autism. *Proceedings of the National Academy of Science of the United States of America, 95*, 13982–13987.

Thelen, E., Corbetta, D., & Spencer, J. P. (1996). Development of reaching during the first year: Role of movement speed. *Journal of Experimental Psychology: Human Perception and Performance, 22*(5), 1059–1076.

Van Waelvelde, H., Oostra, A., Dewitte, G., Van Den Broeck, C., & Jongmans, M. J. (2010). Stability of motor problems in young children with or at risk of autism spectrum disorders, ADHD, and or developmental coordination disorder. *Developmental Medicine & Child Neurolology, 52*(8), 174–178.

Vapnik, V. N. (1995). *The nature of statistical learning theory*. New York, NY: Springer.

Vapnik, V. N. (1998). An overview of statistical learning theory. *IEEE Transactions on Neural Networks, 10*(5), 988–999.

Vapnik, V. N., & Chapelle, O. (1999). Bounds on error expectation for support vector machines. *Neural Computation, 12*(9), 2013–2036.

Von Hofsten, C. (1991). Structuring of early reaching movements: A longitudinal study. *Journal of Motor Behavior, 23*(4), 280–292.

Whyatt, C. P., & Craig, C. M. (2013). Sensory-motor problems in autism. *Frontiers in Integrative Neuroscience, 7*(51), 1–12.

Yerys, B. E., & Pennington, B. F. (2011). How do we establish a biological marker for a behaviorally defined disorder? Autism as a test case. *Autism Research, 4*(4), 239–241.

Other Publications

Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2014, Article ID 804723, 13 pages
http://dx.doi.org/10.1155/2014/804723

Research Article

Computerized Neuropsychological Assessment in Aging: Testing Efficacy and Clinical Ecology of Different Interfaces

Matteo Canini,[1] Petronilla Battista,[1] Pasquale Anthony Della Rosa,[1] Eleonora Catricalà,[2] Christian Salvatore,[1] Maria Carla Gilardi,[1] and Isabella Castiglioni[1]

[1] *Institute of Molecular Bioimaging and Physiology, National Research Council (IBFM-CNR), Segrate, 20090 Milan, Italy*
[2] *Laboratory of Neuropsychology, The Foundation of the Carlo Besta Neurological Institute, IRCSS, 20133 Milan, Italy*

Correspondence should be addressed to Pasquale Anthony Della Rosa; pasquale.dellarosa@ibfm.cnr.it

Received 18 April 2014; Revised 2 July 2014; Accepted 2 July 2014; Published 24 July 2014

Academic Editor: Fabio Babiloni

Digital technologies have opened new opportunities for psychological testing, allowing new computerized testing tools to be developed and/or paper and pencil testing tools to be translated to new computerized devices. The question that rises is whether these implementations may introduce some technology-specific effects to be considered in neuropsychological evaluations. Two core aspects have been investigated in this work: the efficacy of tests and the clinical ecology of their administration (the ability to measure real-world test performance), specifically (1) the testing efficacy of a computerized test when response to stimuli is measured using a touch-screen compared to a conventional mouse-control response device; (2) the testing efficacy of a computerized test with respect to different input modalities (visual versus verbal); and (3) the ecology of two computerized assessment modalities (touch-screen and mouse-control), including preference measurements of participants. Our results suggest that (1) touch-screen devices are suitable for administering experimental tasks requiring precise timings for detection, (2) intrinsic nature of neuropsychological tests should always be respected in terms of stimuli presentation when translated to new digitalized environment, and (3) touch-screen devices result in ecological instruments being proposed for the computerized administration of neuropsychological tests with a high level of preference from elderly people.

1. Introduction

Computerized neuropsychological tests have been used in research for almost fifty years [1]. Although many different test batteries have been developed and new batteries are introduced every year for clinical screening, not sufficient normative data and standardized psychometric measures are yet available [2]. Conversely, paper and pencil tests are widely approved and are still regarded as keynote tools for neuropsychological assessment, due to their high validity and reliability [3]. Paper and pencil neuropsychological tests are based on the presence of a neuropsychologist, essential for the assessment of cognitive abilities, especially for the evaluation of a person with brain injury or cognitive impairment and for the selection, administration, and interpretation of tests. Although suffering from some levels of subjectivity, variability, and long times (due to the fact that it is often necessary to do a screening and also a diagnostic deepening), paper and pencil tests have been validated for the administration of reliable tests able to pinpoint a potential deficit involving a specific cognitive ability, or to discriminate among impairments in different cognitive domains [4]. However, neuropsychological evaluation can also provide information concerning normal brain functioning and allows monitoring the cognitive status of an individual, especially throughout older age. Therefore, its results are extremely important to trace a continuum of normal functioning in the aging population, not only in presence of pathologies. Documenting changes in cognition is, indeed, an important issue in neuropsychological assessment, as the clinician/researcher is often called upon to determine if and when cognitive functioning has changed. Another important advantage of the conventional paper and pencil neuropsychological assessment is their ecological validity. In the context of neuropsychological testing, ecological validity refers to the degree to which test performance corresponds to real-world performance. Validity does not

apply to the test itself, but to the inferences that are drawn from the test [5, 6]. Most importantly, clinical sensitivity allowing neuropsychologist to capture potential shades in a specific domain and to trace a specific cognitive profile may results difficult to be translated in a computer-based assessment. Despite these numerous advantages, traditional paper and pencil tools show some limitations particularly when assessing cognitive changes in a relatively short follow-up period. The most commonly administered tests usually do not provide alternative forms of administration [7], thus precluding to repeat testing over short intervals (e.g., <6 months) [8]. Other specific limitations concern the intrinsic nature of the tools and include: assisting the setting and manually computing scoring by the experimenter [9], long duration of the assessment, potential bias related to different examiners [10], impossibility to provide a precise time control on stimuli presentation and/or the lack of an accurate measurement of motor response accuracy [11], and greater costs. Although there are specific tests for the assessment of attentional and executive functions which can evaluate individual components and can diagnose specific deficits [12], some executive/attention abilities could take advantage from the use of a computerized tool, in particular in the assessment of response inhibition, resistance to distraction, planning, problem solving, working memory operations, and mental set shifting divided attention.

All of these limitations could be overcome by a computerized assessment, on condition that efficacies and ecological measurements are carried out. About these issues, the American Psychological Association (APA) has recognized the importance of computerized psychological testing and has suggested how to implement and interpret computerized test results in its guidelines [13]. Furthermore, computerized assessment of cognitive functions can be self-administered and can have a shorter duration (e.g., by reducing "dead" times in stimuli presentation). They may have great validity and reliability due to their great objectivity, precision, and standardization. Computerized performance can also minimize the so called "floor and ceiling effects", occurring when differences among participant performance are not fully captured; thus, they can provide more standardized measures of subject performance, crucial for example, for an accurate and early detection of specific pathological disease (e.g., dementia) [14].

It appears clearly that computerized testing will represent an essential part of the clinical setting in the nearest future, above all, in screening procedures, on condition that these new instruments and their results are governed by experts.

Feasible, efficacious, and ecological computerized testing could allow clear pictures of normal cognition to be measured and monitored also at home, pinpointing specific deficits in each cognitive domain in aging people. Validated computerized tools could also provide stronger grounding to overcome the lack of a consensus regarding the feasibility and testing efficacy related to the different types of technological solutions and settings and/or response layouts chosen for the assessment. For instance, a more ecological technological solution can overcome one important obstacle to the wide use of computerized assessment attributable to the familiarity with technological devices [11], while a more ecological setting in the nearest future could partially allow the administration of the tests without the support of a specialized clinician, or the tests could be potentially self-administered or assessed by a caregiver at home.

The aims of this study were to evaluate the following:

(1) the testing efficacy of a computerized test (in a representative case of executive function assessment) when response to stimuli is measured using a touch screen compared to a conventional mouse-control response device;

(2) the testing efficacy (the efficacy of the neuropsychological test) of a computerized test (in a representative case of memory function assessment) with respect to two different input modalities: a visual presentation modality of the test, replicating the most diffused digital versions of the test, and a verbal presentation modality, replicating the classical clinical administration of the test;

(3) the ecology of both computerized assessment modalities (touch-screen tablet and mouse-control PC), including preference measurements of participants.

2. Related Works

Computerized cognitive batteries are already available and used for the screening and the assessment of dementia. Some of these instruments were appropriately created for the assessment of cognitive decline in dementia; other ones were adapted to fulfill this role in aging. Some tools are designed for research use, and others tools have been designed mainly for clinical use, some of them already implemented in clinical guidelines.

An interesting review [2] reported seventeen computerized test batteries used in the measurement of cognitive abilities of adults. Some of these tools are able to run only on a PC/laptop and others are available only on web sites.

Among these tools:

(1) CNS vital signs [15] battery is developed as a routine clinical screening instrument. It includes seven tests: verbal and visual memory, finger tapping, symbol digit coding, Stroop test, test of shifting attention, and continuous performance test.

(2) CogState [16] battery was developed as a dementia screening instrument and it is implemented as a card game form. The participant plays different games that are adapted accordingly to performance. CogState requires an active internet connection to generate a report. The participant's data are uploaded and analyzed. Then a report is generated and e-mailed back to the provider.

(3) NeuroTrax [17] includes custom software on the local testing computer and serves as a platform for interactive cognitive tests that provide precise accuracy and reaction time data. The level of difficulty is graded. The NeuroTrax tests different cognitive

Table 1: Features of computerized cognitive batteries cited in Section 2. Hardware, input modality, and context of use information are provided.

Battery name	Hardware used	Input modality	Use
CNS vital signs	PC/laptop	Keyboard	Research use
CogState	PC/laptop/tablet (Web based)	Keyboard	Research use
Neurotrax	PC/laptop (web based)	Keyboard/mouse	Clinical use (AACN consensus)
IntegNeuro	PC/laptop	Keyboard	Research use
Touch panel dementia assessment scale	PC/laptop	Touch screen	Research use
CADi	iPad	Touch screen	Research use
CANTAB mobile	iPad	Touch screen	Research use

domains, memory (verbal and nonverbal), executive function, visual spatial skills, verbal fluency, attention, information processing, and motor skills. The tests were designed for use with the older adults. Responses are collected with the mouse or with the number pad on the keyboard. NeuroTrax also generates a report containing raw and standardized scores immediately following testing. Administrative features are web based.

(4) IntegNeuro [18], in brief, investigates the following domains of cognitive function: sensorimotor, verbal and language, memory, executive planning, and attention. Scoring of responses is obtained by using an automated software program. Trained research assistants conducted the hand scoring of some tests and oversight is implemented to monitor accuracy.

(5) Touch panel dementia assessment scale [19] hardware comprises a 14-inch touch panel display and computer devices built into one case. The TDAS runs on Windows OS and was bundled with a custom program made with reference to the ADAS-cog with the elderly under the control of a physician.

More recently, screening tools for assessment of cognitive impairment able to run on an iPad tablet device have made their appearance as well. They can be downloaded and are self-administrable (see Table 1).

CADi [20] and CANTAB mobile [21] are the first two tools which exploited this technology. The CADi consists of 10 very brief tests; the purpose is to provide a mass screening in the Japanese population in a relatively short time and with overall cost substantially lower than paper and pencil based examination. The most important limit of this tool is the cultural background underlying the validation, as it is available only in Japanese language. Notwithstanding, there are 18 different language versions of the CANTAB mobile tool; it comprises focused screening tests able to investigate only episodic memory and learning abilities (PAL task).

3. Materials and Methods

3.1. Participants. A group of 38 healthy participants was recruited among the Italian elderly population (20 males; age range = 53 : 87; age mean = 64.474; standard deviation [SD] = ±8.462; education range = 5 : 19; education mean = 11,263; [SD] = ±4.131).

Participants with any history of neurological illnesses and a Mini mental state examination (MMSE) score lower than 26 were excluded from the study (MMSE range = 26 : 30; MMSE mean = 28,711; [SD] = ±1.183) (see Table 2). Informed consent was obtained from all participants.

3.2. Hardware Device. In order to compare the participant responses from neuropsychological assessment delivered through mouse-control PC and touch-screen tablet, we used an Asus T100T notebook PC (CPU: quad core Intel Atom processor, RAM: 2 GB, Screen: 10.1″ HD, and screen resolution 1366∗768 with multitouch). This device consists of a 10.1 inches tablet running a Windows 8.1 OS which can be used as a standalone touch-screen device or, combined with a mobile dock, as a "standard" PC with conventional input peripherals (mouse and keyboard).

3.3. Software and Scripting. The experimental tasks were implemented and administered using the presentation software (http://www.neurobs.com/). This is an object-oriented programming language allowing a sharp control on stimuli presentation (e.g., objects presentation timing and randomization of trials) and on response tracking. Experimental paradigms are set up through scripts consisting of two logical components: (i) a Scenario Definition Language (SDL) where the objects of the task (i.e., stimuli organized with hierarchical levels of complexity) are defined along with their specific timing and response matching and a (ii) Presentation Control Language (PCL) where objects presentation parameters are controlled (e.g., possibility to present stimuli through loops, conditionals, and subroutines for trials randomization). An additional third component, the SDL header, is used to specify general parameters, that is, those values which will be used as default for the presentation of all stimuli unless a different specification is given in the stimulus definition (e.g., default font size, background color, stimuli duration and timing, logfiles specifications, and so forth).

Presentation tracks and stores information about stimuli administration timing and participant response behaviors (i.e., the response given by the subject) in terms of response

Table 2: Age, education, and MMSE scores are provided for the whole ($N = 38$) group (left column). Independent sample Student's t-tests values for these variables are shown for the AVLT (verbal versus visual) input modality groups (center column) and for the ANT (touch versus mouse) response modality groups.

	Overall group $N = 38$ 20 M/18 F	AVLT groups $N = 38$ visual group (11 M/8 F) verbal group (9 M/10 F) Student's t-test (P values)	ANT groups $N = 38$ touch group (10 M/9 F) mouse group (9 M/10 F) Student's t-test (P values)
Age	mean 64.474 st dev 8.462 range 53 : 87	0.970	0.831
Education	mean 11.263 st dev 4.131 range 5 : 19	0.487	0.970
MMSE	mean 28.91 st dev 0.67 range 28 : 30	0.344	0.272

classification and reaction times. Information is subsequently analyzed using *ad hoc* data processing procedures in order to assess participant performance.

Parameters to be measured for response behaviors of participants are defined by the experimenter and specified in terms of (i) input device/s used by the participant and (ii) response buttons, in GUI's dedicated response setting panel, associated to the selected device to be recalled in the SDL component. Contextually to our experimental setting, two "response layouts" (i.e., mouse-control or touch-screen) were configured in the following manner.

(i) A "mouse response layout" where the docking-station built-in trackpad mouse was specified as input device, and pressing one of the two mouse buttons was specified as response behavior; two distinct classes of responses were coded, each one associated to a specific button press (i.e., pressing the left and right mouse buttons). Each response class was recorded and directly matched with the stimulus target response in order to assess the responses accuracy.

(ii) A "touch response layout" where the capacitive screen surface was specified as input device, and the screen press was specified as response behavior: in this case this was the only available class of response. The two (i.e., left and right) response classes were obtained through data after processing: X and Y coordinates were registered each time the subject gave a response to a target stimulus and the X coordinate was used to estimate whether the response occurred on the left or right portion of the screen; positive X corresponded to right touch screen press, and negative X corresponded to left touch screen press.

3.4. The Neuropsychological Tests

3.4.1. The Attentional Networks Test (ANT). The attentional network test (ANT) [22] is designed to test three different attentional networks, namely, the executive control and the alerting and the orienting components of attention.

Each trial starts with a fixation point and ends with the target stimulus, consisting of an array of five contiguous arrows; the participant is asked to state as fast as possible the direction (i.e., left or right) of the central arrow (see Figure 1).

The array can be defined according to a congruency factor (depending on the direction of the arrows flanking the central, target stimulus) and a spatial factor depending upon whether the array can occur either above or below the fixation point or at the center of the screen; further the presentation of the array can be cued or not by an asterisk (see Figure 1 for a detailed description of all experimental conditions). Different combinations of these experimental factors are selectively used to assess the efficiency of the three different networks (i.e., executive control and alerting and orienting component) through accuracy and reaction times (RT) analysis as follows.

(i) Conflict effect (CE) is assessed by subtracting RTs belonging to congruent trials (flanking arrows pointing in the same direction of the central target arrow) from RTs belonging to incongruent trials (flanking arrows pointing in the opposite direction of the central target arrow).

(ii) Alerting effect (AE) is calculated by subtracting mean RT of the double cued trials (target trial presentation is preceded by two spatial cues occurring contemporarily above and below the fixation point) from the mean RT of the no cue trials (target trial is not preceded by any cue).

(iii) Orienting effect (OE) is calculated by subtracting the mean RT of the central cue condition (target trial

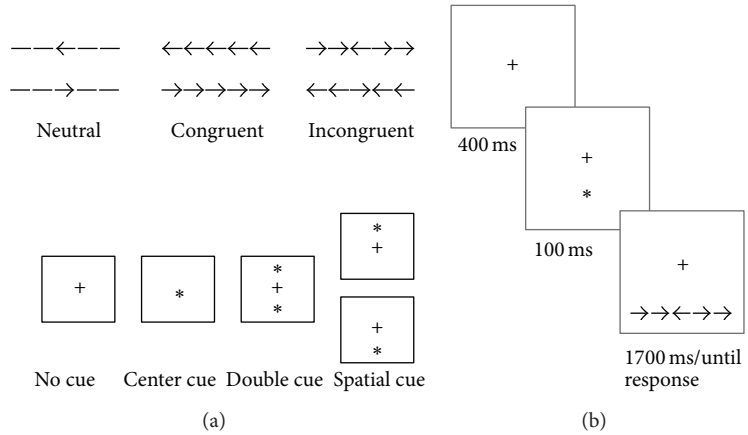


Figure 1: (a) Attentional network test experimental conditions are shown; each trial presented during the task is a combination of congruency (3 levels, on top) and cue (4 levels, below) conditions (adapted from [22]). (b) An example of, spatially cued, incongruent trial is presented; stimuli timings and interstimulus interval of this customized version of the task are provided below each stimulus (adapted from [23]).

is preceded by a cue occurring at the center of the screen) from the mean RT belonging to the spatial cue condition (target trial is preceded by a spatial cue occurring either above or below the fixation point, depending on the position of the incoming target trial).

3.4.2. The Auditory Verbal Learning Test (AVLT). The Rey Auditory Verbal Learning Test [24] assesses the long-term verbal memory or the ability to learn and store in long-term memory unstructured verbal material. It consists of 5 consecutive repetitions for learning the material and then a long delay free recall 15 minutes later. The test consists of 15 unrelated words presented orally. First, the participant has to memorize this list of 15 words and immediately try to recall them after each presentation (immediate recall, IR). Second, after 15 minutes, in the meanwhile the participant is performing nonverbal tests, the participant is asked to recall the same list (delayed recall, DR). Finally, the subject is presented with a longer list of distracter words and asked to recall the previously learned words (recognition task). The Italian test has three alternatives but completely equivalent lists [25].

An interference task not involving verbal or memory cognition is administered between the third (i.e., last) IR session and the DR session.

3.5. Testing Efficacy I: Executive Function Assessment (ANT). Given our purpose to compare the testing efficacy of a computerized test administered by touch-screen tablet versus conventional mouse-control PC, we created two versions of the ANT task: the first one for the *mouse layout response* and the second for the *touch layout response (see Section 3.3 Software and Scripting).*

Both versions were programmed using the presentation software and shared the exact same picture stimuli, stimuli timings, and randomization criteria; they were thus virtually equal, differing only for the input devices layout configuration.

3.5.1. Mouse-Control PC Implementation. Each experimental condition of the ANT task was represented by 8 trials, resulting in a total of 96 [(3 congruency × 4 cue) × 8] trials for each experimental condition.

To this extent an array of 96 trials was specified, each trial consisting of a variable number of component stimuli, depending on the experimental condition.

Each trial was thus specified in terms of (i) stimuli objects, (ii) time intervals occurring between successive stimuli, and (iii) target response (see Figure 1).

The trial presentation order was specified using a fully randomized design. A set of three different interstimulus intervals (ISI) (min value 1873 ms, max value 4964 ms) was also specified; this means that a different time interval could occur between successive trials in order to avoid habituation effects. Two response buttons were assigned, and the participant was required to press as fast as possible the left or right mouse button (depending on the target trial type, as previously described).

Responses accuracy was coded in a logfile by comparing participant's response to a given target stimulus to the expected response for that stimulus, coded in the stimulus parameters in the SDL.

Stimuli presentation and response timing and accuracy were also recorded.

3.5.2. Touch-Screen Tablet Implementation. A touch version of the ANT task was implemented using the same number of trials, experimental design, and randomization order of the ANT task in the *mouse*-control PC implementation. The only difference between the touch layout version and the mouse layout version resides in the response device assigned to the participant. While in the mouse version the participant was asked to respond by pressing the two mouse buttons, in this touch version of the task he/she was required to press the left most or right most part of the screen with the thumbs, handling the touch device with both the hands. Touch location was coded as previously described

(e.g., negative X means a touch occurred in the left portion of the screen and corresponds to the left mouse button press).

3.6. Testing Efficacy II: Memory Function Assessment (AVLT). Given our purpose to assess the efficacy of different modalities (i.e., verbal versus visual implementation) of a computerized test administered by a touch-screen tablet, two versions of the AVLT were created: (1) one verbal version (Verbal AVLT Task), consisting in the presentation of auditory stimuli and requiring the subjects to verbally *recall* answers (which were both recorded and coded by the tester), thus replicating the conventional administration of AVLT used in clinical context, and (2) one visual version (Visual AVLT Task), consisting in the presentation of visual list of words on the screen and requiring the subjects to *recognize* target stimuli by touching the screen when a presented target word appeared. We have chosen not to calibrate the level of difficulty for the recognition test stimuli to be equivalent to that of the recall stimuli in order to detect any differences between the two stimuli presentation modalities, including the level of difficulty.

3.6.1. AVLT Equivalent Lists Creation. Since both lexical and psycholinguistic variables can influence behavioral performance, the characteristics of the original 15 Rey list words were extracted and analyzed in order to create 5 alternatives but completely equivalent lists. Words frequency values for the original 15 words were determined from the Italian lexicon [26], while for the familiarity (FAM), concreteness (CNC), age of acquisition (AoA), and imageability (IMG), values were extracted from the MoA database [27]. For each variable the mean of the distribution of values for the fifteen words and the interquartile range (75–25) was computed. Only the words falling in this range and with values for at least three variables overlapping with those of each word were selected. Following this criteria, 50 words were selected and matched to 10 original words included in the original Rey list. The 50 words were subsequently divided into 5 lists, each containing 10 words. Lists were balanced and statistically matched to the original Rey list. A one-way ANOVA analysis revealed no significant difference for (i) word length ($F(5, 54) = 0.095$, $P = 0.993$), (ii) FAM ($F(5, 54) = 0.856$, $P = 0.517$), (iii) CNC ($F(5, 54) = 1.146$, $P = 0.348$), (iv) AoA ($F(5, 54) = 1,416$, $P = 0.233$), and (v) IMG ($F(5, 54) = 1.096$, $P = 0.373$). Bonferroni post hoc corrections for pairs of lists revealed no significant difference between lists for none of the above-mentioned variables (all P values <1). Semantic or phonemic similarities between words within each list were also excluded.

Among these five lists, two were selected as target stimuli lists, while words belonging to the remaining three lists were used as filler words for the recognition part of the visual version of AVLT (see Section 3.6.3). Each of the two-target lists was pseudorandomized according to three different word orders to control for possible word list sequence effects. Thus, one out of the resulting 6 lists (i.e., two words sets, each randomized three times) was selected as target list for each participant in both versions of the task.

3.6.2. Verbal AVLT Task by the Touch-Screen Tablet. A list of 10 target words was presented verbally to the participants. Specifically, the words were administered as auditory signal, and the sound was generated using a text-to-speech software (Audacity, the free cross-platform sound editor software, http://www.audacity.sourceforge.net/), and recorded, refined and normalized in order to equalize all words in terms of voice quality and volume.

After each presentation the participant was asked to verbally recall as much words as possible.

The participant was tested three times consecutively as in [19] for the immediate memory recall and once for the delayed memory component. Vocal recordings of target words were used as stimuli in this task version and were administered through the integrated speakers of the touch screen tablet. Stimuli were presented consecutively with an interstimulus interval (ISI) of 3000 ms occurring between one stimulus and the next one. Recall sessions were tested using a voice recording script in order to store an audio logfile of participant responses. Parallel to the sound recording, the 10 target words were constantly displayed on the screen only for the experimenter view, allowing making an online monitoring of given responses by the experimenter (this made the task not purely computerized).

3.6.3. Visual AVLT Task by the Touch-Screen Tablet. In the visual version of the task a total number of 10 target words were presented to the participant. Immediately after viewing these 10 target stimuli he/she was presented with the same 10 words randomized together with another set of 10 fillers (i.e., words not present in the targets lists) and asked to touch the screen each time he/she recognized an item belonging to the previously presented list. The same setting was used for the delayed part of the task, with the only exception that he/she was presented with the same 10 words randomized together with another set of 20 fillers, including the 10 fillers presented before. We chose 20 fillers for the delayed task because we tripled the number of the target words (10). This approach has been adopted in the original paper and pencil Rey auditory verbal test [24].

Prior to task administration, the experimenter handed the tablet to the participant, who was therefore actively required to use the testing apparatus while the experimenter would only passively control upon subject's performance.

Stimuli were presented as white words (with a font size of 36 points) at the center of a black screen and lasted for 3000 ms.; an ISI of 2000 ms occurred between one stimulus and the next one.

Accuracies were estimated by comparing each stimulus code (i.e., target or filler) to subject response and stored in a logfile.

3.7. Experimental Design for Testing Efficacy. A combination of one AVLT version (i.e., visual or verbal) and one ANT (i.e., touch or mouse) layout was administered to each subject, resulting in a factorial design in which (i) 19 subjects (10 males) performed the touch response layout of ANT and the other 19 subjects (9 males) performed the mouse response layout, and (ii) 19 out of 38 subjects (11 males) performed

the visual version of AVLT while the remaining 18 subjects (9 males) performed the verbal version of AVLT.

Groups were statistically matched for age, education, and MMSE and no significant difference emerged when comparing the (i) mouse versus touch group (age $P = 0.831$; education $P = 0.970$; and MMSE $P = 0.272$) and the (ii) verbal versus visual group (age $P = 0.970$; education $P = 0.487$; and MMSE $P = 0.344$).

AVLT lists were randomly and evenly distributed across subjects for both verbal and visual versions of the task. Each subject was first asked to complete the immediate memory recall/recognition of AVLT and was tested 15 minutes later for the delayed, long-term retrieval component [24].

Two experimental runs of ANT (7 minutes each) were administered in between the two AVLT components as interference task, that is, a task critically involving neither the verbal nor the learning cognitive resources recruited during AVLT.

3.8. Ecology: Satisfaction Survey. After the administration of the tests, independently from the test (AVLT or ANT) and response layout (touch or mouse), participants were required to complete a satisfaction survey. The survey is a 16-item self-report questionnaire that uses a 5-point Likert scale (for the complete list of items of the questionnaire see Supplementary Materials available online at http://dx.doi.org/10.1155/2014/804723).

The survey measured the following: the participant frequency of use of touch-screen tablet and of mouse-controlled PC (items 1, 2, and 3), the participant qualitative perception of the familiarity with the touch-screen tablet (item 4), the participant qualitative perception of the comfortableness with the touch-screen tablet (items 5, 6, and 7), the participant qualitative perception of the testing environment (items 8 and 9), the participant fatigue of using sensory functions while interacting with the touch-screen tablet (items 10, 11, 12, 13, and 14), the participant fatigue of maintaining the concentration while interacting with the touch-screen tablet (item 15), and the participant time perception of the neuropsychological tests (item 16).

Items administration was customized depending on the experimental (i.e., different combinations of touch or non-touch versions of AVLT and ANT) setting of each participant.

4. Data Analysis

4.1. Testing Efficacy I: Executive Function Assessment. In order to assess possible differences between the two settings (touch-screen tablet versus mouse-control PC), we evaluated the recorded hardware uncertainties to control for specific setting effects (touch versus mouse) on stimulus presentation. Hardware uncertainties are provided by the presentation software giving information on how the hardware is managing stimuli presentation and if some hardware-based source of variability is altering the script management by the software. Specifically, for each Presentation event in the logfile (except for pause, resume, and quit events), presentation provides a time of occurrence T (ms) and an uncertainty dT (ms). These two numbers provide bounds on the time of occurrence of a presentation event. If Hardware uncertainties dT remain less than 0.6 ms the response script is running as expected (http://www.neurobs.com/, Presentation help). Metrics measured by hardware uncertainties refer to time responses of the device to the stimuli software presentation, independently from the subject response to stimuli presentation.

4.1.1. Data Screening. Concerning the ANT, we considered the distributions of reaction times (RT) and accuracies. Specifically, with respect to reaction times (RT), any trial with recorded RT that fell two SDs above or below the calculated RT mean ($RT \geq \text{mean} + 2SD$) ($RT \leq \text{mean} \pm 2SD$) was rejected.

With respect to accuracy each experimental condition with measured accuracy that fell under 80% was rejected.

4.1.2. Statistical Analysis. Based on the selected datasets, we performed a 3×2 repeated measures ANCOVA with effect type (i.e., conflict effect, alerting effect, and orienting effect) as within-subjects factor with three levels.

A between-subjects factor was considered, namely, the participant group with two levels (mouse versus touch layout). Although groups were sampled with comparable values of MMSE, age, and education (see Section 3.7 *Experimental Design for Testing Efficacy* for groups comparisons statistics) these variables were included as covariates, in order to account for their potential influence on task performance. We assessed the main effect of group of participants, covariates, and exclusively interactions with the between-subjects factor (i.e., group of participants).

4.2. Testing Efficacy II: Memory Function Assessment. Concerning the verbal AVLT, vocal recordings were listened to and classified as correct or incorrect, while, for the visual AVLT, responses to target stimuli were stored in a logfile and successively coded as correct or incorrect.

Accuracies for the three immediate recall sessions (i.e., IR-1, IR-2, and IR-3) and the delayed recognition session (DR) were coded as percentages. Analyses of equality between verbal and visual AVLT were performed with a Mann-Whitney U test (nonparametric) for independent samples (i.e., according to AVLT verbal or visual condition) on the arcsine-transformed percentages of accuracy (for each recall) and the delayed session independently. Correlations between the MMSE scores and the arcsine-transformed percentages of correct responses for IR-1, IR-2, IR-3, and DR were then evaluated using Kendall's tau correlation coefficient split by AVLT verbal and AVLT visual conditions.

4.3. Ecology. Firstly, we performed one sample Wilcoxon signed rank test for each of the 16 items, measured by the survey administered to the participants, versus the middle level of perceived scale quality (3 with respect of a scale maximum of 5). Furthermore, we investigated the relationship between the level of preference of participants for the testing environment (i.e., item 7, using the touch-screen tablet with respect to an external device such as a mouse or a keyboard)

and the perceived degree of easiness when using a touch-screen tablet (i.e., item 5) or the perceived degree of easiness when touching the screen (i.e., item 6).

5. Results

5.1. Testing Efficacy I: Executive Function Assessment. No anomaly linked to hardware management of stimuli presentation was detected when screening both mouse and touch logfiles.

5.1.1. Data Screening. All subjects had an RT outlier percentage <30% on each experimental condition and accuracies were >80% for all participants; no subject was therefore excluded from the analyses.

5.1.2. Statistical Analysis Results. Mean RTs for trial conditions for the calculation of each of the three effects are summarized in Table 3. The main effect of group of participants was found not to be significant ($F(1, 33) = 0.008$, $P = 0.929$) revealing that performances are comparable in terms of overall RTs for the ANT with mouse or touch response layout. No significant main effects of covariates were found (MMSE $F(1, 33) = 0.049$, $P = 0.826$; age $F(1, 33) = 1.221$, $P = 0.277$; and education $F(1, 33) = 0.359$, $P = 0.553$).

The interaction between group of participants and effects showed a trend toward significance ($F(2, 66) = 2.357$, $P = 0.116$, Greenhouse-Geisser correction) and a plot of this 2-way interaction showed a reduced CE and an incremental AE and OE effect for the touch group (see Figure 2); pairwise, post hoc comparisons revealed a trend towards significance when considering CE differences between groups (mean difference, Touch − Mouse = −37.280; $P = 0.170$) in terms of a reduced CE for the touch, while the two groups differed to a lesser degree for the OE (Mean Difference, Touch − Mouse = 18,218; $P = 0.232$) and AE (Mean Difference, Touch − Mouse = 16,418; $P = 0.261$).

5.2. Testing Efficacy II: Memory Function Assessment. No significant differences were observed between verbal and visual recall/recognition performances for IR-1 (Mann-Whitney test; $P = 0.365$), IR-2 (Mann-Whitney test; $P = 0.293$), IR-3 (Mann-Whitney test; $P = 0.694$), and DR (Mann-Whitney test; $P = 0.988$).

For AVLT verbal, MMSE scores correlated with performance on IR-1, IR-2 immediate and with the delayed recall session; a trend towards correlation was found for the third immediate recall session.

For AVLT visual, MMSE scores showed a correlation trend with IR-1 and a significant correlation with IR-2, IR-3, and DR.

For AVLT verbal all immediate and delayed recall sessions significantly correlated with age.

For AVLT visual, recognition sessions did not significantly correlate with age.

For AVLT verbal, education scores significantly correlated with all the immediate and the delayed recall sessions. For AVLT visual, education scores did not correlate with immediate recall sessions while a significant correlation was found with education for the delayed recall session. Table 4 summarizes correlations coefficients and statistical significance between each AVLT recall session and socio demographics (age and education) and cognitive index (MMSE) variables tested.

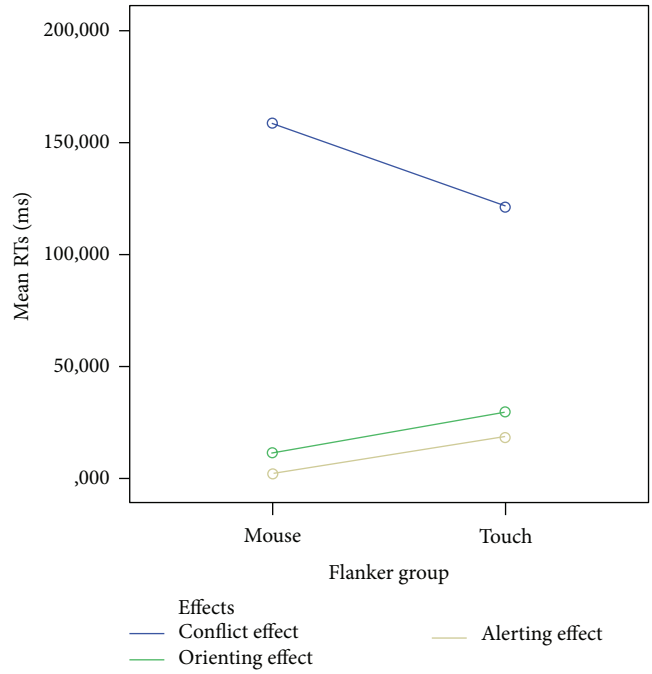


Figure 2: Plot of 2-way interaction group (mouse or touch) X effect (conflict effect, orienting effect, and alerting effect). Mean RTs (y-axis) for conflict Effect (Blue Line), orienting Effect (Green Line), and alerting Effect (Brown Line) are shown for mouse (leftmost dots) and touch (rightmost dots) performances (x-axis).

5.3. Ecology. Wilcoxon one sample signed rank test indicated that the middle percentage of quality ratings was significantly lower than 3 for item 1 ("how much do you use the tablet in your daily life?" $P = 0.003$), for item 4 (did you feel uncomfortable using the tablet?, $P = 0.000$), for item 10 (did you feel fatigued while handling the tablet?, $P = 0.000$), for item 12 (did you feel fatigued while touching the screen?, $P = 0.000$), for item 13 (did you feel fatigued while listening to vocal recordings?, $P = 0.005$), for item 14 (did you feel fatigued producing a vocal response?, $P = 0.001$), and for item 16 (did you feel the experiment had a too long duration?, $P = 0.000$), which is indicative of very low levels of uneasiness (i.e., item 4) or fatigue of using sensory functions (i.e., items 12, 13, and 14) and a perceived very long duration of test administration (i.e., item 16), notwithstanding a very low use of a touch-screen tablet in everyday life (i.e., item 1). The relationship between the preference for using the touch screen (i.e., item 7) and the perceived degree of easiness when using a tablet (i.e., item 5, Kendall's tau = 0.484, $P = 0.009$) or the perceived degree of easiness when touching the screen (i.e., item 6, Kendall's tau = 0.397, $P = 0.025$) were both significant (Kendall's tau = 0.83, $P = 0.02$) (see Supplementary Materials online for the complete items list).

Table 3: Mean RTs (along with SD) for each experimental condition used to calculate effects of ANT are shown in the table for both touch (upper panel) and mouse (lower panel) versions of the task. Congruent trials are subtracted from incongruent trials in order to calculate conflict effect, double cue trials are subtracted from no cue trials in order to calculate the alerting effect, and spatial cue trials are subtracted from central cue trials in order to calculate the orienting effect.

Trial type	Mean	SD	Trial type	Mean	SD	Effect	Mean	SD
				Touch layout				
Incongruent	1016.229	122.973	Congruent	890.436	125.071	Conflict effect	125.793	61.692
No cue	936.42	122.244	Double cue	914.954	127.472	Alerting effect	21.466	49.6
Central cue	932.806	117.579	Spatial cue	908.124	129.387	Orienting effect	24.682	34.646
				Mouse layout				
Incongruent	941.746	181.163	Congruent	786.627	109.758	Conflict effect	155.118	89.207
No cue	833.05	125.772	Double cue	828.475	135.419	Alerting effect	4.57	34.988
Central cue	830.022	129.443	Spatial cue	820.197	140.739	Orienting effect	9.825	54.498

Table 4: Results show correlations coefficients (upper values in cells) and statistical significance (lower values in cells) between each AVLT recall session (i.e., IR-1, IR-2, IR-3, and DR) and sociodemographics (age and education) and cognitive index (MMSE) variables tested. The left most panel of table shows statistics belonging to the group tested with verbal version of the AVLT while the rightmost panel shows statistics belonging to the group tested with the visual version of the task.

	Verbal				Visual			
	IR-1 (τ)	IR-2 (τ)	IR-3 (τ)	DR (τ)	IR-1 (τ)	IR-2 (τ)	IR-3 (τ)	DR (τ)
MMSE	0.633	0.492	0.373	0.406	0.246	0.579	0.451	0.407
	($P = 0.001^{**}$)	($P = 0.013^{*}$)	($P = 0.040^{*}$)	($P = 0.069$)	($P = 0.195$)	($P = 0.003^{**}$)	($P = 0.020^{*}$)	($P = 0.035^{*}$)
Age	−0.54	−0.479	−0.434	−0.353	−0.137	−0.19	0.007	−0.216
	($P = 0.031^{**}$)	($P = 0.008^{**}$)	($P = 0.050^{*}$)	($P = 0.240$)	($P = 0.448$)	($P = 0.303$)	($P = 0.970$)	($P = 0.240$)
Education	0.448	0.527	0.408	0.404	0.183	0.25	0.164	0.38
	($P = 0.017^{*}$)	($P = 0.006^{**}$)	($P = 0.039^{*}$)	($P = 0.034$)	($P = 0.332$)	($P = 0.197$)	($P = 0.409$)	($P = 0.049^{*}$)

(τ = Kendall's tau. *Correlation is significant at the 0.05 level, 2-tailed).
(**Correlation is significant at the 0.01 level, 2-tailed).

6. Discussion

As general consideration, for the purpose of our work, we have chosen two representative tests to assess two main objectives.

We investigated the memory domain, given that the earlier cognitive symptoms reported in Alzheimer's disease, the most common form of dementia, involve memory [28]. Therefore, the majority of tools have focused mainly on this cognitive ability and have implemented tests tailored at investigating memory impairments, of which, the most clinically validated and commonly employed in a clinical setting is the AVLT test. Furthermore, memory impairments, in particular, are a cardinal feature of the majority of dementia syndromes.

We investigated the executive function domain by the use of the ANT, as a useful test for assessing differences between responses, given that it encompasses three different effects relying on three different cognitive mechanisms (i.e., conflict effect, alerting effect, and orienting effect) [29].

The first aim of the present study was to evaluate the testing efficacy of a computerized neuropsychological assessment when implemented on a touch-screen device.

To this aim, we created and tested two different experimental settings in terms of response layout; specifically, two identical versions of an experimentally validated attentional task (i.e., ANT; [22, 23]), differing only in terms of response modality were implemented: one version of the task required subjects to give a response with a mouse device while the other one by using a touch screen. This was made in order to directly compare the testing efficacy of a psychological test (in our case the evaluation of executive function) when administered by a touch-screen device with respect to a more conventional mouse-control PC.

Comparisons of reaction times between subjects using touch screen or mouse and of their test performance revealed no significant overall differences, suggesting that touch screen and mouse can be equally chosen as response devices, since they grant the same experimental outcome. These findings strengthen the results highlighted by Sears and Shneiderman [30] who, although under different experimental conditions, compared touch-screen response layout versus mouse response layout. Their results suggested substantial comparability between these two input devices. Our finding, consistently with results from other authors, is of particular interest, given that the touch-screen technology is currently widely spreading, also among the elderly population. Touch-screen tablets are innovative technological solutions which are emerging also as devices for healthcare intervention. Healthcare services are indeed progressively showing an increasing interest in translating services into touch-screen based environment [31, 32]. From this perspective it is an

important topic to test if such devices can guarantee the same testing efficacy of more conventional and extensively validated devices, as mouse-control PC, which are still today used in clinical environment to administer neuropsychological tests. Our work, despite being validated on a limited number of subjects, suggests that this technological solution is feasible for test administration.

The second aim of our work was to evaluate the testing efficacy of a computerized test (in a representative case of memory function assessment) when administered by the touch-screen tablet with respect to two different experimental settings in terms of stimuli presentation.

To this aim, we created and tested two different versions of AVLT, a widely standardized and validated neuropsychological test: (i) a visual version, replicating the visual porting of this task which is currently used by a set of different digital neuropsychological training batteries [19, 33]: in this versions stimuli were presented visually and subjects were asked to recognize memorized stimuli among other nontarget stimuli and (ii) a verbal version, replicating the "classical" administration of the task in the clinical context: in this version stimuli were presented verbally and subjects were asked to freely recall all the memorized target stimuli. This was done with the purpose to assess the effects of the stimuli presentation modality on the efficacy of test (in our case the evaluation of memory function).

Our results for the verbal version of AVLT showed significant correlations between MMSE scores and performance on IR-1, IR-2 immediate and with the delayed recall session. A trend towards significance was found for the third immediate recall session (IR-3). Similarly, results for the visual version of AVLT showed a correlation trend between MMSE scores and performance with IR-1 and a significant correlation with IR-2, IR-3 and with the delayed recall session. For AVLT verbal, all immediate and delayed recall sessions showed an inverse and significant correlation with age (i.e., lower scores of age correspond to higher values of recall performance); education scores significantly and positively correlated with all immediate and the delayed recall sessions (i.e., higher values of education correspond to higher values of recall performance). For AVLT visual, a significant and direct correlation with recognition performance scores was found only with education for the delayed recall session and no significant correlations were found between age and performance on both immediate and delayed recognitions (see Table 4 for detailed results).

Overall these results suggest that both implementations of the test (i.e., visual and verbal) are affordable measures of the general cognitive status, directly correlated with a measure of general cognitive status assessment (i.e., MMSE); from this point of view they can be both considered affordable tools for a broad cognitive assessment. In spite of this only the verbal version of the task showed a correlation with the sociodemographical data of our sample (i.e., an inverse correlation with age and direct correlation with education).

It should be in fact recognized that we have not compared verbal with "pure" visual stimuli presentation, since filler words were presented during the task: under this light the two tasks share some common features but differ for others in terms of both experimental setting and underlying cognitive processes.

In fact the verbal version requires an active retrieval from memory of the presented stimuli, while the visual version requires the recognition and discrimination of the memorized target stimuli among other nontarget stimuli. From this point of view they require the subject to use different strategies to be solved and thus may involve different brain networks (e.g., [34]). For example from the memory point of view the visual modality could be easier since the subject is provided with a cue (namely, the target stimulus is directly presented to the subject) but at the same time it requires the inhibition of distracters and may be more difficult in terms of executive functioning cognitive load (namely, the subject has to discriminate the target stimulus among other nontarget words).

Another crucial aspect concerning the introduction of new technological solutions in the everyday life is the degree of ecology and the level of preference of the computerized assessment modalities regarding the administration of the neuropsychological tests. Although some authors found significant draw backs to touch screens in the elderly [35] others (e.g., [31]) reported that touch-screen devices are ideal instruments for assessing populations with low technological familiarity, such as elders and patients. Our results on elderly and healthy participants confirm this finding, considering that our subjects felt comfortable using the touch-screen device and did not experience unease or fatigue feelings while performing the tests. Crucially, all subjects possessed low familiarity with such devices and, in some cases, it was their first experience of physical interaction with a touch-screen tablet.

Given the performance comparability between responses using mouse and touch, it is important to introduce some considerations for future evaluation regarding whether (and under which circumstances) it is preferable to choose one or the other response layout.

While no main effect of group was highlighted, our analysis revealed that subjects performing the ANT task using the touch response layout showed a tendency towards an advantage for all three effects accounted by the task and, namely, a trend towards a significant reduction of conflict effect and slightly larger alerting and orienting effects (see Figure 2 and Table 3 for details).

However, it must be acknowledged that RTs from which the effects are derived showed the same pattern for both the mouse and the touch layout; namely, for the conflict effect RTs are the longer with respect to congruent ones while for alerting effect RTs measured following the double cue were shorter with respect to when no cue was presented, and for orienting effect responses after a spatial cue were faster with respect to a cue presented centrally. This means that there is no difference in terms of performance for all ANT conditions. However, when assessing specific cognitive processes measured through the difference between RTs it appears that the touch device may provide some benefits mainly on cognitive control, enhancing performance on the more demanding trial type. Namely, conflict effect is calculated by subtracting RTs belonging to congruent trials (flanking arrows pointing in

the same direction of the central target arrow) from RTs belonging to incongruent trials (flanking arrows pointing in the opposite direction of the central target arrow). To this extent lower values of conflict effect indicate a higher performance on cognitive conflict resolution.

This finding becomes of particular interest when considering the kind of trial associated to this reduction in the effect size. Incongruent trials require the resolution of a cognitive conflict to be solved and they are known to be the slowest ANT experimental conditions in terms of response speed [22].

Forlines et al. [36] showed that tasks requiring the use of two hands (bimanual tasks, as the ANT) are better performed when using touch-screen devices with respect to mouse-control PC. To a similar extent Rogers et al. [37] found that touch-screen devices are particularly suitable for response collection when compared to another indirect response device (namely, a rotary encoder).

Given our results, it is important to acknowledge this proposed dichotomy between direct (the touch-screen, in our case) and indirect response devices. When using a direct input device, the distance between the subject (his/her fingers) and the causal effect he/she carries on the environment modification (touching stimuli on the screen, as required by the task) is reduced. Touch-screen devices, in this framework, lead a virtual environment to a more tangible and ecological dimension. One possible consequence of such phenomenon could be an increase in self-commitment or in self-perceived efficacy towards the task, and this could lead to an enhancement by establishing a direct link between the subject and the task reality. In other words, a different perception of the self-commitment could be associated with responses given with direct input devices, shifting the task environment perception into a more concrete entity on which the subject acts as a physical agent. Thus, critically, the subject involvement into the task could have been enhanced.

Under this light one would expect to observe a greater effect for those trials requiring a greater cognitive demand (i.e., incongruent trials). A greater involvement could translate into greater resources dedicated to task solution. To this extent, cognitively simpler trials (i.e., congruent trials) would benefit less, since they do need less work to be solved; on the contrary, trials requiring greater cognitive effort to be solved, such as incongruent trials, would greatly benefit from such resource availability.

Although this scenario is suggestive, some dedicated experimental investigation is needed to shed light on the cognitive basis of this behaviorally observed phenomenon.

These evidences, taken together with results on the ease of their use highlighted by the survey, indicate touch-screen devices as an ecological and suitable tool for the computerized administration of neuropsychological tests. Furthermore, other authors [35] showed that alternative response input devices, such as a light pen or touch screen are highly intuitive, and have the advantage of bypassing the keyboard. They demonstrate how these devices allow subjects to focus their attention directly on the video display terminal and not have to shift their attention from the monitor to the keyboard to locate a response key. Nevertheless, light pens and touch screens also have their disadvantages. They require the subject to hold his or her arm in an "up" position and move it along the screen. This can produce fatigue and some variation in reaction time.

It should be noted that the computerized assessment does not represent an alternative to the clinical setting. However, it can contribute in a significant manner to the traditional evaluation. Nevertheless, there is a need to further detail some aspects of our investigation: (i) in order to increase the inferential power and experimental validity of our findings, the tests will need to be administered to a larger number of participants; (ii) among these participants, specific cognitively-impaired populations and physically-impaired populations (e.g., those subjects with motor function deficit from a brain injury that could affect the test performance) will need to be tested in order to assess if these instruments can be a valid and accessible tools in the clinical context, and, (iii) most importantly, a dedicated version of cognitive domain-specific tests will need to be implemented and case-wise tested in order to detail whether, and to which extent, they can be a valid alternative to more conventional pc-based and/or pencil and paper testing approaches.

7. Conclusions and Future Perspectives

This work provides new data on the experimental feasibility and clinical ecology of computerized neuropsychological assessment by addressing the impact of the implementation of different user interfaces and different stimuli presentation modality.

In order to set up an innovative computerized testing environment, while keeping it feasible and ecological, it is fundamental to detail how this conversion process impacts the experimental and clinical neuropsychological settings. Although limited on approximately 40 healthy subjects and experimented only on representative, not exhaustive, neuropsychological tests (on memory and attention functions) our evidences suggest that touch-screen devices can be considered for the computerized administration of neuropsychological tests.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Authors' Contribution

Matteo Canini, Petronilla Battista, and Pasquale Anthony Della Rosa contributed equally to the paper.

Acknowledgment

This work was supported by the fund for research of the Italian Ministry of University and Research, within a framework agreement between Lombardy Region and National Research Council of Italy (no. 17125, 27/09/2012).

References

[1] R. E. Schlegel and K. Gilliland, "Development and quality assurance of computer-based assessment batteries," *Archives of Clinical Neuropsychology*, vol. 22, no. 1, pp. 49–61, 2007.

[2] S. Zygouris and M. Tsolaki, "Computerized cognitive testing for older adults. A review," *American Journal of Alzheimer's Disease & Other Dementias*, 2014.

[3] M. D. Lezak, *Neuropsychological Assessment*, Oxford University Press, Oxford, UK, 2004.

[4] O. Spreen, *A Compendium of Neuropsychological Tests: Administration, Norms, and Commentary*, Oxford University Press, 1998.

[5] M. D. Franzen and P. A. Arnett, "The validity of neuropsychological assessment procedures," in *Biological and Neuropsychological Mechanisms: Life-Span Developmental Psychology*, H. W. Resse and M. D. Franzen, Eds., pp. 51–69, Erlbaum, Mahwah, NJ, USA, 1997.

[6] R. W. Heinrichs, "Current and emergent applications of neuropsychological assessment: problems with validity and utility," *Professional Psychology: Research and Practice*, vol. 21, no. 3, pp. 171–176, 1990.

[7] H. C. Fichman, R. Nitrini, P. Caramelli, and K. Sameshima, "A new brief computerized cognitive screening battery (CompCogs) for early diagnosis of Alzheimer's disease," *Dementia e Neuropsychologia*, vol. 2, no. 1, pp. 13–19, 2008.

[8] C. Bartels, M. Wegrzyn, A. Wiedl, V. Ackermann, and H. Ehrenreich, "Practice effects in healthy adults: a longitudinal study on frequent repetitive cognitive testing," *BMC Neuroscience*, vol. 11, no. 1, article 118, 2010.

[9] J. Fredrickson, P. Maruff, M. Woodward et al., "Evaluation of the usability of a brief computerized cognitive screening test in older people for epidemiological studies," *Neuroepidemiology*, vol. 34, no. 2, pp. 65–75, 2010.

[10] E. Woo, "Computerized neuropsychological assessments," *CNS Spectrums*, vol. 13, no. 10, supplement 16, pp. 14–17, 2008.

[11] J.-A. Witt, W. Alpherts, and C. Helmstaedter, "Computerized neuropsychological testing in epilepsy: overview of available tools," *Seizure*, vol. 22, no. 6, pp. 416–423, 2013.

[12] R. C. K. Chan, D. Shum, T. Toulopoulou, and E. Y. H. Chen, "Assessment of executive functions: review of instruments and identification of critical issues," *Archives of Clinical Neuropsychology*, vol. 23, no. 2, pp. 201–216, 2008.

[13] American Psychological Association, Committee on Professional Standards, American Psychological Association, Board of Scientific Affairs, and Committee on Psychological Tests and Assessment, *Guidelines for Computer-Based Tests and Interpretations*, The Association, 1986.

[14] K. Wild, D. Howieson, F. Webbe, A. Seelye, and J. Kaye, "Status of computerized cognitive testing in aging: a systematic review," *Alzheimer's and Dementia*, vol. 4, no. 6, pp. 428–437, 2008.

[15] C. T. Gualtieri and L. G. Johnson, "Reliability and validity of a computerized neurocognitive test battery, CNS Vital Signs," *Archives of Clinical Neuropsychology*, vol. 21, no. 7, pp. 623–643, 2006.

[16] P. Maruff, E. Thomas, L. Cysique et al., "Validity of the CogState brief battery: relationship to standardized tests and sensitivity to cognitive impairment in mild traumatic brain injury, schizophrenia, and AIDS dementia complex," *Archives of Clinical Neuropsychology*, vol. 24, no. 2, pp. 165–178, 2009.

[17] T. Dwolatzky, V. Whitehead, G. M. Doniger et al., "Validity of a novel computerized cognitive battery for mild cognitive impairment," *BMC Geriatrics*, vol. 3, article 1, 2003.

[18] R. H. Paul, J. Lawrence, L. M. Williams, C. C. Richard, N. Cooper, and E. Gordon, "Preliminary validity of "integneuro": a new computerized battery of neurocognitive tests," *International Journal of Neuroscience*, vol. 115, no. 11, pp. 1549–1567, 2005.

[19] M. Inoue, D. Jimbo, M. Taniguchi, and K. Urakami, "Touch panel-type dementia assessment scale: a new computer-based rating scale for Alzheimer's disease," *Psychogeriatrics*, vol. 11, no. 1, pp. 28–33, 2011.

[20] K. Onoda, T. Hamano, Y. Nabika et al., "Validation of a new mass screening tool for cognitive impairment: cognitive assessment for Dementia, iPad version," *Clinical Interventions in Aging*, vol. 8, pp. 353–360, 2013.

[21] A. Blackwell, "Tag Archives: CANTAB mobile," 2011.

[22] J. Fan, B. D. McCandliss, J. Fossella, J. I. Flombaum, and M. I. Posner, "The activation of attentional networks," *NeuroImage*, vol. 26, no. 2, pp. 471–479, 2005.

[23] P. A. Della Rosa, G. Videsott, V. M. Borsa et al., "A neural interactive location for multilingual talent," *Cortex*, vol. 49, no. 2, pp. 605–608, 2013.

[24] A. Rey, *L'Examenclinique en Psychologie*, Presses Universitaires de France, Paris, France, 1964.

[25] C. Caltagirone, G. Gainotti, C. Masullo, and G. Miceli, "Validity of some neuropsychological tests in the assessment of mental deterioration," *Acta Psychiatrica Scandinavica*, vol. 60, no. 1, pp. 50–56, 1979.

[26] A. Laudanna, A. M. Thornton, G. Brown, C. Burani, and L. Marconi, "Un corpus dell'italiano scritto contemporaneo dalla parte del ricevente," in *III Giornate Internazionali di Analisi Statistica dei Dati Testuali*, S. Bolasco, L. Lebart, and A. Salem, Eds., vol. 1, pp. 103–109, Centro Informazione Stampa Universitaria, Roma, Italy, 1995.

[27] P. A. Della Rosa, E. Catricalà, G. Vigliocco, and S. F. Cappa, "Beyond the abstract-concrete dichotomy: mode of acquisition, concreteness, imageability, familiarity, age of acquisition, context availability, and abstractness norms for a set of 417 Italian words," *Behavior Research Methods*, vol. 42, no. 4, pp. 1042–1048, 2010.

[28] A. J. Larner, *Neuropsychological Neurology: The Neurocognitive Impairments of Neurological Disorders*, Cambridge University Press, Cambridge, UK, 2013.

[29] J. W. MacLeod, M. A. Lawrence, M. M. McConnell, G. A. Eskes, R. M. Klein, and D. I. Shore, "Appraising the ANT: psychometric and theoretical considerations of the attention network test," *Neuropsychology*, vol. 24, no. 5, pp. 637–651, 2010.

[30] A. Sears and B. Shneiderman, "High precision touchscreens: design strategies and comparisons with a mouse," *International Journal of Man-Machine Studies*, vol. 34, no. 4, pp. 593–613, 1991.

[31] A. Holzinger, "Finger instead of mouse: touch screens as a means of enhancing universal access," in *Universal Access Theoretical Perspectives, Practice, and Experience*, pp. 387–397, Springer, Berlin, Germany, 2003.

[32] P. Thekkumpurath, C. Venkateswaran, M. Kumar, A. Newsham, and M. I. Bennett, "creening for psychological distress in palliative care: performance of touch screen questionnaires compared with semistructured psychiatric interview," *Journal of Pain and Symptom Management*, vol. 38, no. 4, pp. 597–605, 2009.

[33] H. C. Fichman, R. Nitrini, P. Caramelli, and K. Sameshima, "A new Brief computerized cognitive screening battery (CompCogs) for early diagnosis of Alzheimers disease," *Dementia & Neuropsychologia*, vol. 2, no. 1, pp. 13–19, 2008.

[34] G. Radvansky, *Human Memory*, Pearson Education, Boston, Mass, USA, 2006.

[35] E. Wood, T. Willoughby, A. Rushing, L. Bechtel, and J. Gilbert, "Use of computer input devices by older adults," *Journal of Applied Gerontology*, vol. 24, pp. 419–438, 2005.

[36] C. Forlines, D. Wigdor, C. Shen, and R. Balakrishnan, "Direct-touch vs. mouse input for tabletop displays," in *Proceedings of the 25th SIGCHI Conference on Human Factors in Computing Systems (CHI '07)*, pp. 647–656, ACM, May 2007.

[37] W. A. Rogers, A. D. Fisk, A. C. McLaughlin, and R. Pak, "Touch a screen or turn a knob: choosing the best device for the job," *The Journal of the Human Factors and Ergonomics Society*, vol. 47, no. 2, pp. 271–288, 2005.

Hindawi Publishing Corporation
BioMed Research International
Volume 2013, Article ID 780458, 12 pages
http://dx.doi.org/10.1155/2013/780458

Research Article

A Partial Volume Effect Correction Tailored for ^{18}F-FDG-PET Oncological Studies

F. Gallivanone,[1] C. Canevari,[2] L. Gianolli,[2] C. Salvatore,[3] P. A. Della Rosa,[1] M. C. Gilardi,[1] and I. Castiglioni[1]

[1] *IBFM-CNR, Via F.lli Cervi 93, 20090 Segrate, Milan, Italy*
[2] *H San Raffaele, Via Olgettina 62, 20090 Segrate, Milan, Italy*
[3] *University of Milan-Bicocca, Milan, Italy*

Correspondence should be addressed to I. Castiglioni; castiglioni.isabella@hsr.it

Received 30 April 2013; Revised 2 August 2013; Accepted 2 August 2013

Academic Editor: Noriyoshi Sawabata

We have developed, optimized, and validated a method for partial volume effect (PVE) correction of oncological lesions in positron emission tomography (PET) clinical studies, based on recovery coefficients (RC) and on PET measurements of lesion-to-background ratio (L/B_m) and of lesion metabolic volume. An operator-independent technique, based on an optimised threshold of the maximum lesion uptake, allows to define an isocontour around the lesion on PET images in order to measure both lesion radioactivity uptake and lesion metabolic volume. RC are experimentally derived from PET measurements of hot spheres in hot background, miming oncological lesions. RC were obtained as a function of PET measured sphere-to-background ratio and PET measured sphere metabolic volume, both resulting from the threshold-isocontour technique. PVE correction of lesions of a diameter ranging from 10 mm to 40 mm and for measured L/B_m from 2 to 30 was performed using measured RC curves tailored at answering the need to quantify a large variety of real oncological lesions by means of PET. Validation of the PVE correction method resulted to be accurate (>89%) in clinical realistic conditions for lesion diameter > 1 cm, recovering >76% of radioactivity for lesion diameter < 1 cm. Results from patient studies showed that the proposed PVE correction method is suitable and feasible and has an impact on a clinical environment.

1. Introduction

Molecular imaging by positron emission tomography (PET) and ^{18}F-fluorodeoxyglucose (^{18}F-FDG) radiotracer is currently the most commonly used method for the detection and metabolic characterisation of several oncological pathologies, given the possibility to detect foci with an increased ^{18}F-FDG metabolism as those characterising tumour cells (e.g., [1, 2]).

In the PET clinical environment, diagnosis and tumor staging are commonly assessed by qualitative visual inspection of ^{18}F-FDG PET images [3–5]. Nevertheless, a quantitative analysis of ^{18}F-FDG uptake in oncological lesions has been proven to be useful to differentiate benign and malignant tissues (e.g., [6]), to assess response to therapy [7–9], and to predict tumour aggressiveness [10–13].

Despite these benefits, a quantitative approach for the evaluation of PET oncological studies is not a common practice in clinical routine due to the presence of partial volume effect (PVE) on the PET images. PVE is a physical limitation resulting from the poor spatial resolution of PET systems (4-5 mm) that strongly affects the estimation of radioactivity concentration within structures less than two or three times the PET spatial resolution [14, 15].

Several techniques have been advanced to compensate for PVE in PET [15–20]. Among all PVE correction methods, more common ones are based on multiplicative numerical factors (recovery coefficients, RC), recovering the local radioactivity concentration within any small structure which uptakes ^{18}F-FDG. RC can be derived from PET experimental measurements of small radioactive objects in a priori known object-to-background radioactivity concentration ratio.

PET experimental measurements of RC have been carried out by using ^{18}F-FDG radioactive spheres (hot spheres) [14]. RC coefficients were obtained as the ratio between PET measured-and-actual radioactivity concentration within the hot spheres. This approach was applied to the PVE correction of PET oncological lesions in real patients [21], since radioactive spheres were considered suitable to simulate metabolic active oncological lesions. Unfortunately, the method was able to compensate only for the spread out (spill out) of lesion ^{18}F-FDG uptake into the surrounding background of the patient body not accounting for the spread in (spill in) of the background into the lesion, as it occurs in the body tissues surrounding oncological lesions in a real scenario.

More realistic models were developed by combining RC derived from hot spots in cold background, RC from cold spots in hot background, and RC from hot spots in warm background, allowing both spill out and spill in effects to be accounted for (e.g., [15]), but were never applied to real clinical studies.

In all cases, the applicability of RC-based PVE correction methods to PET real oncological studies is still constrained by two problems: the impossibility to estimate both the actual lesion-to-background ratio (L/B) and the actual lesion volume of oncological lesions [22–25]. For instance, measured PET images result intrinsically affected by PVE, and no a priori known information about actual L/B is available for in vivo patient studies. Furthermore, the estimation of the actual volume of an oncological lesion is one of the most debated issues in both the nuclear medicine and radiology community even though it has been coped with from different perspectives.

An RC-based PVE correction method devoted to oncological studies which overcomes the need to actually determine L/B was proposed by Srinivas et al. [26]. They performed PET measurements of hot spheres in hot background and obtained RC as a function of measured L/B (L/B_m), derived from the maximum value of lesion uptake. However, RC curves were obtained as a function of the actual lesion volume of the hot spheres representing a strong limit imposed by the need to know the actual volume of lesions. As Srinivas et al. suggest, when lesion density is different from the density of the surrounding tissues, a CT study in the region of interest can provide lesion anatomical volume. Current generation multimodal computerized tomography (CT)-PET systems allow to obtain anatomical volume of a lesion temporally and spatially coregistered with the metabolic volume. Unfortunately, a lesion is not always visible on CT images and often CT anatomical volume and PET metabolic volume can deviate [27–30].

The applicability of RC-based PVE correction method to real oncological PET-CT images needs an estimation of L/B from measured data. Therefore, another limit of RC-based PVE correction methods is that the accuracy of the chosen RC depends on the accuracy of the technique used for the measurements of the lesion uptake [31]. For instance, operator-dependent techniques for sphere uptake measurements [24, 32–35] can induce operator-dependent differences in the estimation of RC [16]. On the other hand, operator-independent techniques [36–39] are more sensitive to the noise level of PET images and require optimisation strategies and accurate validation [16].

The aim of this work was the development of a method for PVE correction tailored for clinical application to PET-CT oncological studies. Our method is based on RC curves as functions of PET L/B_m and of PET measured lesion volume, both estimated by an operator-independent technique. The proposed PVE correction method was assessed on both anthropomorphic phantoms and in clinical ^{18}F-FDG PET-CT studies.

2. Materials and Methods

2.1. ^{18}F-FDG PET Studies. ^{18}F was produced by a cyclotron (RDS Eclipse, Siemens Healthcare) with a fixed proton beam of 11 MeV. ^{18}F-FDG synthesis was obtained by nucleophilic substitution in acidic medium and subsequent purification.

A dose measurement system (Dose calibrator Pet Dose, Comecer) provided measurements of the amount of ^{18}F-FDG radioactivity (administered and residual) for all phantoms and patient studies.

The multimodal PET-CT system (Discovery STE, General Electric Medical System), cross-calibrated with the dose measurement system, was used for PET-CT measurements. D-STE is a 3D hybrid system that combines a 16 multislice helical CT scanner with a PET scanner of 280 bismuth oxygen germinate crystals ($4.7 \times 6.3 \times 30\ \text{mm}^3$) arranged in 24 rings. Transaxial field of view is 60 cm and 50 cm for PET and CT, respectively. Axial field of view is 15.7 cm for PET.

Oncological protocol was set as follows: a SCOUT scan at 40 mA, followed by a CT scan at 140 mV and 150 mA (10 sec), and 3D PET scans (2.5 min/scan) for adjacent bed positions. For each bed position, CT data were reconstructed into a $512 \times 512 \times 47$ matrix with a voxel size of $0.97 \times 0.97 \times 3.27\ \text{mm}^3$ [40]. For each bed position, PET data were sampled into a $128 \times 128 \times 47$ matrix with a voxel size of $4.7 \times 4.7 \times 3.27\ \text{mm}^3$ and reconstructed using a 3D ordered subset expectation maximization algorithm (OSEM) with corrections for random, scatter, and attenuation incorporated into the iterative process.

2.2. Synthetic Oncological Lesions. Perspex spheres of different diameters were used to simulate oncological lesions.

Six spheres (diameter = 10 mm, 13 mm, 17 mm, 23 mm, 29 mm, and 37 mm) within an elliptical perspex cylinder (d_1 = 24 cm, d_2 = 30 cm, and h = 21 cm) [41] were used for the estimation of RC.

Three spheres (diameter = 9.8, 12.3, and 15.6 mm) were placed in different regions of different anthropomorphic phantoms (thorax, breast, and brain) and were used for the validation of the proposed PVE correction method in clinical-like oncological studies. Specifically, the three spheres were placed in

(1) a thorax-like phantom (d_1 = 20 cm, d_2 = 30 cm, and h = 21 cm) with two cork parts simulating lungs and a cardiac insert;

(2) a breast-like phantom consisting into the previously described thorax phantom (no cardiac insert) and

Table 1: Characteristics and available data of patients.

Patient group	N	Purpose	Available data
Gastro	49	Tumor staging	^{18}F-FDG PET-CT study (basal), tumor histotype (SRC, SC)
Breast	40	Tumor staging	^{18}F-FDG PET-CT study (basal), Mib-1
Head-neck	19	Tumor staging	^{18}F-FDG PET-CT study (basal), DFS
Skeleton	29	Therapy monitoring	Basal and follow-up ^{18}F-FDG PET-CT studies

into two plastic containers (cylinder equivalent radius = 3 cm, h = 10 cm) miming breasts;

(3) brain-like phantom: the Hoffman 3D brain phantom [15].

Three additional nonspherical lesions consisting of zeolites were considered. Zeolites are porous aluminosilicate minerals already used to simulate oncological lesions in anthropomorphic phantoms assessed by ^{18}F-FDG PET-CT studies. When soaked into an aqueous solution of ^{18}F-FDG, they are able to absorb and not release ^{18}F-FDG molecule, in a nonhomogeneous way for short soaking duration and in a homogeneous way for long soaking duration [42]. Zeolites with nonspherical shape and sphere-equivalent diameters = 10.3 mm, 9.9 mm, and 7.9 mm were placed in the breast phantom and were used for the estimation of bias induced by the proposed PVE correction method specifically for nonspherical and nonuniform lesions. In particular, we simulated one lesion with nonspherical shape and uniform uptake and two lesions with nonspherical shape and nonuniform uptake.

2.3. Patients. One hundred and thirty-seven oncological patients (46 males, 91 females, age: 28–86 years) were considered, requiring diagnostic investigation involving small lesions (diameter < 4 cm) in different body districts.

All patients signed informed consent. They fasted for twelve hours before the PET-CT exam. ^{18}F-FDG administered dose was prepared based on patient weight considering an amount of 37 MBq for each 10 kg. Administered and residual radioactivity concentrations, administration time, and patient body weight were recorded for each PET-CT study.

108 patients underwent one basal ^{18}F-FDG PET-CT study for tumor staging purpose and they were subjected to radical therapy (surgical intervention or radical radiotherapy); 29 patients underwent two ^{18}F-FDG PET-CT studies, before and after receiving chemotherapy, for therapy monitoring purpose. All PET-CT studies were performed according to the oncological protocol (Section 2.1) and started 60 minutes after the injection. A total of 149 oncological lesions were assessed by ^{18}F-FDG PET-CT images (49 lesions in gastric and gastro-oesophageal regions, 40 lesions in breast, 19 lesions in head and neck regions, and 42 lesions in skeleton).

Histological and therapy-outcome data were considered. Histological data were obtained from surgical intervention of 89 patients, for example, tumour histotype. In particular, for the gastric and gastro-oesophageal lesions, two histotypes were considered: signet ring cell (SRC) carcinaoma and squamous cell (SC) carcinaoma. For the breast lesions, proliferation cell index MiB-1 was provided. Disease-free survival (DFS) data at 24 months after therapy were obtained for 19 patients with cancer in the head and neck regions and treated with radical radiotherapy.

Table 1 describes the characteristics and the available data of the considered patients.

2.4. The PVE Correction Method. The PVE correction method is based on recovery coefficients (RC) derived from PET measured hot-lesion-to-hot-background ratio (L/B_m) and PET measured lesion metabolic volume of the six spheres within the elliptical perspex cylinder.

L/B_m is obtained by the ratio between the PET measured sphere uptake and the PET measured background surrounding the sphere, resulting from the average over several circular regions of interest (4) around the lesion.

RC are plotted as a function of L/B_m and of PET measured sphere metabolic volume.

The proposed PVE correction method acts at a regional level and compensates the lesion uptake underestimation on PET clinical images due to PVE by multiplying it by a proper factor (F) defined as $F = 1/\mathrm{RC}$.

For each lesion detected on the PET clinical images of an oncological patient, F is assigned based upon the PET measured L/B_m and the PET measured lesion metabolic volume.

The PET measured sphere uptake, the PET measured sphere metabolic volume, the PET measured lesion uptake, and the PET measured lesion metabolic volume are all obtained by the an operator-independent technique described as follows.

2.5. The Operator-Independent Technique. An operator-independent technique was developed allowing to obtain an isocontour on that PET image including the maximum lesion/sphere uptake. The isocontour is defined at a definite threshold of the maximum lesion/sphere uptake. Such isocontour defines either the region of interest for the PET measurement of sphere/lesion uptake or the circle-equivalent section of a PET measured sphere/lesion spherical metabolic volume (isocontour volume).

The threshold is chosen by an optimisation procedure such that the PET measured metabolic volumes of spheres match their actual metabolic volumes.

2.6. Optimization of the Operator-Independent Technique. PET-CT independent measurements with the six spheres and the PET-CT DSTE scanner were performed according to

Table 2: Six spheres, one representative measurement: actual diameter, GS radioactivity concentration in the spheres and in the background and the derived L/B_{GS}.

d (mm)	$C_{GS\text{-sphere}}$ (MBq × mL^{-1})	$C_{GS\text{-background}}$ (MBq × mL^{-1})	L/B_{GS}
10	0.07844 ± 0.00666	0.01258 ± 0.000555	6.3 ± 0.65
13	0.07363 ± 0.00555	0.01258 ± 0.000555	5.9 ± 0.56
17	0.06475 ± 0.00222	0.01258 ± 0.000555	5.2 ± 0.35
23	0.06438 ± 0.00185	0.01258 ± 0.000555	5.2 ± 0.34
29	0.05550 ± 0.00111	0.01258 ± 0.000555	4.4 ± 0.27
37	0.05550 ± 0.00037	0.01258 ± 0.000555	4.4 ± 0.26

the oncological protocol (Section 2.1) using an acquisition time of 30 min/scan in order to minimize the noise level on the PET images and considering 2 PET scans at 2 adjacent bed positions (phantom $h = 20$ cm).

For each independent measurement, the spheres were filled with different radioactivity concentrations of ^{18}F-FDG and dipped into the elliptical cylinder filled with a radioactivity concentration of ^{18}F-FDG of 0.01258 MBq × mL^{-1} (background).

PET measured metabolic volumes were calculated on the PET images according to the described operator-independent technique for thresholds at 50, 60, 70, and 80% of the maximum sphere uptake. The percentage differences between the actual sphere diameter and the derived sphere diameter were calculated using a different threshold from each PET measured volume.

The optimal threshold was chosen as the threshold giving the lowest positive percentage differences. This procedure warrants the actual sphere metabolic volume to be represented by the PET measured volume in the best possible way and at the same time allows to exclude background components.

2.7. RC Estimation. PET-CT independent measurements with the six spheres and the PET-CT DSTE scanner were performed as in Section 2.1.

Sphere and background radioactivity concentration obtained with the dose measurement system was regarded as the gold standard (GS), namely the best estimate of the actual radioactivity concentration. L/B_{GS} ranged from 4 to 35 (B_{GS} concentration from 0.0018 MBq × mL^{-1} to 0.024 MBq × mL^{-1}).

As a representative example, Table 2 shows the GS radioactivity concentrations in the spheres ($C_{GS\text{-sphere}}$) and in the background ($C_{GS\text{-background}}$) and the derived L/B_{GS} for one of the measurements.

For all the independent PET-CT measurements, L/B_m was calculated according to the operator-independent technique at the optimal threshold.

RC were calculated as the ratio between L/B_m and L/B_{GS}.

RC curves were obtained by combining RC values as a function of L/B_m and of sphere "isocontour" diameter.

The RC curves were fitted using a three-parameter hyperbolic function.

2.8. Validation of the PVE Correction Method

2.8.1. RC Noise Sensitiveness. The sensitiveness to noise level on the PET images of the method to estimate RC was assessed.

PET-CT measurements were performed with the six spheres with L/B_{GS} ranging from 7 to 10, following oncological protocol but using different acquisition times (2.5 min, 5 min, 10 min, 15 min, and 30 min).

For each sphere, RC was calculated at each acquisition time, and percentage differences of RC over time were obtained.

2.8.2. Residual Errors after PVE Correction. The accuracy of the PVE correction method was assessed by evaluating residual errors after PVE correction.

PET-CT measurements were performed with the three synthetic spherical lesions and with the three synthetic zeolites within the anthropomorphic phantoms, and background was filled with different concentrations of ^{18}F-FDG. L/B_{GS} ranged from 4 to 35 for 31 independent experiments (B_{GS} concentration from 0.005 MBq × mL^{-1} to 0.0012 MBq × mL^{-1}).

Zeolites were prepared as described in [42]. They were soaked into an aqueous solution of ^{18}F-FDG with an actual radioactivity concentration of 0.17 MBq × mL^{-1}. One zeolite was soaked for 15 minutes to simulate a nonspherical but homogeneous tumor. The other two zeolites were soaked only for 5 seconds to simulate nonspherical heterogenous tumors.

Zeolite weights (dry weight before soaking and wet weight after soaking) were measured by means of an analytic balance. Absorbed radioactive solution volume was estimated as the difference between wet and dry weights. Zeolite volume was measured using Archimedes' principle. Radioactivity within zeolites was calculated as radioactivity concentration of the ^{18}F-FDG soaking solution multiplied by the absorbed radioactive solution weight. Radioactivity concentration within each zeolite was calculated as the ratio between radioactivity within zeolite and zeolite volume. Sphere-equivalent diameters were obtained from zeolite volumes.

For each phantom lesion (both spheres and zeolites), lesion optimised "isocontour" volume and L/B_m were measured on PET images.

The PVE-corrected radioactivity concentration within spheres was obtained by multiplying the measured PVE-affected radioactivity concentration by the proper $F = 1/\text{RC}$.

Percentage residual errors, as the differences between the GS and PVE-corrected radioactivity concentration, were calculated.

2.9. Feasibility of the PVE Correction Method. Feasibility of the PVE correction method was assessed by applying the PVE correction to the PET-CT studies of the selected oncological patients.

Qualitative and quantitative assessment was performed under the guide of one expert nuclear medicine physician. Body-weighted standardized uptake value (SUV) was provided and calculated as the tissue radioactivity concentration corrected for the injected activity and body weight of the patient [32]. SUV quantification with PVE correction was performed for all considered lesions (149). During the measurement of L/B_m, for each considered lesion, the nuclear medicine physician was informed not to include any adjacent high uptake organ in the background measurement.

Statistical correlation analysis was performed between SUV (with and without PVE correction) and the histological and therapy outcome data available for the 108 patients subjected to radical therapy.

For the 29 patients subjected to chemotherapy, the EORTC classification of response to treatment was provided [43].

Table 3 briefly describes the kind of analysis performed for the patient groups.

Table 3: Statistical analysis performed for the patient groups.

Patient group	Analysis
Gastro	Correlation between SUV and histological grade (Mann-Whitney test)
Breast	Correlation between SUV and Mib-1 (Mann-Whitney test)
Head-neck	Correlation between SUV and DFS (Log-rank test)
Skeleton	Classification of response to treatment (EORTC evaluation)

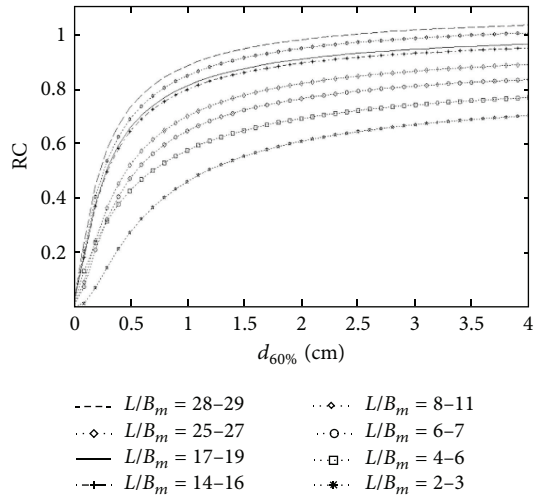


Figure 1: RC curves, threshold = 60%.

3. Results

3.1. Optimization of the Operator-Independent Technique. Table 4 shows, for the PET measurements of the six spheres, the percentage differences (%) between the actual sphere diameter (d) and the sphere diameter derived from "isocontour" volumes at 50, 60, 70, and 80%, averaged over L/B_m.

The optimal threshold giving the lowest positive percentage difference was found to be the threshold at 60%. This value represents a well compromise between a good sample of the lesion actual volume and a good sample of the lesion uptake, minimizing the possibility to include radioactivity background in the sample. Indeed a 50% threshold for the 10 mm sphere gives a negative difference between the actual sphere diameter and the sphere diameter derived from the "isocontour" volume, estimating a lesion volume that is larger than the true volume, thus bringing to include nontumour tissues adjacent to the lesion.

3.2. RC Estimation. Figure 1 shows, for the six spheres, RC curves (8) obtained for L/B_m from 2 to 29, with sphere "isocontour" diameter derived from the optimal threshold (60%) up to 4 cm. The fit was accurate (r square > 0.93) for all RC curves.

Figure 2 shows, for sphere measurements, RC, error bar, and fitting curve for $L/B_m = 3$. The accuracy of the fit can be observed also qualitatively.

Results show that the underestimation of radiotracer uptake due to PVE ranged from 26% up to 70% for the sphere of 10 mm diameter, from −3% up to 32% for the sphere of 37 mm diameter, and from 30 to 2 for L/B_m, respectively. This confirms the severity of the error and the need for PVE correction.

Table 5 shows the percentage differences between the GS ($C_{\text{GS-sphere}}$) and measured ($C_{60\%}$) radioactivity concentrations for the six spheres (one representative PET-CT measurement). $C_{\text{GS-sphere}}$, $C_{60\%}$, and L/B_m are also presented.

3.3. Validation of the PVE Correction Method

3.3.1. RC Noise Sensitiveness. Figure 3 shows, for the sphere with $d = 13$ mm, the percentage difference of RC over the acquisition time (2.5, 5, 10, 15, and 30 min).

RC was found poorly sensitive to the noise level on the PET images for acquisition times in the order of 30 min down to 2.5 min (percentage difference < 5%), proving the noise independency of the method to estimate RC. This guarantees the feasibility of our RC-based PVE correction method for clinical studies of acquisition time from 2.5 min (standard whole-body PET scan/bed) up to 30 min.

3.3.2. Residual Errors after PVE Correction. Figure 4 shows PET-CT representative images of the oncological phantoms used for the validation of the PVE correction method.

Table 4: Six spheres: percentage differences (%) between actual sphere diameter (d) and sphere diameter derived from "isocontour" volumes at 50, 60, 70, and 80%.

d (mm)	$d_{50\%}$ (mm)	% diff	$d_{60\%}$ (mm)	% diff	$d_{70\%}$ (mm)	% diff	$d_{80\%}$ (mm)	% diff
10	12	−20 ± 6.9	9	10 ± 1.8	7	30 ± 8.9	5	50 ± 17.4
13	12	7.7 ± 0.9	10	23.1 ± 2.0	8	38.5 ± 3.2	7	46.2 ± 5.1
17	16	5.9 ± 0.7	13	23.5 ± 0.9	12	29.4 ± 0.9	10	41.2 ± 1.5
23	19	17.4 ± 0.5	17	26.1 ± 0.8	15	34.8 ± 1.4	13	40.9 ± 2.5
29	27	6.9 ± 0.5	25	13.8 ± 0.8	23	20.7 ± 0.9	19	17.4 ± 1.0
37	34	8.1 ± 0.03	32	13.2 ± 0.1	30	18.9 ± 0.4	28	24.3 ± 1.1

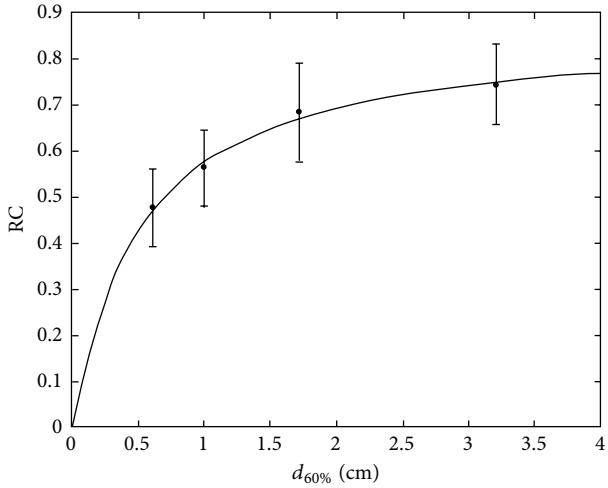


Figure 2: RC curves, threshold = 60%, $L/B_m = 3$.

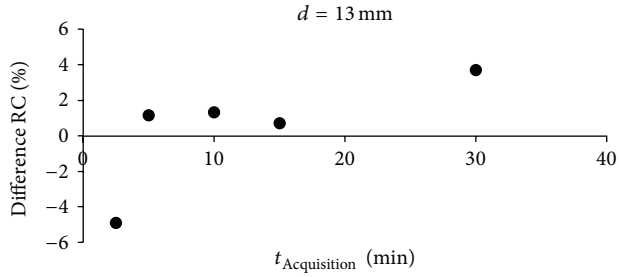


Figure 3: Percentage difference of RC over acquisition time.

Figure 5 shows PET-CT representative images of the oncological nonuniform and nonhomogeneus lesions (the three zeolites) used for the estimation of bias of the PVE correction method for nonspherical lesions. The uniform uptake of the zeolite soaked in the radioactive solution for 15 minutes and the nonuniform uptake of the two zeolites soaked for few seconds can be observed.

$d_{60\%}$ of the spherical lesions ranged from 6 mm up to 12 mm and L/B_m ranged from 8 to 18.

Table 6 shows residual errors (%) after PVE correction, for all the lesions of the validation phantoms as percentage differences between the GS and PVE-corrected radioactivity concentration within the lesions. $C_{\text{GS-sphere}}$, $C_{60\%}$, L/B_m, and actual diameter are also presented.

For lesions with diameter > 1 cm, the PVE correction method was found with an accuracy > 91% in the thorax and breast. The method revealed an accuracy greater than 89% in the brain.

For lesions with diameter < 1 cm, the residual error is of 24%, from an initial error of 70%. Thus, the method allows to recover 76% of radioactivity.

In case of zeolites, the PVE correction method confirms a good accuracy in the uniform lesion (% residual error < 17%). The method is not accurate for nonuniform lesions (zeolites with nonuniform uptake (% residual error > 30%)).

3.4. Feasibility of the PVE Correction Method. For all 149 lesions, it was possible to define the metabolic volume on PET images. 100% of lesions were found to have an L/B_m in the range of L/B_m measured from the spheres and lesion sphere-equivalent diameters in the range of sphere-diameters of RC curves.

97% of lesions were found to have a spherical functional volume; 83% of lesions were found to have a uniform lesion uptake.

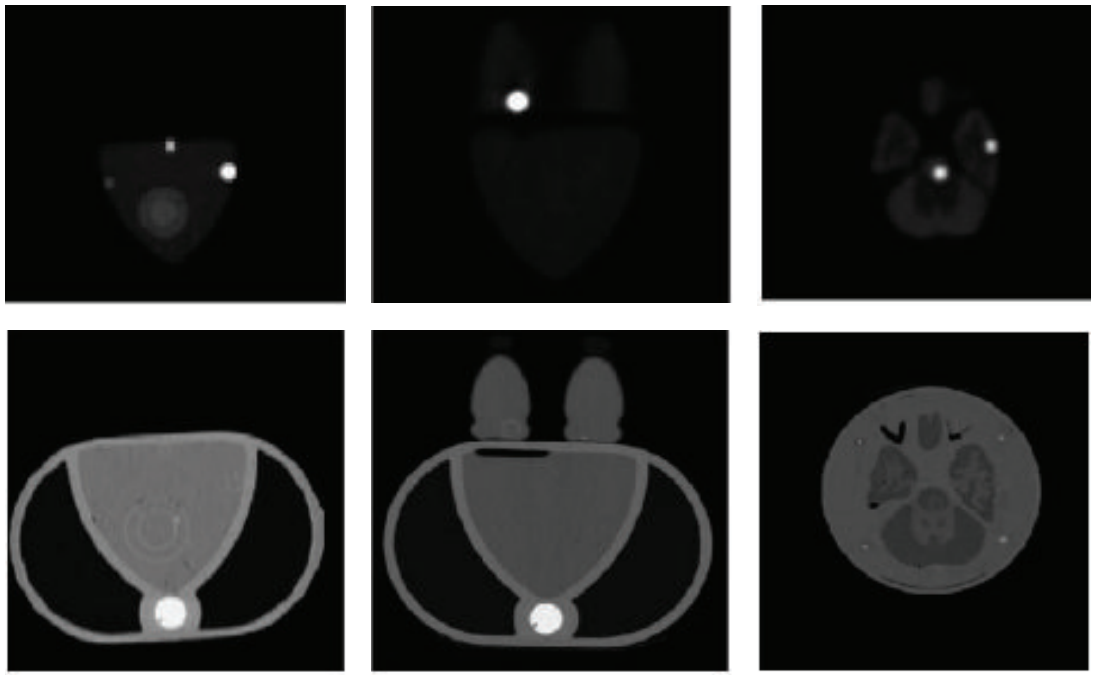
Only for 25% of lesions, the lesion volume was visible on CT images.

PVE correction was found to modify both the value of SUV and of SUV variations during patient followup. After PVE correction, SUV was found to be increased more than 25% in 31% of lesions with a percentage difference between PVE-affected SUV and PVE-corrected SUV up to 120%. SUV variations during followup were also found to be modified by PVE correction of >50% for 67% of lesions and up to 200%.

PVE correction was found to increase the statistical significance of statistical correlation tests (P changed significantly) between SUV and prognostic factors as histopathological indexes (histological grade, cell proliferation index, and therapy outcome indexes), allowing to identify a prognostic value of SUV for the considered cohort of oncological patients. As a consequence, SUV corrected with the proposed PVE was able to stratify different groups of patients.

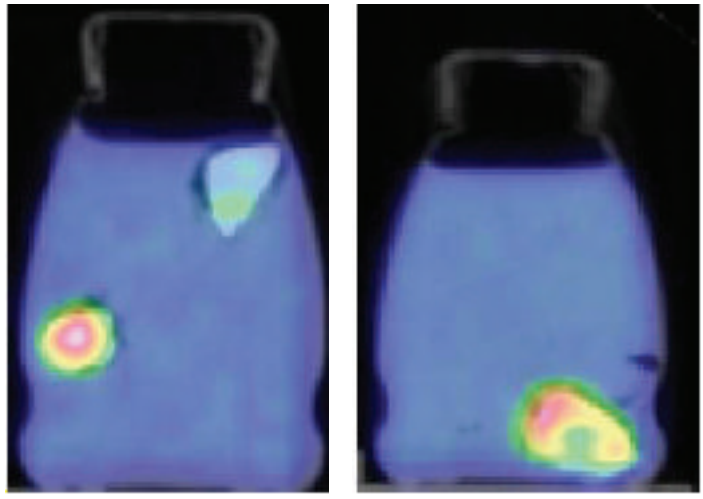
Table 7 summarizes the main results of the impact of PVE correction on the considered correlation studies in the oncological patients.

PVE was also found to have an impact on the classification of patient response to treatment based on EORTC recommendations. Noteworthy, PVE correction changed the response classification of 3 of the 19 patients with bone metastasis (EORTC response classification: partial metabolic



(a) (b) (c)

Figure 4: PET-CT images for (a) thorax phantom, (b) breast phantom, and (c) brain phantom.



(a) (b)

Figure 5: PET-CT images for the oncological nonspherical lesions (zeolites).

response, PMR; stable metabolic disease, SMD; progressive metabolic disease, PMD). In particular, one patient changed from PMR to SMD, one patient from SMD to PMD, and one patient from SMD to PMR.

Table 8 summarizes the main results on the impact of PVE correction on the considered therapy response in the oncological patients. Applying PVC, the average SUV values increased more than 45%, proving the need for correction.

4. Discussion and Conclusions

Two aspects mainly characterise the proposed PVE correction method and differentiate it from other RC-based PVE correction procedures.

(1) The Clinical Approach for the Design of PVE Correction. The approach for the design of the PVE correction method

Table 5: Six spheres, one representative measurement: % difference between GS radioactivity concentration and measured radioactivity concentration and the derived L/B_m.

d (mm)	$C_{\text{GS-sphere}}$ (MBq × mL^{-1})	$C_{60\%}$ (MBq × mL^{-1})	% diff	L/B_m
10	0.078449 ± 0.00666	0.02331 ± 0.0037	70.3 ± 12.7	2.0 ± 0.4
13	0.07363 ± 0.00555	0.03774 ± 0.0074	42.5 ± 6.4	3.3 ± 0.5
17	0.06475 ± 0.00222	0.03959 ± 0.0037	39.0 ± 4.9	3.5 ± 0.4
23	0.06438 ± 0.00185	0.04033 ± 0.0074	37.4 ± 5.8	3.5 ± 0.6
29	0.5550 ± 0.00111	0.03330 ± 0.0037	40.0 ± 6.0	2.9 ± 0.4
37	0.5550 ± 0.00037	0.037774 ± 0.0037	31.9 ± 3.8	3.3 ± 0.4

Table 6: Validation phantoms: % residual errors after PVE correction.

Phantom	d (mm)	$C_{\text{GS-sphere}}$ (MBq × mL^{-1})	$C_{60\%}$ (MBq × mL^{-1})	% res	L/B_m
Thorax	9.8	0.8214	0.5883 ± 0.1036	24 ± 5.0	17.8
	12.3	0.3626	0.3293 ± 0.0333	9.8 ± 1.0	8.9
	12.3	0.6993	0.666 ± 0.0629	4.9 ± 0.5	16.8
	15.6	0.9065	0.8473 ± 0.0481	6.7 ± 0.4	30
	15.6	0.46028	0.45917 ± 0.06845	0.3 ± 0.04	9.2
Breast	9.8	0.0962	0.0777 ± 0.0074	16.6 ± 2.4	4.9
	12.3	0.1184	0.1073 ± 0.0148	9.3 ± 1.4	13.3
	15.6	0.2479	0.2590 ± 0.0148	−4.5 ± 0.3	8.3
	15.6	0.4884	0.4662 ± 0.0592	4.7 ± 0.8	20.3
	13.3*	0.0048*	0.0056 ± 0.00001*	−16.6 ± 2.9*	3.1*
	10.3**	0.0100**	0.0070 ± 0.00001**	30.0 ± 5.1**	2.8**
	9.9**	0.0128**	0.0049 ± 0.00007**	62.7 ± 11.1**	2.4**
Brain	9.8	0.5402	0.4070 ± 0.0666	24 ± 4.1	8.8
	12.3	0.4555	0.4033 ± 0.1184	11.4 ± 3.3	12.8
	12.3	0.4144	0.3663 ± 0.0703	11.2 ± 2.2	11.9
	15.6	0.3737	0.3552 ± 0.037	4.8 ± 0.5	14.2

* Represents zeolites with uniform uptake; ** represents zeolites with nonuniform uptake.

moved from considering information from real clinical PET-CT studies, which is always available. The purpose of the work was the development of a PVE correction method allowing quantification of glucose metabolism in tumour cells with the primary objective to be easily implementable and usable in a clinical environment. Several studies showed that PET-detected oncological lesions are not always visible on CT images [27, 28] and this has also been confirmed by our nuclear medicine physicians. Thus, information which is always available is represented by data measurable on PET images, for instance, PET L/B_m and PET measured lesion volume. Following this consideration, our PVE correction method was based on RC factors derived from PET measurements of hot spheres in hot background, simulating lesions in body tissue under PET study. RC curves were thus obtained from PET L/B_m and from PET measured sphere volume, and not from actual L/B or from actual sphere volume, as in the case of all the rest of RC-based PVE correction methods. This strategy allows to overcome the problem of being aware of the actual L/B and the actual lesion volume.

(2) The Technique for the PET Measurement of L/B_m and Lesion Metabolic Volume. A technique allowing PET measurement of both lesion uptake (and thus PET L/B_m) and lesion metabolic volume was developed, based on a technique of threshold isocontours. Such technique is not aimed at extracting the actual lesion metabolic volume but it is able to provide a PET measurement of a lesion metabolic volume (the "isocontour" volume) which is strongly dependant on the actual metabolic volume (the larger the actual metabolic volume, the "larger" the isocontour volume), however, being independent of the operator. The optimal volume, that is the volume defined by the optimal threshold, is the best metabolic volume matching the actual lesion volume and excluding at the same time the background uptake [30]. In this work, we present results of an optimal threshold relative to a 60% threshold. From the current literature [16], among the studied thresholds (in the range of 50–80%), the "isocontour" volume derived from a 60% threshold has been shown to represent a well compromise between a good sample of the lesion actual volume and a good sample of the lesion uptake, minimizing the possibility to include radioactivity background in the sample [36]. As assessed also by Krak et al. [30], a 50% threshold leads to include nontumour tissues, and this increases the possibility to include within the lesion normal high uptake in localized areas (e.g., liver, heart, and inflammatory tissues) that could be adjacent to the lesion. Furthermore, a threshold greater than 75% shows

Table 7: Oncological patients: results of SUV quantification with PVE correction on correlation studies between SUV and prognostic factors (P is the result of statistical tests).

Patient group	Lesion d (cm)	SUV (g/cc)	P	PVE-corrected SUV (g/cc)	P after PVE correction
Gastro	2.15 ± 1.17 (0.99–6.25)	3.27 ± 1.22 (SRC) 7.93 ± 5.01 (SC)	$P > 0.05$	5.57 ± 3.22 (SRC) 9.90 ± 1.91 (SC)	$P < 0.05$
Breast	1.57 ± 0.5 (1.1–3.2)	2.28 ± 1.02 (Mib+) 7.64 ± 6.08 (Mib−)	$P > 0.05$	4.52 ± 2.92 (Mib+) 9.30 ± 7.40 (Mib−)	$P < 0.05$
Head-neck	1.52 ± 0.5	<10.8 (lymph−) >10.8 (lymph+)	$P > 0.05$	<13.3 (lymph−) >13.3 (lymp+)	$P < 0.05$

Table 8: Oncological patients: results of SUV quantification with PVE correction on therapy response classification (EORTC). I means pretreatment SUV; II means posttreatment SUV.

Patient group	Lesion d (cm)	SUV (g/cc)	PVE-corrected SUV (g/cc)	SUV% difference
Skeleton	1.55 ± 0.5 (0.9–3.4)	4.7 ± 1.9 (I) 4.2 ± 1.9 (II)	6.6 ± 2.3 5.8 ± 2.6	46.4 ± 29.7 45.9 ± 28.7

"less reproducibility" than lower thresholds (the difference in lesion metabolic volumes measured by PET at consecutive days is >50% for a threshold of 75% and <25% for a threshold of 60%, resp.)—both in terms of lesion metabolic volume and SUV [30]. Our results, relative to a threshold of 60%, show that the proposed PVE correction technique is accurate for lesion diameter > 1 cm, considering that previous studies on SUV reproducibility from oncological patients showed SUV percentage errors up to 17% [40].

The advantages of our approach are as follows.

(A1) Consistency. There is a full consistency between the direct procedure of obtaining RC from PET measurements with hot spheres in hot background and the inverse procedure that applies $F = 1/\text{RC}$ factors for PVE correction of PET-detected oncological lesions. This allows for the clinical implementation of the PVE correction method to real oncological studies.

(A2) Operator Independency. The operator independency of the threshold technique for the PET measurement of quantitative parameters (PET L/B_m and PET measured "isocontour" volume) required by our PVE correction method guarantees reproducible measurements. Furthermore, the use of metabolic volumes defined by a threshold technique in clinical follow-up studies is suitable to show the effect of metabolic change due to therapy. This instead is not true for the CT detected anatomical volumes that may result unmodified at followup. As a result, our PVE correction method is more feasible for quantification of follow-up studies than alternative strategies based on actual lesion volume (e.g., [30]).

(A3) Applicability. The PVE correction method can be applied for any PET-CT scanner in a simple manner, given that it lies upon experimental measurements easy to be performed with a PET scanner and a standard phantom of easy availability. Anthropomorphic phantoms miming oncological lesions in specific regions of the human body could be used to extract RC factors more accurately for specific body regions (e.g., brain) or for a specific radiotracer (e.g., ^{11}C-choline).

The disadvantages of our approach are as follows.

(D1) Local Correction. As for all RC-based PVE correction methods, our method applies PVE correction only at a regional level, on the PET images. This means that the lesion uptake is corrected using some information (PET L/B_m and PET measured lesion volume) of that particular region. As opposed to PVE correction methods which process PET images for the creation of PET corrected images (e.g., [20, 44, 45]), our method requires the correction to be applied separately to different lesions.

(D2) Noise Dependency. One of the drawbacks of the threshold technique for the PET measurement of lesion uptake and lesion metabolic volume is that the resulting defined region can be dependent on the noise present in the PET images. The threshold value for the radioactivity concentration (thus the corresponding isocontour) is dependent on the maximum value of the lesion uptake, being the threshold defined as a percentage of this value. Optimisation strategies based on smoothing or averaging techniques over the maximum could be applied [16] in order to reduce this effect.

(D3) Lesion Roundness and Uniformity. RC values have been obtained for hot spheres miming spherical and uniform lesions. This limits the application of the proposed PVE correction to oncological lesions which can be assumed to be spherical and with a uniform uptake. Preliminary results from our simulations on lesions with nonuniform uptake (zeolites) indicated that the PVE correction method is very sensitive to nonspherical and nonuniform lesions, while it can work well in nonspherical but uniform lesions, consistently with some results from Monte Carlo simulations proving the suitability of RC-based PVE correction for nonspherical lesions (e.g., [46, 47]). Considering our PET-CT clinical studies, we found that this occurs for a limited number of cases (96% of lesions were spherical and 80% with a uniform uptake). For those lesions that have hypometabolic characteristics (e.g., low grade tumour in the cerebral white matter), other PVE correction methods (e.g., based on image-guided

segmentation or preprocessing) can be applied (e.g., [19, 20, 27, 48–51]).

However, for lesions that cannot be approximated to spheres, our PVE correction approach should be used carefully and it needs optimization (e.g., new RC from nonspherical objects) as well as validation (e.g., with anthropomorphic phantoms including nonspherical objects). The same care in the use of the considered RC-based PVE correction must be applied to heterogeneous lesions. A recent study that focused on the impact of PVE correction on tumor heterogeneity suggests in this case the use of local image deconvolution approach with expectation maximization and spatially variant point spread function (e.g., [52]).

(D4) Background Uniformity. An important problem in practice is that the background is usually not uniform. High uptake in localized areas (e.g., liver, heart, and inflammatory tissues) could be present in regions adjacent to the lesion. The use of a single threshold to segment metabolic lesion volume, as proposed in our method, could include these normal tissues. In the latter case, manual intervention could be needed in order to exclude background tissues, thus making our method more observer dependent.

We have developed, implemented, and assessed a method for PVE correction of oncological lesions in PET clinical studies, based on RC factors and PET L/B_m and PET measured lesion metabolic volume.

Phantom measurements proved that PVE strongly affects lesion quantification (up to 70%) and needs to be corrected. Consistently with previous findings [26, 27, 53], we found this effect to be increasing when sphere volume and L/B_m decrease.

Measured RC curves allowed PVE correction to be applied to lesions of diameter up to 40 mm and for PET L/B_m from 2 to 30, answering the need of PET quantification for a large variety of oncological lesions.

An operator independent technique was developed and optimised for the PET measurement of lesion uptake and of lesion metabolic volume. The technique is based on a threshold that defined an isocontour with respect to the maximum uptake on PET image. Such isocontour defines either the region of interest for the PET measurement of sphere/lesion uptake or the circle-equivalent section of a PET measured sphere/lesion spherical metabolic volume (isocontour volume).

Our residual errors obtained after the application of the PVE correction method to anthropomorphic oncological phantoms, compared with the errors on the measurement of SUV (12%-13%) obtained by Krak et al. [30], proved that our method is accurate (>89%) in clinical realistic conditions for lesion diameter > 1 cm and it is able to recover 76% of radioactivity for lesions diameter < 1 cm in a consistent way with the errors on the measurement of lesion metabolic volume (>23%) estimated by Krak et al. Other methods based on postreconstruction iterative techniques [44], iterative deconvolution [43], image segmentation [18], or multiresolution approach [20] implemented for PVE correction mainly in neurodegenerative diseases show an accuracy up to 98% for lesion diameter > 1 cm and up to 86% for lesion diameter < 1 cm. However, these methods require images to be processed by dedicated software and are more complex to be implemented in clinical routine than RC-based methods, as previously discussed ((*A3) Applicability*) and also commented by Soret et al. [16].

Patient studies showed that the proposed PVE correction method is suitable and feasible in a clinical environment. L/B_m and "optimal" isocontour volume at 60% threshold of the maximum were used to obtain proper RC in order to correct the PVE-affected SUV for all considered patient lesions. The quantitative analysis was performed under the guide of an expert nuclear medicine physician. We found that at least 80% of selected lesions met the requirements of roundness and uniformity for an accurate use of the proposed PVE correction method. As expected, only few lesions were clearly visible on CT images, confirming the need to define lesion volume from PET images.

Considerations on SUV increase or decrease during patient followup as an effect of a therapy is beyond the purpose of this paper. However, our results suggest that the use of PVE correction can be fruitful in staging oncological disease and in monitoring oncological disease progression.

Our results suggest that the PVE correction has to be applied if SUV is used to stratify patients on the basis of an SUV cut-off value and/or to classify lesion metabolic response by means of SUV variations during followup. When SUV is considered for diagnostic purposes (i.e., an absolute cut-off value of SUV to differentiate benign from malignant tumor), the cutoff should be defined by accounting for PVE; otherwise it could be inappropriate.

In conclusion, in this work, we developed a method for PVE correction tailored for clinical application to PET-CT oncological studies. Our method overcomes the problem of considering actual L/B and actual lesion volume, being grounded in RC curves determined as functions of PET L/B_m and measured lesion volume, both estimated by an optimized and validated operator-independent technique. The proposed PVE correction method was applied to clinical oncological ^{18}F-FDG PET-CT studies showing to have an impact on the metabolic assessment of lesions.

Acknowledgment

The authors thank all the staff of the Nuclear Medicine Department of H San Raffaele, Milan, Italy, for their collaboration during quantitative clinical PET-CT studies.

References

[1] D. A. Mankoff and W. B. Eubank, "Current and future use of positron emission tomography (PET) in breast cancer," *Journal of Mammary Gland Biology and Neoplasia*, vol. 11, no. 2, pp. 125–136, 2006.

[2] C. Plathow and W. A. Weber, "Tumor cell metabolism imaging," *Journal of Nuclear Medicine*, vol. 49, no. 6, pp. 43S–63S, 2008.

[3] M. Lapela, A. Eigtved, S. Jyrkkiö et al., "Experience in qualitative and quantitative FDG PET in follow-up of patients with suspected recurrence from head and neck cancer," *European Journal of Cancer*, vol. 36, no. 7, pp. 858–867, 2000.

[4] E. L. Rosen, W. B. Eubank, and D. A. Mankoff, "FDG PET, PET/CT, and breast cancer imaging," *Radiographics*, vol. 27, pp. S215–S229, 2007.

[5] F. Castell and G. J. R. Cook, "Quantitative techniques in ^{18}FDG PET scanning in oncology," *British Journal of Cancer*, vol. 98, no. 10, pp. 1597–1601, 2008.

[6] K. Strobel, U. E. Exner, K. D. M. Stumpe et al., "The additional value of CT images interpretation in the differential diagnosis of benign vs. malignant primary bone lesions with ^{18}F-FDG-PET/CT," *European Journal of Nuclear Medicine and Molecular Imaging*, vol. 35, no. 11, pp. 2000–2008, 2008.

[7] H. Hoshikawa, T. Mitani, Y. Nishiyama, Y. Yamamoto, M. Ohkawa, and N. Mori, "Evaluation of the therapeutic effects and recurrence for head and neck cancer after chemoradiotherapy by FDG-PET," *Auris Nasus Larynx*, vol. 36, no. 2, pp. 192–198, 2009.

[8] L. K. Shankar, J. M. Hoffman, S. Bacharach et al., "Consensus recommendations for the use of ^{18}F-FDG PET as an indicator of therapeutic response in patients in national cancer institute trials," *Journal of Nuclear Medicine*, vol. 47, no. 6, pp. 1059–1066, 2006.

[9] A. Stahl, K. Ott, M. Schwaiger, and W. A. Weber, "Comparison of different SUV-based methods for monitoring cytotoxic therapy with FDG PET," *European Journal of Nuclear Medicine and Molecular Imaging*, vol. 31, no. 11, pp. 1471–1479, 2004.

[10] A. Gil-Rendo, F. Martínez-Regueira, G. Zornoza, M. J. García-Velloso, C. Beorlegui, and N. Rodriguez-Spiteri, "Association between [^{18}F] fluorodeoxyglucose uptake and prognostic parameters in breast cancer," *British Journal of Surgery*, vol. 96, no. 2, pp. 166–170, 2009.

[11] M. Schmidt, E. Bollschweiler, M. Dietlein et al., "Mean and maximum standardized uptake values in [^{18}F]FDG-PET for assessment of histopathological response in oesophageal squamous cell carcinoma or adenocarcinoma after radiochemotherapy," *European Journal of Nuclear Medicine and Molecular Imaging*, vol. 36, no. 5, pp. 735–744, 2009.

[12] A. Berriolo-Riedinger, C. Touzery, J.-M. Riedinger et al., "[^{18}F]FDG-PET predicts complete pathological response of breast cancer to neoadjuvant chemotherapy," *European Journal of Nuclear Medicine and Molecular Imaging*, vol. 34, no. 12, pp. 1915–1924, 2007.

[13] N. Avril, M. Menzel, J. Dose et al., "Glucose metabolism of breast cancer assessed by ^{18}F-FDG PET: histologic and immunohistochemical tissue analysis," *Journal of Nuclear Medicine*, vol. 42, no. 1, pp. 9–16, 2001.

[14] E. J. Hoffman, P. D. Cutler, T. M. Guerrero, W. M. Digby, and J. C. Mazziotta, "Assessment of accuracy of PET utilizing a 3-D phantom to simulate the activity distribution of [^{18}F]fluorodeoxyglucose uptake in the human brain," *Journal of Cerebral Blood Flow and Metabolism*, vol. 11, no. 2, pp. A17–A25, 1991.

[15] R. M. Kessler, J. R. Ellis Jr., and M. Eden, "Analysis of emission tomographic scan data: Limitations imposed by resolution and background," *Journal of Computer Assisted Tomography*, vol. 8, no. 3, pp. 514–522, 1984.

[16] M. Soret, S. L. Bacharach, and I. Buvat, "Partial-volume effect in PET tumor imaging," *Journal of Nuclear Medicine*, vol. 48, no. 6, pp. 932–945, 2007.

[17] E. J. Hoffman, S. C. Huang, and M. E. Phelps, "Quantitation in positron emission computer tomography: effect of object size," *Journal of Computer Assisted Tomography*, vol. 3, no. 3, pp. 299–308, 1978.

[18] O. G. Rousset, Y. Ma, and A. C. Evans, "Correction for partial volume effects in PET: principle and validation," *Journal of Nuclear Medicine*, vol. 39, no. 5, pp. 904–911, 1998.

[19] H. W. Müller-Gärtner, J. M. Links, J. L. Prince et al., "Measurement of radiotracer concentration in brain gray matter using positron emission tomography: MRI-based correction for partial volume effects," *Journal of Cerebral Blood Flow and Metabolism*, vol. 12, no. 4, pp. 571–583, 1992.

[20] N. Boussion, M. Hatt, F. Lamare et al., "A multiresolution image based approach for correction of partial volume effects in emission tomography," *Physics in Medicine and Biology*, vol. 51, no. 7, pp. 1857–1876, 2006.

[21] H. Vesselle, R. A. Schmidt, J. M. Pugsley et al., "Lung cancer proliferation correlates with [F-18]fluorodeoxyglucose uptake by positron emission tomography," *Clinical Cancer Research*, vol. 6, no. 10, pp. 3837–3844, 2000.

[22] A. van Baardwijk, B. G. Baumert, G. Bosmans et al., "The current status of FDG-PET in tumour volume definition in radiotherapy treatment planning," *Cancer Treatment Reviews*, vol. 32, no. 4, pp. 245–260, 2006.

[23] J. Yu, X. Li, L. Xing et al., "Comparison of tumor volumes as determined by pathologic examination and FDG-PET/CT images of non-small-cell lung cancer: a pilot study," *International Journal of Radiation Oncology Biology Physics*, vol. 75, no. 5, pp. 1468–1474, 2009.

[24] W. A. Weber, S. I. Ziegler, R. Thödtmann, A.-R. Hanauske, and M. Schwaiger, "Reproducibility of metabolic measurements in malignant tumors using FDG PET," *Journal of Nuclear Medicine*, vol. 40, no. 11, pp. 1771–1777, 1999.

[25] Y. E. Erdi, O. Mawlawi, S. M. Larson et al., "Segmentation of lung lesion volume by adaptive positron emission tomography image thresholding," *Cancer*, vol. 80, no. 12, supplement, pp. 2505–2509, 1997.

[26] S. M. Srinivas, T. Dhurairaj, S. Basu, G. Bural, S. Surti, and A. Alavi, "A recovery coefficient method for partial volume correction of PET images," *Annals of Nuclear Medicine*, vol. 23, no. 4, pp. 341–348, 2009.

[27] W. Jentzen, L. Freudenberg, E. G. Eising, M. Heinze, W. Brandau, and A. Bockisch, "Segmentation of PET volumes by iterative image thresholding," *Journal of Nuclear Medicine*, vol. 48, no. 1, pp. 108–114, 2007.

[28] D. A. Schinagl, J. H. Kaanders, and W. J. Oyen, "From anatomical to biological target volumes: the role of PET in radiation treatment planning," *Cancer Imaging*, vol. 6, pp. S107–S116, 2006.

[29] D. C. Crawford, M. A. Flower, B. E. Pratt et al., "Thyroid volume measurement in thyrotoxic patients: comparison between ultrasonography and iodine-124 positron emission tomography," *European Journal of Nuclear Medicine*, vol. 24, no. 12, pp. 1470–1478, 1997.

[30] N. C. Krak, R. Boellaard, O. S. Hoekstra, J. W. R. Twisk, C. J. Hoekstra, and A. A. Lammertsma, "Effects of ROI definition and reconstruction method on quantitative outcome and applicability in a response monitoring trial," *European Journal of Nuclear Medicine and Molecular Imaging*, vol. 32, no. 3, pp. 294–301, 2005.

[31] F. Gallivanone, A. Stefano, E. Grosso et al., "PVE correction in PET-CT whole-body oncological studies from PVE-affected images images," *IEEE Transactions on Nuclear Science*, vol. 58, no. 3, pp. 736–747, 2011.

[32] M. M. Graham, L. M. Peterson, and R. M. Hayward, "Comparison of simplified quantitative analyses of FDG uptake," *Nuclear Medicine and Biology*, vol. 27, no. 7, pp. 647–655, 2000.

[33] N. Avril, S. Bense, S. I. Ziegler et al., "Breast imaging with fluorine-18-FDG PET: quantitative image analysis," *Journal of Nuclear Medicine*, vol. 38, no. 8, pp. 1186–1191, 1997.

[34] A. Dimitrakopoulou-Strauss, L. G. Strauss, T. Heichel et al., "The role of quantitative ^{18}F-FDG PET studies for the differentiation of malignant and benign bone lesions," *Journal of Nuclear Medicine*, vol. 43, no. 4, pp. 510–518, 2002.

[35] I. C. Smith, A. E. Welch, A. W. Hutcheon et al., "Positron emission tomography using [^{18}F]-fluorodeoxy-D-glucose to predict the pathologic response of breast cancer to primary chemotherapy," *Journal of Clinical Oncology*, vol. 18, no. 8, pp. 1676–1688, 2000.

[36] J. R. Lee, M. T. Madsen, D. Bushnel, and Y. Menda, "A threshold method to improve standardized uptake value reproducibility," *Nuclear Medicine Communications*, vol. 21, no. 7, pp. 685–690, 2000.

[37] A. C. Kole, O. E. Nieweg, and J. Pruim, "Standardized uptake value and quantification of metabolism for breast cancer imaging with FDG and L-[1-11C]Tyrosine PET," *Journal of Nuclear Medicine*, vol. 38, no. 5, pp. 692–696, 1997.

[38] Y. Nakamoto, K. R. Zasadny, H. Minn, and R. L. Wahl, "Reproducibility of common semi-quantitative parameters for evaluating lung cancer glucose metabolism with positron emission tomography using 2-deoxy-2-[^{18}F]fluoro-D-glucose," *Molecular Imaging and Biology*, vol. 4, no. 2, pp. 171–178, 2002.

[39] R. Boellaard, N. C. Krak, O. S. Hoekstra, and A. A. Lammertsma, "Effects of noise, image resolution, and ROI definition on the accuracy of standard uptake values: a simulation study," *Journal of Nuclear Medicine*, vol. 45, no. 9, pp. 1519–1527, 2004.

[40] M. Teräs, T. Tolvanen, J. J. Johansson, J. J. Williams, and J. Knuuti, "Performance of the new generation of whole-body PET/CT scanners: Discovery STE and Discovery VCT," *European Journal of Nuclear Medicine and Molecular Imaging*, vol. 34, no. 10, pp. 1683–1692, 2007.

[41] M. E. Daube-Witherspoon, J. S. Karp, M. E. Casey et al., "PET performance measurements using the NEMA NU 2-2001 standard," *Journal of Nuclear Medicine*, vol. 43, no. 10, pp. 1398–1409, 2002.

[42] F. Zito, E. De Bernardi, C. Soffientini, C. Canzi, R. Casati, and P. Gerundini, "The use of zeolites to generate PET phantoms for the validation of quantification strategies in oncology," *Medical Physics*, vol. 39, no. 9, pp. 5353–5361, 2012.

[43] H. Young, R. Baum, U. Cremerius et al., "Measurement of clinical and subclinical tumour response using [^{18}F]- fluorodeoxyglucose and positron emission tomography: review and 1999 EORTC recommendations," *European Journal of Cancer*, vol. 35, no. 13, pp. 1773–1782, 1999.

[44] N. Boussion, C. Cheze-Le Rest, M. Hatt, and D. Visvikis, "Incorporation of wavelet-based denoising in iterative deconvolution for partial volume correction in whole-body PET imaging," *European Journal of Nuclear Medicine and Molecular Imaging*, vol. 36, no. 7, pp. 1064–1075, 2009.

[45] B.-K. Teo, Y. Seo, S. L. Bacharach et al., "Partial-volume correction in PET: validation of an iterative postreconstruction method with phantom and patient data," *Journal of Nuclear Medicine*, vol. 48, no. 5, pp. 802–810, 2007.

[46] P. Tylski, S. Stute, N. Grotus et al., "Comparative assessment of methods for estimating tumor volume and standardized uptake value in ^{18}F-FDG PET," *Journal of Nuclear Medicine*, vol. 51, no. 2, pp. 268–276, 2010.

[47] I. Castiglioni, G. Rizzo, A. Panzacchi, M. C. Gilardi, and F. Fazio, "A MC-based PV correction method for PET/CT oncological studies," in *Proceedings of the IEEE Nuclear Science Symposium and Medical Imaging Conference Record*, pp. 11–153, 2005.

[48] O. G. Rousset, A. Rahmim, A. Alavi, and H. Zaidi, "Partial volume correction strategies in PET," *PET Clinics*, vol. 2, no. 2, pp. 235–249, 2007.

[49] C.-H. Chen, R. F. Muzic Jr., A. D. Nelson, and L. P. Adler, "Simultaneous recovery of size and radioactivity concentration of small spheroids with PET data," *Journal of Nuclear Medicine*, vol. 40, no. 1, pp. 118–130, 1999.

[50] K. Baete, J. Nuyts, K. V. Laere et al., "Evaluation of anatomy based reconstruction for partial volume correction in brain FDG-PET," *NeuroImage*, vol. 23, no. 1, pp. 305–317, 2004.

[51] D. Strul and B. Bendriem, "Robustness of anatomically guided pixel-by-pixel algorithms for partial volume effect correction in positron emission tomography," *Journal of Cerebral Blood Flow and Metabolism*, vol. 19, no. 5, pp. 547–559, 1999.

[52] D. L. Barbee, R. T. Flynn, J. E. Holden, R. J. Nickles, and R. Jeraj, "A method for partial volume correction of PET-imaged tumor heterogeneity using expectation maximization with a spatially varying point spread function," *Physics in medicine and biology*, vol. 55, no. 1, pp. 221–236, 2010.

[53] L. Geworski, B. O. Knoop, M. L. de Cabrejas, W. H. Knapp, and D. L. Munz, "Recovery correction for quantitation in emission tomography: a feasibility study," *European Journal of Nuclear Medicine*, vol. 27, no. 2, pp. 161–169, 2000.