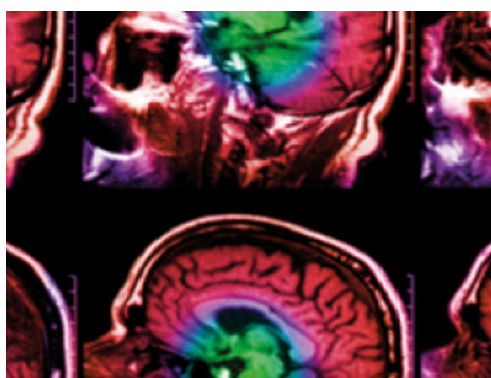


PAPER

Mid-sagittal plane detection for advanced physiological measurements in brain scans

To cite this article: Francesca Bertacchini *et al* 2019 *Physiol. Meas.* **40** 115009

View the [article online](#) for updates and enhancements.



IPEM | IOP

Series in Physics and Engineering in Medicine and Biology


Your publishing choice in medical physics,
biomedical engineering and related subjects.

Start exploring the collection—download the
first chapter of every title for free.



PAPER

Mid-sagittal plane detection for advanced physiological measurements in brain scans

RECEIVED
7 June 2019REVISED
1 October 2019ACCEPTED FOR PUBLICATION
18 October 2019PUBLISHED
3 December 2019Francesca Bertacchini¹, Rossella Rizzo^{2,4,5} , Eleonora Bilotta², Pietro Pantano², Angela Luca³, Alessandro Mazzuca³, Antonio Lopez³ and for the Alzheimer's Disease Neuroimaging Initiative⁶¹ Evolutionary System Group, Department of Mechanical, Energy and Management Engineering, University of Calabria, Rende (CS), Italy² Evolutionary System Group, Department of Physics, University of Calabria, Rende (CS), Italy³ Radiological Unit, Cetraro Hospital, Cetraro (CS), Italy⁴ Postal address: Physics Department, Cubo 17B, University of Calabria, 87036, Arcavacata di Rende (CS), Italy⁵ Author to whom any correspondence should be addressed.E-mail: rossella.rizzo@unical.it and rossella.rizzo3108@gmail.com**Keywords:** image segmentation, *k*-means algorithm, machine learning, magnetic resonance imaging, mid-sagittal plane**Abstract**

Objective: The process of diagnosing many neurodegenerative diseases, such as Parkinson's and progressive supranuclear palsy, involves the study of brain magnetic resonance imaging (MRI) scans in order to identify and locate morphological markers that can highlight the health status of the subject. A fundamental step in the pre-processing and analysis of MRI scans is the identification of the mid-sagittal plane, which corresponds to the mid-brain and allows a coordinate reference system for the whole MRI scan set. *Approach:* To improve the identification of the mid-sagittal plane we have developed an algorithm in Matlab[®] based on the *k*-means clustering function. The results have been compared with the evaluation of four experts who manually identified the mid-sagittal plane and whose performances have been combined with a cognitive decisional algorithm in order to define a gold standard. *Main results:* The comparison provided a mean percentage error of 1.84%. To further refine the automatic procedure we trained a machine learning system using the results from the proposed algorithm and the gold standard. We tested this machine learning system and obtained results comparable to medical raters with a mean absolute error of 1.86 slices. *Significance:* The system is promising and could be directly incorporated into broader diagnostic support systems.

1. Introduction

Magnetic resonance imaging (MRI) can provide valuable information for the detection of degenerative diseases, not just qualitatively but even measurement of volumes, areas and distances between different sections, especially when magnitudes vary, due to the presence of severe deformation. In these cases, one of the main problems is the identification of the optimal slice on which to make these measurements. In this framework, identification of the mid-sagittal plane (MSP) in brain MRI scans is crucial for detecting many of the most important neurodegenerative diseases such as Parkinson's disease (PD) (Nigro *et al* 2014), Huntington's disease (HD) (Di Paola *et al* 2010, 2012), multiple sclerosis (MS) (Bilotta *et al* 2010, 2012, Cerasa *et al* 2012) and Alzheimer's disease (AD) (Di Paola *et al* 2015). In this paper, we present a fully automated method for identifying the MSP in brain MRI scans of subjects with progressive supranuclear palsy (PSP), PD, MS and AD as well as healthy control subjects.

⁶ Some of the data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (<http://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/wpcontent/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

PD, for example, presents with a variety of neurological malfunctions resulting from pyramidal cerebellar, vegetative and cognitive degeneration. The disease affects the nigro-striatal pyramid and involves the cerebellum and deep cerebral structures, but also neuronal degeneration in the neo-striatum. Accurate early diagnosis of PD is important for both therapeutic purposes—to target therapy more precisely to the various symptoms—and in terms of prognosis. However, although advanced diagnostic techniques have recently been developed for PD (Oba *et al* 2005, Quattrone *et al* 2008) it suffers from a lack of universally accepted diagnostic criteria, making it difficult to distinguish; therefore PD is characterized by a high rate of misdiagnosis (Litvan *et al* 1996). Structural MRI is routinely used to detect early signs of PD, from hyper-intensity of the lateral edge of the putamen and atrophy of the brainstem; cross-shaped hyper-intensity of the bridge and middle cerebral peduncles (Bhattacharya *et al* 2002) is also an indicator of this disease. Axial T2-weighted MRI is used to measure the arrangement of basal ganglia. It is particularly worth noting that MRI morphometry (Oba *et al* 2005) allowed a series of studies that led to the creation of the Quattrone index (Quattrone *et al* 2008). To allow this index to work properly, it is very important to correctly detect the mid-sagittal slice in MRI. This slice, usually seen as an indicator of variation, allows one to observe the main internal anatomical structures in the MSP (Ruppert *et al* 2011), taking advantage of the mirror image symmetry of the human brain. Determination of the exact location of this plane is required for many applications; however, there is no universal agreement about the identification of the MSP, as the dividing plane between the brain hemispheres often does not correspond to the plane of symmetry of the head.

Changes in the neurotransmitter systems and signal transduction mechanism are very frequent in patients with AD, altering the cholinergic signaling system and the production of the neurotransmitter acetylcholine (Crews and Masliah 2010). Moreover, we can observe other cerebral alterations both macroscopic (a decrease in the weight and volume of the brain, due to cortical atrophy and ventricular dilatation) and microscopic (neuronal loss, glial and astroglial reaction, microvessel alteration). As a consequence of these brain modifications it becomes impossible for the neurons to transmit nerve impulses; these neurons then die and progressive atrophy of the brain as a whole ensues (Crews and Masliah 2010).

MS is a complex neurodegenerative disease characterized by inflammation which results in multifocal demyelinating lesions and degeneration, with diffuse axonal loss leading to brain atrophy in the central nervous system (Lombardo *et al* 2017). Given the cyclical relapsing/remitting behavior of MS, MRI is fundamental in the diagnosis and monitoring of treatment. Traditional quantitative parameters include whole brain and white and gray matter volumes, as well as the brain lesions load, with the use of sequences and complex post-processing techniques, which are usually time-consuming procedures if they are not automated by particular segmentation algorithms (Bilotta *et al* 2010, Cerasa *et al* 2012).

The improvement of MRI techniques is also useful for the novel field of network physiology, in particular to reach its main goal of building the first complex atlas of dynamic interactions between different brain locations and organ systems (Bartsch *et al* 2015). The human organism comprises a complex and integrated network of different organ systems, each with its own regulatory dynamic mechanism and dynamic interactions between each system that define different physiological states (Ivanov and Bartsch 2014). Changes in these networks of interactions indicate not only changes between different physiological states but also the transition between a physiological situation and a pathological one. Since the different organ systems are closely connected, a failure in one organ can lead to total failure of the organism; therefore mapping and studying changes in the network of interactions could aid early diagnosis neurodegenerative diseases such as PD and MS that involve other organ systems. Further steps in brain imaging could help to reconstruct anatomical brain connectivity, providing a powerful tool for diagnosis of neurodegenerative disorders as well as for extracting information about the functional network connectivity of the brain.

The problem with identifying the MSP (and in general with morphometric measurements of the brain) is that measurements are made in an environment with variability characteristics that are relevant. Consider, for example, the difference in the resolution of brain scans: this depends on the type of brain scan employed, the time taken for shooting, the variability of the morphology of individual patients and the multiplicity of motion artifacts due to technical problems or casual movements of the skull during recording. Moreover, the method most often used to analyze these changes in measurements of volumes, areas and distances is to return the set of images to the standard model in order to segment the new dataset. But very often this approach is not useful because it reveals that interpolation techniques modify original data and alter brain images, making subsequent measurements unsuitable for the correct identification of the proper disease markers. This happens when, for instance, an entire set of MRI scans is tilted in order to make the scan plane parallel to the sagittal plane. In this case, rigid rotation is the first step in the pre-processing of the MRI scans. Then, an interpolation is required in order to represent the new MRI set as an imaginary cube, the three-dimensional (3D) reconstructed image of whole brain, and the same is done for each voxel. In this last operation some information is lost, for example information about the ratio between the different dimensions of some brain areas—the distances and volumes change, especially in the mid-brain, the area of interest in the diagnostic process for the above-mentioned diseases.

To meet these needs, and to support medical diagnosis, we have implemented a computational method for the automatic identification of the MSP from the raw data. Developed in Matlab, it uses a classic k -means based algorithm to localize the slice of the DICOM file containing the MSP among all the structural MRI (sMRI) brain scans. To validate the method, we compared its performance with manual measurements conducted by four expert raters who carefully analyzed the MRI brain scans of the healthy, PSP, PD, AD and MS subjects. In order to carefully compare the manual segmentation results with the results of our proposed algorithm we introduced a definition of ‘gold standard’ for MSP location, by imposing majority rules and developing a ‘cognitive’ decisional algorithm for the human raters’ measurements. Furthermore, we trained a machine learning system with the results of our proposed algorithm and the ‘gold standard’. The ultimate purpose of this study is to provide computational tools that can be used to develop fully automated systems with the capability to recognize patterns relevant to medical diagnosis and clinical investigation.

The paper is organized as follows. In section 2 the dataset used and the methods are outlined. The results follow in section 3, with the main conclusions and further developments in section 4.

2. Materials and methods

2.1. Datasets

The datasets comprised a total of 109 MRI scans as described in the following paragraphs.

Brain scans of 37 individuals comprising 14 healthy control subjects (mean age 52 years; six women, eight men), 13 PD subjects (mean age 69 years; four women, nine men) and 10 PSP subjects (mean age 71 years; three women, seven men) provided by the CNR Catanzaro (CZ, Italy). The scans were acquired by a 3.0 T magnetic resonance (MR) scanner (GE Medical Systems Discovery MR750) using a 3D T1-weighted sequence [sagittal acquisition plane, inversion time 650 ms, repetition time (TR) = 9.15 ms, echo time (TE) = 3.67 ms, slice thickness 1.0 mm, resolution 256×256 pixels, voxel size $1.0 \text{ mm} \times 1.0 \text{ mm} \times 0.5 \text{ mm}$]. This set of data was used in a previous paper by some of the authors (Nigro *et al* 2014).

To test if our methods are independent with respect to the specific MRI scanner used and to PD and its variants, data from 15 MS subjects (mean age 45 years; 12 women, three men) were used. Data related to MS were collected at the neurodiagnostic unit of Cetraro Hospital (CS), in compliance with the Privacy Act and current legislation (Declaration of Helsinki), which provided MRI files. Brain scans were acquired using a 1.5 T MR scanner (Philips Achieva Rev R5 v3-rev.00) with slice thickness 1.0 mm, resolution 336×336 pixels, voxel size $0.762 \text{ mm} \times 0.762 \text{ mm} \times 1.0 \text{ mm}$, TR = 7.0286 ms and TE = 3.178 ms. Subjects’ data were treated according to the current laws on privacy. The ethics committee of the Cetraro Hospital approved the research.

Furthermore, MRI scans for 57 AD subjects (mean age 75 years, 29 women, 28 men) were obtained from the Alzheimer’s Disease Neuroimaging Initiative (ADNI) database (<http://adni.loni.usc.edu/>). The ADNI was launched in 2003 as a public–private partnership, led by principal investigator Michael W Weiner, MD. The primary goal of ADNI has been to test whether serial MRI, positron emission tomography (PET), other biological markers and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early AD (for up-to-date information see <http://adni.loni.usc.edu/>). All these brain MRIs were acquired using a 3.0 T MR scanner (Siemens) with inversion time 900 ms, TE = 2.98 ms, TR = 2300.0 ms, resolution 240×256 pixels, slice thickness 1.0 mm and voxel size $1.0 \text{ mm} \times 1.0 \text{ mm} \times 1.0 \text{ mm}$).

In summary: for each group of subjects there are different technical specifications related to the type of brain scan performed and the physical characteristics of the system used. To test our methods, we used 3D T1-weighted sequences (sagittal acquisition plane), restricting our interest to a range of 100/101 central slices, depending on whether the total number of slices was even or odd, respectively, in order to always choose a central interval and not computationally overburden the software. Technical features of the datasets are summarized in table 1. All data scans were anonymized to comply with the current ethical requirements.

2.2. k -means

The k -means algorithm is an exclusive or partitioning-type algorithm. Given a set of n objects D and a number of clusters k , it organizes objects into separate partitions k ($k \leq n$), where each one represents a cluster (MacQueen 1965). Clusters are used in order to optimize a grouping criterion, generally a function based on the distance between the objects. In this case, the similarity measure is based on the average value of the objects in a cluster, which can be seen as the centroid or center of gravity of the cluster.

Given a set of n elements $S = \{x_{i,i=1,\dots,n}\}$ defined in a space where it is possible to state a metric d , and the number of clusters k in which to partition the set, k elements $c_j \in S, j \in \{1, \dots, k\}$ are randomly chosen among all the elements of S . Each c_j will be at first the centroid of the corresponding cluster C_j , with $j \in \{1, \dots, k\}$. Then, another element $x_i \in S$ is randomly chosen. x_i will be associated with the cluster C_{j_0} whose centroid c_{j_0} is the closest of all the centroids c_j , in accordance with the metric d :

Table 1. Technical features of the datasets used in this work.

Dataset	Subsets	Scanner machine	Slice resolution (pixels)	Voxel size (mm)	Slice thickness (mm)	Total number of slices	Range considered (slice number)
CNR Catanzaro	Healthy	GE Medical Systems Discovery MR750 3.0 T	256 × 256	1.0 × 1.0 × 0.5	1.0	367/368	133–233/134–233
	PD		256 × 256	1.0 × 1.0 × 0.5	1.0	367/368	133–233/134–233
	PSP		256 × 256	1.0 × 1.0 × 0.5	1.0	367/368	133–233/134–233
Cetraro Hospital	MS	Philips Achieva Rev R5 v3-rev.00 1.5 T	336 × 336	0.762 × 0.762 × 1.0	1.0	210	55–154
ADNI	AD	Siemens 3.0 T	240 × 256	1.0 × 1.0 × 1.0	1.0	176/208	38–137/54–153

$$\underline{x}_i \in C_{j_0} \quad \text{so that} \quad d(\underline{x}_i, c_{j_0}) = \min_{1 \leq j \leq k} d(\underline{x}_i, c_j)$$

Cluster C_{j_0} will have a new centroid, calculated considering both c_{j_0} and \underline{x}_i . This is repeated by identifying the cluster to which another random element $\underline{x}_i \in S$ belongs. The process is iteratively repeated until the whole set S has been partitioned as follows:

$$\forall i \in \{1, \dots, n\} \quad \exists j \in \{1, \dots, k\} \quad \text{such that} \quad \underline{x}_i \in C_j.$$

For the aim of this work, we considered the T1-weighted sequences of sMRI brain scans of the experimental subjects acquired along the sagittal plane, and using a k -means clustering based algorithm we segmented each 2D image, corresponding to a slice, into four clusters (figure 1).

2.3. Image pixel intensity (IPI)

In order to reduce the computational cost, and since we are trying to locate the mid-sagittal reference slice, we can just consider the central 100/101 slices. These will be placed in an INPUT folder and we repeat the procedure for each subject. The goal is to identify the slice in which the difference in term of gray-scale pixel intensity between the different brain tissues is more marked. For this reason we called our algorithm the image pixel intensity (IPI) algorithm. In our proposed algorithm we used two main Matlab scripts: *k_mean.m* and *peaks.m*.

1. *k_mean.m*

We can divide this script into three sub-parts:

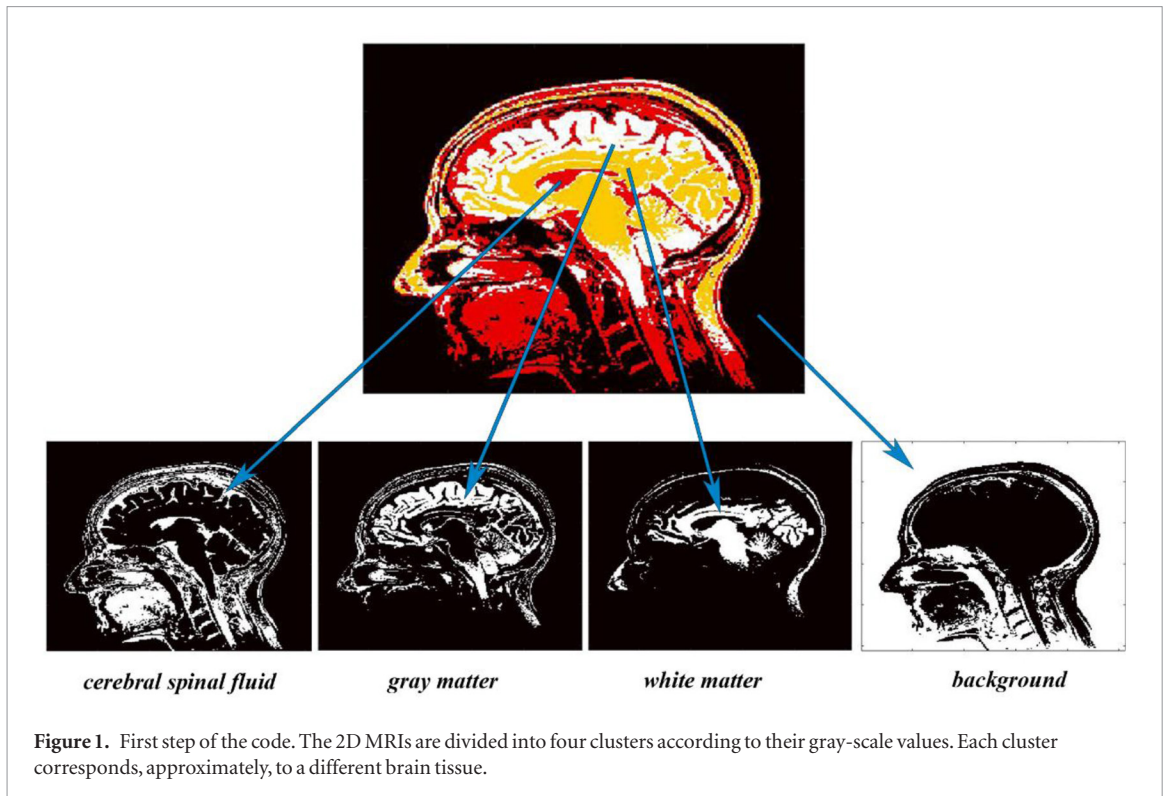
a. Iteration of the k -means method to all DICOM files in the INPUT folder of each subject. In this case:

- the set S is a 2D image (the slice on sagittal plane);
- the elements x_i are the points in the image corresponding to the pixels;
- the metric d is defined by

$$d(x_i, x_j) = |v(x_i) - v(x_j)|.$$

- $k = 4$

We chose a cluster number equal to four because experimentally we observed that with this choice we can obtain the best slice partition. Indeed, by partitioning the image into four clusters it is possible to distinguish quite well the different areas of the mid-brain, the region of greatest interest in defining the MSP (figure 1). Moreover, with this choice we can distinguish the different brain tissues from each other. Next, the clusters are sorted in ascending order depending on the number of pixels they contain. Therefore, cluster 1 will contain the points corresponding to pixels representing the cerebrospinal fluid, cluster 2 will contain the points corresponding to pixels representing gray matter, cluster 3 will contain the points corresponding to pixels representing white matter, and finally cluster 4 will contain the points corresponding to pixels representing the background. It is possible to determine this order because at the variation of the slice, within the central slices, the ratio between the numbers of pixels of the different brain tissues is always the same. The process is repeated iteratively for all files (sMRI central slices) in the INPUT folder for each subject.



b. Graph creation.

This part of the script concerns the creation of the graphs for each subject. Each graph represents the number of pixels contained in each cluster by varying the slice index, always considering only the 100/101 central slices. Note the clusters for each slice are sorted in increasing order. This step is crucial since the choice of the number associated with each cluster in the k -means iteration occurs randomly as the initial centroids in the first part are randomly chosen among all points in the image. Therefore, once number 1 can be associated with the cluster containing the points representing the cerebrospinal fluid, another time, when applying the same method to another slice, the number 1 can be associated with the cluster corresponding to the white matter and so on for the other brain tissues. Sorting clusters every time in ascending order can ensure two-way correspondence between the cluster identification number and the cerebral tissue (figure 2(a)).

- c. Finally, the script creates a graph representing the differences in the number of pixels in different tissues, by varying the slice index (figure 2(b)). Note that the difference in the number of pixels between region 4 and the others is not considered because cluster 4 contains pixels corresponding to the background, and for the central slices there are no substantial differences between one slice and another in the number of pixels representing the background.

2. *peaks.m*

This script allows us to automatically identify the peaks of each ‘difference curve’ for each subject, whether they are absolute or relative maximum or minimum, provided that the jump between the function valuated at the critical point and the average of the values that the function assumes elsewhere is relevant. In particular, suppose we analyze a ‘difference curve’ that we call v . Now we can consider two cases:

- a. The peak of the curve is situated within the 40 central slices. In this case only the 40 central slices are considered to be established if the curve has a maximum or a minimum in the formulae
- if $|\min_{30 \leq i \leq 70} v(i) - \frac{v(30)+v(70)}{2}| > |\max_{30 \leq i \leq 70} v(i) - \frac{v(30)+v(70)}{2}|$, then v has an absolute minimum in j such that $v(j) = \min_{30 \leq i \leq 70} v(i)$;
 - if $|\min_{30 \leq i \leq 70} v(i) - \frac{v(30)+v(70)}{2}| < |\max_{30 \leq i \leq 70} v(i) - \frac{v(30)+v(70)}{2}|$, then v has an absolute maximum in j such that $v(j) = \max_{30 \leq i \leq 70} v(i)$.

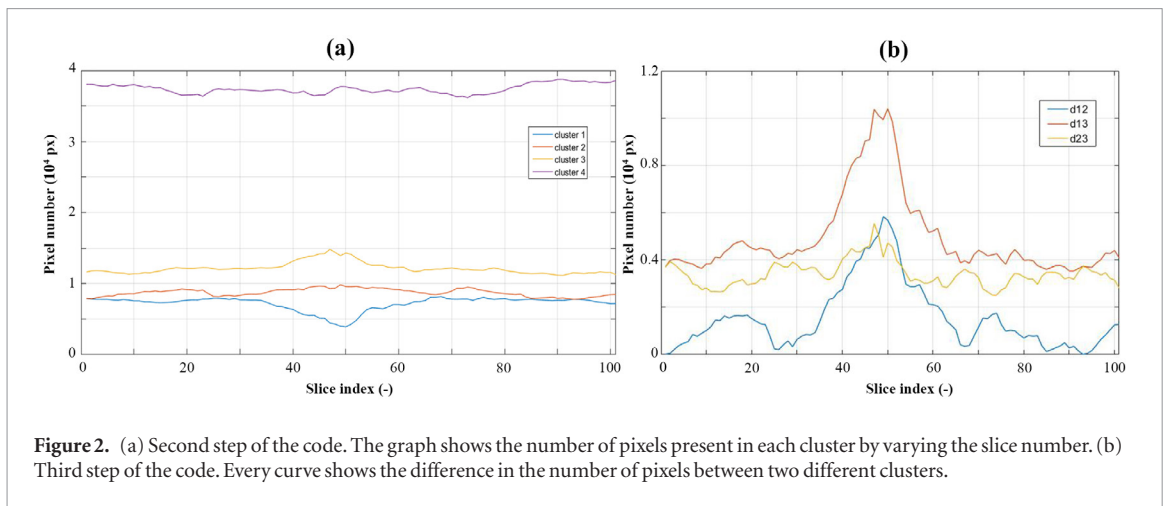


Figure 2. (a) Second step of the code. The graph shows the number of pixels present in each cluster by varying the slice number. (b) Third step of the code. Every curve shows the difference in the number of pixels between two different clusters.

Note that we do not consider the average of the values that ν assumes elsewhere to reduce the computational cost and also because we notice that there are no substantial changes between $\nu(30)$ and $\nu(i)$, $i \in \{1, \dots, 30\}$, or between $\nu(70)$ and $\nu(i)$, $i \in \{71, \dots, 100\}$.

- b. The peak of the curve is situated outside the 40 central slices or too close to the edges of the central range (in particular if the peak is located in $[m, m + 3]$ or in $[M - 3, M]$ where m and M are, respectively, the minimum and the maximum extremes of the 40-slice central range) to surely establish if that point is a maximum or minimum for the curve (this point could be part of the ascending or descending section of the curve before reaching the maximum or minimum outside the central range). In this case we consider the whole interval of the central 100/101 slices. The minimum or maximum of the curve is chosen using the same procedure explained in 2a.

We indicate the ‘difference curve’ between clusters i and j as d_{ij} , $i, j \in \{1, 2, 3\}$. Finally, we compute the arithmetic average between the indices of the slice corresponding to the peak of each curve. The output of the previous script is a vector $p = (p_1, p_2, p_3)$, where $p_{h, h \in \{1, 2, 3\}}$ is the index of the slice corresponding to the maximum or minimum of the ‘difference curve’ d_{12} , d_{13} , d_{23} , respectively. We calculated the average $m = \frac{1}{3}(p_1 + p_2 + p_3)$ and the value is approximated to the nearest integer. The same procedure is repeated for all subjects.

2.4. Implementation of the IPI algorithm

A block diagram of our proposed IPI algorithm is shown in figure 3.

Regardless of the resolution of the brain scans, the central 100/101 slices for each subject are provided as input to the developed system. The first script *k_mean.m* works on each slice improving the *k*-means method, shown from the second cycle. The first cycle ends when all the slices have been analyzed. Then the script creates a graph representing the number of pixels in the different clusters by varying the slice index and computes the difference between the number of pixels in the clusters. Finally, the script *peaks.m* gives the critical points of ‘difference curves’, thus identifying the index of the mid-sagittal reference slice.

2.5. Inter-rater reliability and gold standard definition

To obtain a gold standard with which to compare the performance of the developed IPI algorithm, we considered an independent evaluation of the sMRI images by human experts. We invited the opinions of four human experts (raters of sMRI images) who manually segmented the MSP for each subject in the considered sample.

In order to arrive at a perfect agreement with the mathematical algorithms in delineating the mid-sagittal slice that could be used as a standard to compare the performance of the algorithm, we used a statistical–mathematical model that allows us to outline an evaluation of the performance of the individual rater and an analysis of the characteristics of each item studied (Lord and Novick 1968). There are two ways of applying this method: *dichotomous* and *polychromous* ratings. For dichotomous rating, values *correct/incorrect* is assigned to each rater’s response to obtain a proportional rating. The results gave us the percentage of agreement among raters. This means, for instance, that if raters agree in 61% of the 109 cases considered they do not agree in the remaining 39% of cases. Some limitations of this method are related to the fact that this measure does not discriminate exactly between agreement on positive and negative ratings and, having such a low percentage of success, it could not be considered an optimal gold standard by which test the performance of our automatic algorithm.

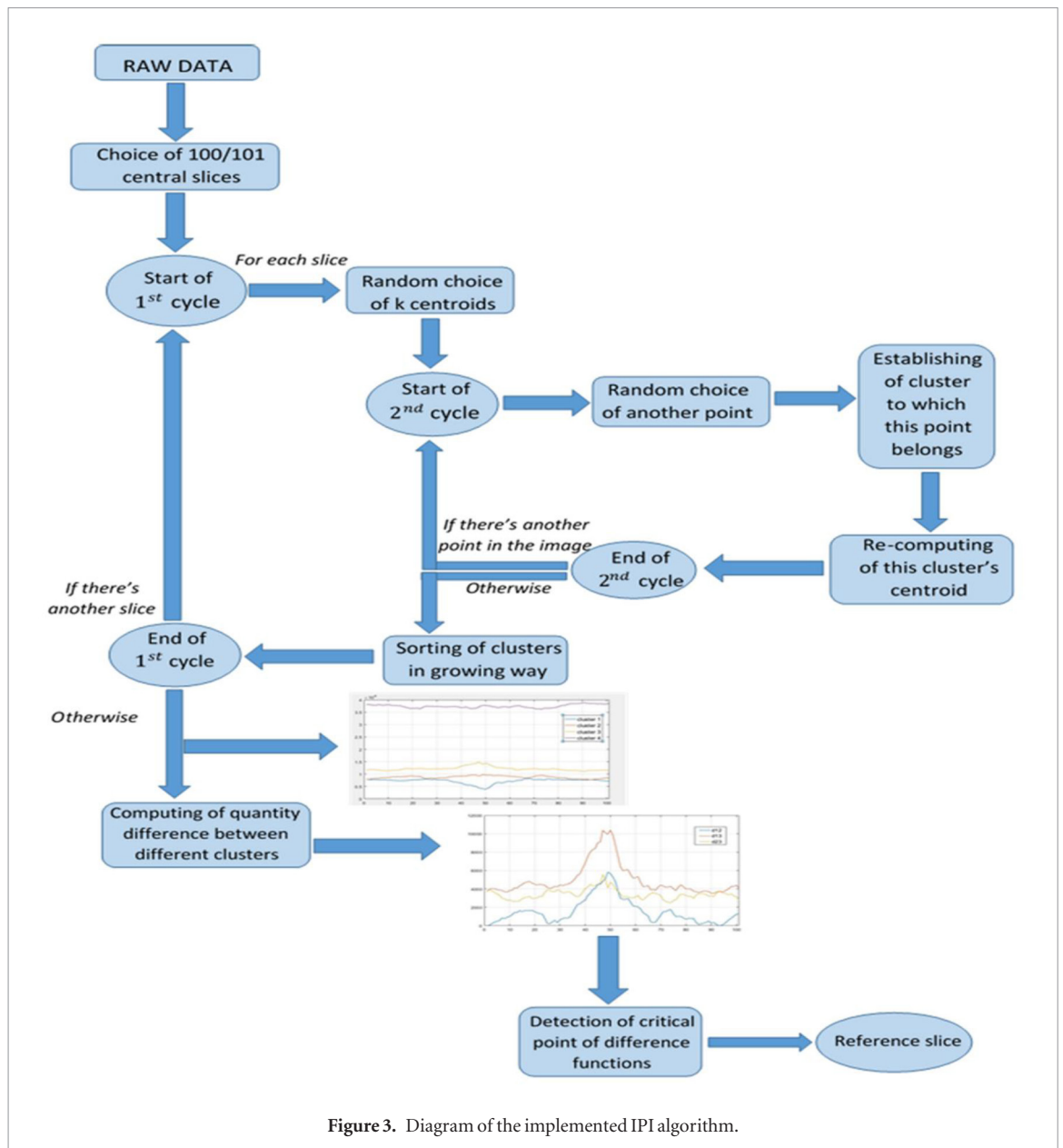


Figure 3. Diagram of the implemented IPI algorithm.

The central problem the raters found in the measurement of the MSP was that there is no single unique slice that identifies the subtle morphological differences of the midbrain pattern in the mid-sagittal, but rather a dynamical interval with the rise and fall of the correct mid-sagittal configuration. Therefore, the choice could be from among those slices belonging to previously defined sets of slices to which we assigned the values for the rating categories or levels. To satisfying this requirement we implemented a polychromous rating, giving scores of 2, 3 and 4, for differences among raters of two, three and four slices. This method, while showing improved percentage agreement, also exhibited many downsides: we can have a high enough percentage agreement only when the raters identify exactly the same slice, and the measurements of the expert raters always differed by at least one to two slices. Therefore, we decided to discard the polychromous approach.

In order to find agreement between raters' measurements we needed to implement a procedure that could not take in account small variations (one to two slices); thus we considered the arithmetic average between the slices identified by raters as mid-sagittal. Nevertheless, this approach is not much more reliable. For instance, if we have three raters who have indicated the same value x and only one who has indicated the value y , it seems reasonable that the correct value is x and not the average of $\{x, x, x, y\}$.

Therefore, we incorporated a decision-making process with majority rule and a 'cognitive' decisional algorithm that we developed to support the choice. The following procedure has been applied to establish the agreement between raters on the identification of the MSP for each subject:

- In the case of majority of agreement among the raters upon a slice as the mid-sagittal reference, that slice is chosen as the gold standard.

- Otherwise, when there is a split decision (two of the raters versus the other two raters, or all four raters have different opinions) a random way of choosing the mid-sagittal reference slice from between the two or four different values proposed by the raters is employed.

We note that in the case where there is no majority in agreement, taking the average between the measurements instead of a random choice between the values provided by the raters may not offer the best decision, since this could happen in the case where one of the raters is not sufficiently precise (due to tiredness or other reasons) for the particular measure and the average would take in account this rater's evaluation but it would be different from the measurements of all the other raters, not picking any of them. To avoid this problem we decided to randomly choose the index of the reference slice for the MSP from the different raters' evaluations.

Finally, the defined gold standard was compared with the results obtained from the IPI algorithm for each subject. In particular, the slice corresponding to the peak for each 'difference curve' was taken into account as well as the arithmetic average of the index of the slices corresponding to the peak of all the three 'difference curves', in order to study the reliability of each curve and the average to the gold standard.

2.6. Machine learning approach

The last step of our work involves the development of a machine learning model to further test the reliability of the outcome of the IPI algorithm to automatically locate the MSP. Our aim is to apply an unsupervised machine learning technique in order to verify if the results of the proposed algorithm are useful as input to train a machine to automatically detect the MSP reference slice, given the gold standard as output.

We chose to apply an ensemble learning method for the regression task, namely the random forest (RF) model (Breiman 2001) with the following parameters:

- number of trees = 10;
- tree depth = 5.

We chose a number of trees equal to 10 since, following the literature (Liaw and Wiener 2002, Oshiro *et al* 2012) and experimentally, this choice offered a good balance between computational cost and high accuracy for the model to predict the expected value for the MSP. Moreover, for simplicity and prevention of over-fitting, we chose a tree depth equal to 5 (Criminisi *et al* 2010, Bozkir and Sezer 2011).

We employed 10-fold cross-validation on the entire set of 109 subjects. In particular, in each iteration a subset of 98/99 random subjects was chosen to be the training set, and the rest (10/11 subjects) were reserved for the test set. Moreover, during each iteration the training set is randomly divided in five subsets of 19/20 subjects. Each sub-subset is used as a training set, then we double the size of the training set, using the previous sub-subset together with another sub-subset, and keep increasing the size of the training set, adding a sub-subset every time until the whole subset of 98/99 subjects has been used as training set. Therefore, we trained the machine with 20-, 40-, 60-, 80- and 98/99-subject sets in sequence in order to check whether or not the IPI results constituted a good set of objects for training the machine (by analogy with human learning, the greater the amount of information used to help someone learn the more he or she should be able to recognize objects).

3. Results

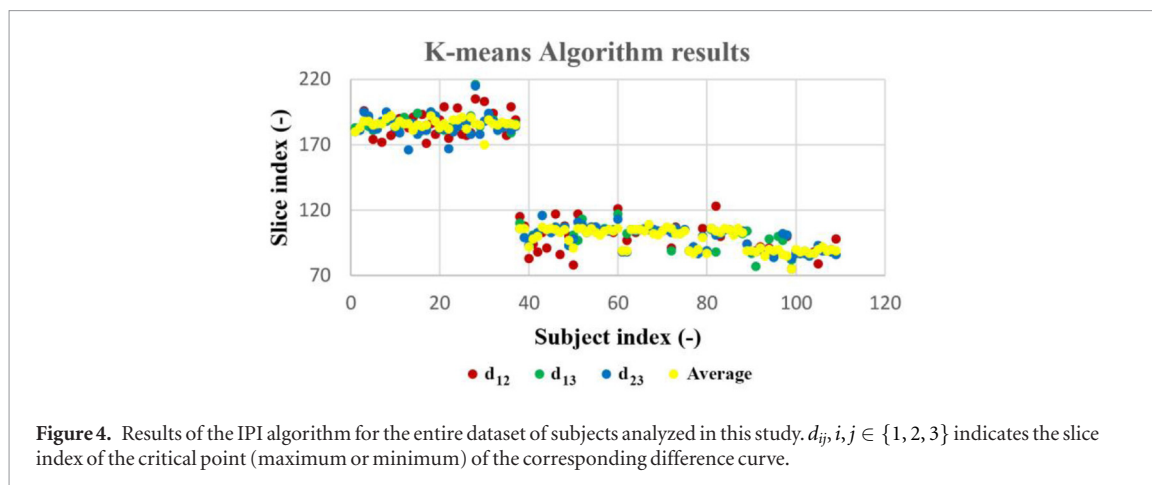
3.1. The performance of the IPI algorithm

Results obtained by the automatic detection of the MSP as explained in the previous section are displayed in figure 4.

The first 37 subjects (CNR Catanzaro dataset in table 1) have a MRI set of 367/368 slices in total. Therefore the IPI algorithm returns a proposed value for the MSP reference slice of 186.0 ± 8.8 , 187.1 ± 6.3 and 184.8 ± 8.3 , respectively, for curves d_{12} , d_{13} and d_{23} , taking into account the arithmetic average and the standard deviation across all 37 subjects. For the other 72 subjects (Cetraro Hospital and ADNI datasets in table 1), who have a MRI set of 176/210 slices, the IPI algorithm returns a value for the MSP reference slice of 98.1 ± 10.3 , 98.5 ± 8.8 and 98.6 ± 8.4 , respectively, for the curves d_{12} , d_{13} and d_{23} , taking into account the arithmetic average and the standard deviation across all 72 subjects.

3.2. Inter-rater agreement and emergence of a gold standard

To verify the outcome of our proposed IPI algorithm we used an independent evaluation of the sMRI images by human experts. In order to arrive at a perfect agreement between raters and from their collective decision to obtain a 'gold standard' with which to compare the performance of our algorithm we employed the 'cognitive' decisional algorithm explained in section 2.5.



Results for the manual segmentation of the mid-sagittal performed by expert raters in the different data subsets are reported in figure 5. The agreement between raters has a standard deviation (SD) of 1.8555 for healthy subject, 1.2780 for PD subjects, 1.9195 for PSP subjects, 2.3208 for MS subjects and 1.4003 for AD subjects.

The raters' opinions differ between themselves in a relevant way, if we consider a dichotomous approach whereby they can reach the identification target or not. To estimate the mutual agreement between raters in the manual identification of the MSP, we calculated the arithmetic average between the evaluations of the raters and obtained the distribution of the performance of each rater with respect to the average of all opinions (figure 6). We realized that each rater has a different performance which can diverge very much regarding the frequency of correct evaluation.

3.3. Comparison 'gold standard'—the IPI algorithm

For definition of the gold standard we incorporated a decision-making process with the majority rule and a 'cognitive' decisional algorithm, described in section 2.5. Then, we used the 'gold standard' thus defined to compare the performance of the IPI algorithm.

Figure 7 shows the results for slice distributions identified as the peak of the three curves compared with the 'gold standard', shown in the histograms (figure 7(a)), the distribution of the results grouped according to the absolute error frequency (figure 7(b)), and the corresponding distribution in quartiles of the algorithm results (figure 7(c)). From the tables in figure 7(b) we see that the first curve nicks the target slice 17 times while the other two algorithms hit the objective slice 26 and 25 times, respectively. The means of absolute errors are $e_{12} = 5.00917$, $e_{13} = 3.33028$, $e_{23} = 3.08257$ respectively for d_{12} , d_{13} , d_{23} . It is evident that the best prediction curve turns out to be the d_{23} , whose error average is 3 slices. The standard deviations of the three prediction curve are $SD_{12} = 6.06523$, $SD_{13} = 4.70826$, $SD_{23} = 4.49718$ respectively. We note that the third algorithm has the least dispersion, as is also clear from figure 7(c). However, it is interesting to observe these averages in relation to the mean absolute errors of the raters and their standard deviations, which are respectively equal to $e_1 = 1.77982$, $e_2 = 1.5412$, $e_3 = 0.834862$, $e_4 = 1$ and $SD_1 = 4.04882$, $SD_2 = 3.18956$, $SD_3 = 2.41502$, $SD_4 = 2.46281$. Indeed, we can consider the results from the three curves as the opinion of three different people that can provide results more or less close to the gold standard. From this it emerges that the best algorithm has a 1.5 slice error compared with the worst rater, whereas the standard deviation of the best algorithm is close enough to the standard deviation of that obtained by the worst rater.

Rather than searching for an arithmetic average of the absolute error, it is more logical to compute a weighted average of the absolute error, taking into account the different number of slices for each subject. We noticed that data from CNR Catanzaro are brain MRI scans composed of 367/368 slices in total and the space between slices is 0.5 mm, while the data from Cetraro Hospital and ADNI are brain MRI scans composed of about 200 slices in total and the space between slices is 1.0 mm (table 1). Therefore, a certain number of slices from the first set of subjects cover a physical space smaller than the space covered by the same number of slices in the other two datasets. In the computation of the error from the comparison between the proposed IPI algorithm results and the 'gold standard', averaged across all subjects, we wanted to assign different weights to the data from different laboratories, according to the different sizes of the physical space covered by a certain number of slices. To each subject s_i we assigned the weight $w_i = 1/N_i$ where N_i is the total number of slices for subject s_i . In this way the errors corresponding to subjects in the first dataset have a smaller weight than the errors for the subjects in the other two datasets. Then we computed the weighted mean absolute error as

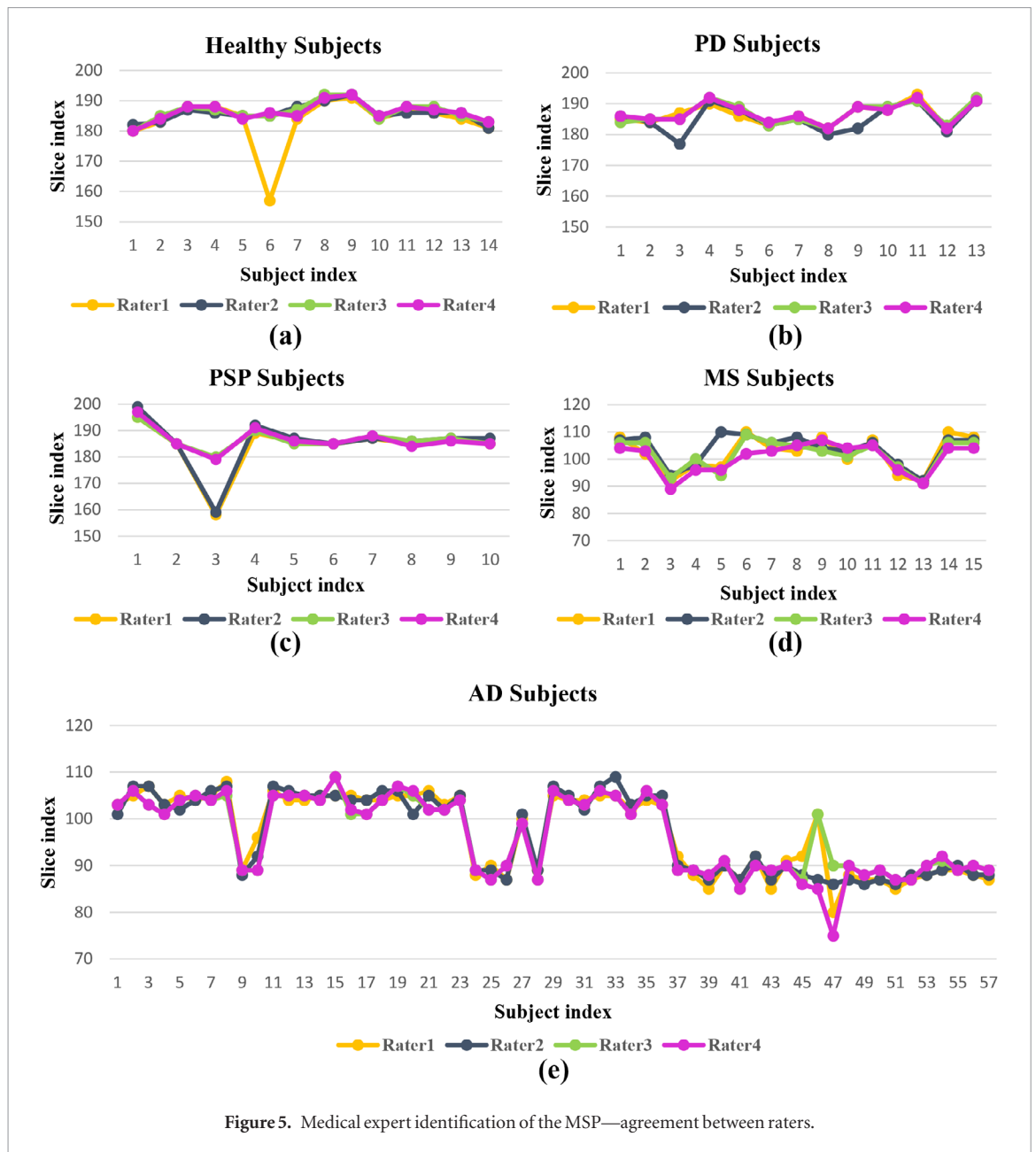


Figure 5. Medical expert identification of the MSP—agreement between raters.

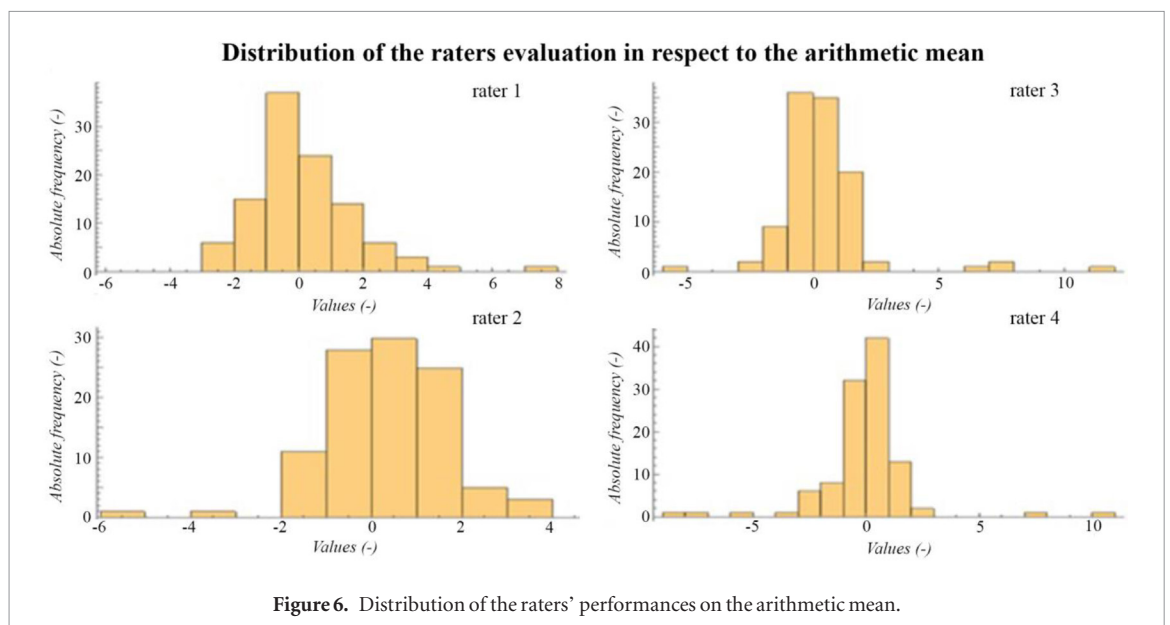


Figure 6. Distribution of the raters' performances on the arithmetic mean.

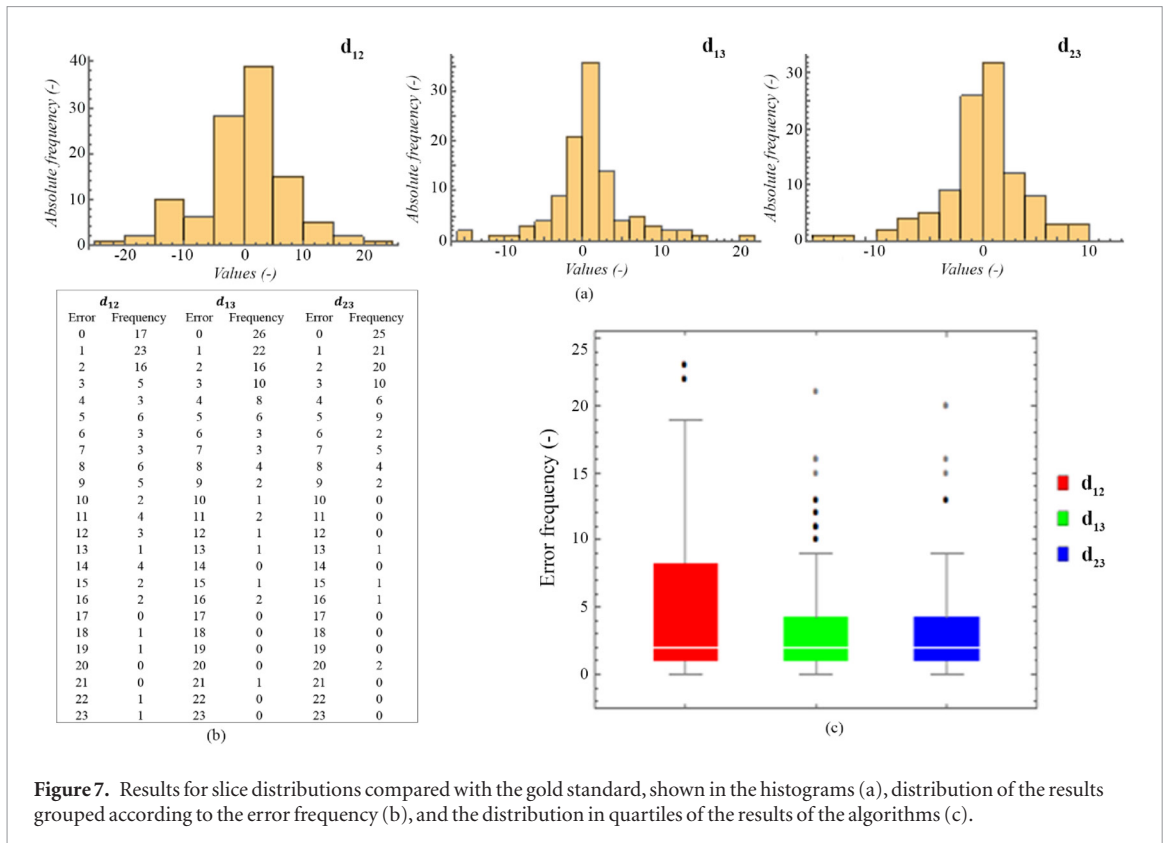


Figure 7. Results for slice distributions compared with the gold standard, shown in the histograms (a), distribution of the results grouped according to the error frequency (b), and the distribution in quartiles of the results of the algorithms (c).

$$\bar{e}_w = \sum_{i=1}^{109} \frac{e_i}{N_i} \cdot \frac{1}{\sum_{i=1}^{109} 1/N_i}.$$

Using the same weights, we computed the weighted standard deviation

$$SD_w = \sqrt{\sum_{i=1}^{109} \frac{(e_i - \bar{e}_w)^2}{N_i} \cdot \frac{109}{108 \sum_{i=1}^{109} 1/N_i}}.$$

We obtained the weighted mean absolute errors $\bar{e}_{W_{12}} = 4.44785$, $\bar{e}_{W_{13}} = 3.18746$, $\bar{e}_{W_{23}} = 2.55357$ and weighted standard deviations $SD_{W_{12}} = 5.52971$, $SD_{W_{13}} = 4.50097$, $SD_{W_{23}} = 3.84922$.

The weighted mean relative errors are $\bar{e}_{W_{r_{12}}} = 0.03765$, $\bar{e}_{W_{r_{13}}} = 0.02874$, $\bar{e}_{W_{r_{23}}} = 0.02039$ respectively. Thus the weighted mean percentage error is about 2.04% for the best algorithm while it is 3.77% for the worst algorithm.

Since we want to localize the reference slice where the differences between brain tissues are more marked and there is no reason to assign a larger weight to one difference curve than another, we computed an arithmetic average of results extracted from the three ‘difference curves’. The weighted mean relative error between the arithmetic average of the different ‘difference curves’ and the gold standard averaged across the subjects is $\bar{e}_{W_{r_{ave}}} = 0.01842$, $\bar{e}_{W_{p_{ave}}} = 1.84\%$, obtaining a better result even than the best algorithm d_{23} . The weighted standard deviation is $SD_{W_{ave}} = 3.67198$, lower than $SD_{W_{23}} = 3.84922$. A comparison with the other distributions (considering the absolute errors on the slices) is shown in figure 8.

All these results show that the developed system is already comparable to the performance of the raters.

For the machine learning approach we applied the RF model, employing a 10-fold cross-validation process on the entire set of 109 subjects. In each iteration a sub training set of 98/99 random subjects is partitioned into five subsets of 19/20 subjects. For each of these we trained the machine, increasing the size of the training set (20, 40, 60, 80, 98/99 subjects), and tested the machine on the test set (10/11 subjects) corresponding to the iteration of the 10-fold cross-validation process. Therefore, in total we repeated the training process 250 times: five times increasing the size of the training set starting from each sub-subset of 19/20 subjects (five sub-subset in total for each iteration in the 10-fold cross-validation process). After the training and testing process we evaluated the performance of the model, computing the mean absolute error between the actual output of the machine and the expected value for the output (‘gold standard’) averaged over all the subjects in the test set. Figure 9 shows the evolution of the accuracy of the RF model, increasing the size of the training set. Starting from the smallest to the

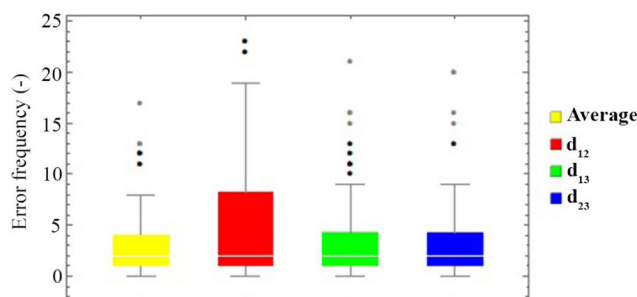


Figure 8. Comparison of the distributions of the three algorithms and their average, considering the absolute error on the slices.

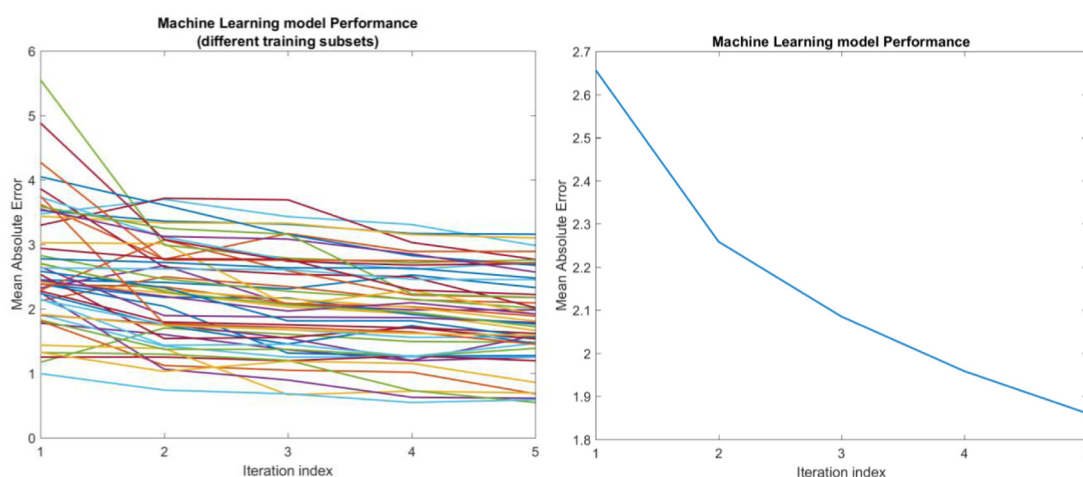


Figure 9. Evolution of the performance of the RF model increasing the size of the training set. On the x -axis the index of the iteration (in sequence 20, 40, 60, 80, 98/99 subjects), on the y -axis the value of the mean absolute error computed between the actual output of the machine and the expected output ('gold standard') averaged over all the subjects in the training set. Left panel: The performance of different iterations, considering different sub-subsets of 19/20 subjects as the starting point, is shown. Right panel: The curve shows the mean absolute error computed for each size of the training set, averaged over all iterations.

biggest training set the accuracy of the model increases by 30%, arriving at a mean absolute error, averaged over all the iterations $\bar{e} = 1.8609$ for the training set of 98/99 subjects.

We demonstrated that the output of our proposed IPI algorithm is good enough to train a machine in order to automatically localize the MSP reference slice, since by increasing the size of the training set the machine actually learns better and improves its performance.

4. Conclusions

The obtained results show how the performance the system has been improved, increasing its accuracy and making the extreme variability of the identification task more flexible. The human brain is highly variable. Although MRI systems are currently the most powerful machines for detecting this variability, they also have many drawbacks in the visual rendering of data. So the problem we face is highly sensitive to the initial data. Consequently, each subject in the sample has been carefully analyzed, adopting the technique of polychromatic ratings, which by enlarging the intervals, better specified the accuracy of the developed tool. The IPI algorithm could be used to automatically segment the brain sMRI images and localize the MSP. Moreover, the machine learning system allows forecasting of the MSP from a MRI file, in an automatic way, without passing through the repetition of the procedure that we have described in this work. In fact, by means of the training set of data in this article, used as a computational benchmark, we can forecast any set of data, independently of the MRI system and neurodegenerative disease. Continuation along this path can provide excellent results, optimizing the system to make it as reliable as a human expert. Indeed, the next step in this framework is to collect a larger dataset and test the machine learning system on it. This method could also find application in the identification of different brain locations, key points in the understanding of brain network interactions and connections with other organ systems. Indeed, detecting particular brain areas responsible for the strongest connection with a particular organ

system during a particular physiological state could help derive a pathological picture of the whole organism from a physiological one (Bashan *et al* 2012, Ivanov *et al* 2016).

Acknowledgments

For the 57 AD subjects, data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant No. U01 AG024904) and DOD ADNI (Department of Defense Grant No. W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association, Alzheimer's Drug Discovery Foundation, Araclon Biotech, BioClinica, Inc., Biogen, Bristol-Myers Squibb Company, CereSpir, Inc., Cogstate, Eisai Inc., Elan Pharmaceuticals, Inc., Eli Lilly and Company, EuroImmun, F Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc., Fujirebio, GE Healthcare, IXICO Ltd, Janssen Alzheimer Immunotherapy Research & Development, LLC, Johnson & Johnson Pharmaceutical Research & Development LLC, Lumosity, Lundbeck, Merck & Co., Inc., Meso Scale Diagnostics, LLC, NeuroRx Research, Neurotrack Technologies, Novartis Pharmaceuticals Corporation, Pfizer Inc., Piramal Imaging, Servier, Takeda Pharmaceutical Company and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (<https://fnih.org/>). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

This research did not receive any specific grant from funding agencies in the public, commercial or not-for-profit sectors.

Declarations of interest

None.

ORCID iDs

Rossella Rizzo  <https://orcid.org/0000-0003-4486-4794>

References

- Bartsch R P, Liu K K, Bashan A and Ivanov P C 2015 Network physiology: how organ systems dynamically interact *PLoS One* **10** e0142143
- Bashan A, Bartsch R P, Kantelhardt J W, Havlin S and Ivanov P C 2012 Network physiology reveals relations between network topology and physiological function *Nat. Commun.* **3** 1–9
- Bhattacharya K, Saadia D, Eisenkraft B, Melvin Y, Warren O, Burton D and Kaufmann H 2002 Brain magnetic resonance imaging in multiple-system atrophy and Parkinson disease: a diagnostic algorithm *Arch. Neurol.* **59** 835–42
- Bilotta E, Cerasa A, Pantano P, Quattrone A, Staino A and Stramandinoli F 2010 A CNN based algorithm for the automated segmentation of multiple sclerosis lesions *EvoApplications 2010* (Berlin: Springer) pp 211–20
- Bilotta E, Cerasa A, Pantano P, Quattrone A, Staino A and Stramandinoli F 2012 Evolving cellular neural networks for the automated segmentation of multiple sclerosis lesions *Variants of Evolutionary Algorithms for Real-World Applications* (Berlin: Springer) pp 377–412
- Bozkir A S and Sezer E A 2011 Predicting food demand in food courts by decision tree approaches *Proc. Comput. Sci.* **3** 759–63
- Breiman L 2001 Random forest *Mach. Learn.* **45** 5–32
- Cerasa A, Bilotta E, Augimeri A, Cherubini A, Pantano P, Zito G, Lanza P, Valentino P, Gioia M C and Quattrone A 2012 A cellular neural network methodology for the automated segmentation of multiple sclerosis lesions *J. Neurosci. Methods* **2013** 193–9
- Crews L and Masliah E 2010 Molecular mechanisms of neurodegeneration in Alzheimer's disease *Hum. Mol. Genet.* **19** R12–20
- Criminisi A, Shotton J, Robertson D and Konukoglu E 2010 Regression forests for efficient anatomy detection and localization in CT studies *Medical Computer Vision. Recognition Techniques and Applications in Medical Imaging. MCV 2010 (Lecture Notes in Computer Science)* vol 6533, ed M Menze *et al* (Berlin: Springer) (https://doi.org/10.1007/978-3-642-18421-5_11)
- Di Paola M *et al* 2012 Multimodal MRI analysis of the corpus callosum reveals white matter differences in presymptomatic and early Huntington's disease *Cereb Cortex* **22** 2858–66
- Di Paola M, Luders E, Di Julio F, Cherubini A, Passafiume D, Thompson P M, Caltagirone C, Toga A W and Spalletta G 2010 Callosal atrophy in mild cognitive impairment and Alzheimer's disease: different effects in different stages *NeuroImage* **49** 141–9
- Di Paola M, Phillips O, Orfei M D, Piras F, Cacciari C, Caltagirone C and Spalletta G 2015 Corpus callosum structure is topographically correlated with the early course of cognition and depression in Alzheimer's disease *J. Alzheimers Dis.* **45** 1097–108
- Ivanov P C and Bartsch R P 2014 Network physiology: mapping interactions between networks of physiologic networks *Networks of Networks: the Last Frontier of Complexity (Series: Understanding Complex Systems Springer Complexity)* ed G D'Agostino and A Scala pp 203–22
- Ivanov P C, Liu K K and Bartsch R P 2016 Focus on the emerging new fields of network physiology and network medicine *New J. Phys.* **18** 100–201
- Liaw A and Wiener M 2002 Classification and regression by randomForest *R News* **2/3**

- Litvan I *et al* 1996 Clinical research criteria for the diagnosis of progressive supranuclear palsy (Steele-Richardson-Olszewski syndrome) *Neurology* **47** 1–9
- Lombardo M C, Barresi R, Bilotta E, Gargano F, Pantano P and Sammartino M 2017 Demyelination patterns in a mathematical model of multiple sclerosis *J. Math. Biol.* **75** 373–417
- Lord F M and Novick M R 1968 *Statistical Theories of Mental Test Scores* (Reading, MA: Addison-Wesley)
- MacQueen J 1965 Some methods for classification and analysis of multivariate observations *Proc. Fifth Berkeley Symp. on Math. Statist. and Prob. (Los Angeles, CA, USA)* pp 281–97
- Nigro S, Cerasa A, Zito G, Perrotta P, Chiaravalloti F, Donzuso G, Fera F, Bilotta E, Pantano P, Quattrone A and the Alzheimer's Disease Neuroimaging Initiative 2014 Fully automated segmentation of the pons and midbrain using human T1 MR brain images *PLoS One* **9** e856182014
- Oba H *et al* 2005 New and reliable MRI diagnosis for progressive supranuclear palsy *Neurology* **64** 2050–5
- Oshiro T M, Perez P S and Baranauskas J A 2012 How many trees in a random forest? *Learning and Data Mining in Pattern Recognition. MLDM 2012 (Lecture Notes in Computer Science)* vol 7376, ed P Perner (Berlin: Springer) (https://doi.org/10.1007/978-3-642-31537-4_13)
- Quattrone A *et al* 2008 MR imaging index for differentiation of progressive supranuclear palsy from Parkinson disease and the Parkinson variant of multiple system atrophy *Radiology* **246** 214–21
- Ruppert G C S, Teverovskiy L, Yu C, Falcão A X and Liu Y 2011 A new symmetry-based method for mid-sagittal plane extraction in neuroimages *ISBI 2011 (Chicago, IL, USA)* pp 285–8