Mid-Sagittal Plane detection for advanced physiological measurements in brain scans

Francesca Bertacchini [1], Rossella Rizzo [2*], Eleonora Bilotta [2], Pietro Pantano [2], Angela Luca [3], Alessandro Mazzuca [3], Antonio Lopez [3] for the Alzheimer's disease Neuroimaging Iniziative ^

[1] Evolutionary System Group, Department of Mechanical, Energy and Management Engineering, University of Calabria, Rende (CS), Italy
[2] Evolutionary System Group, Department of Physics, University of Calabria, Rende (CS), Italy
[3] Radiological Unit, Cetraro Hospital, Cetraro (CS), Italy

*** Corresponding author:** Rossella Rizzo
E-mail addresses: rossella.rizzo@unical.it, rossella.rizzo3108@gmail.com
Full postal address: Physics Department, Cubo 17B, University of Calabria, 87036, Arcavacata di Rende, CS, Italy

^ Some of the 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/wpcontent/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

Abstract

The diagnostic process of many neurodegenerative diseases, such as Parkinson, Progressive Supranuclear Palsy, etc., involves the study of brain MRI scans in order to individuate morphological markers that can highlight on the healthy status of the subject. A fundamental step in the pre-processing and analysis of MRI is the identification of the Mid-Sagittal Plane, which corresponds to the mid-brain and allows a coordinate reference system for the whole MRI scans set. To improve the identification of the Mid-Sagittal, 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 that manually identified the mid-sagittal and whose performances have been crossed with a cognitive decisional algorithm in order to define a gold standard. The comparison provided a mean percentage error of 0.96%. To further refine the automatic procedure, we trained a machine learning considering the results coming from the proposed algorithm and the gold standard. Therefore, we tested the machine learning and obtained results comparable to medical raters with a mean percentage error of 0.65%. Even if the sample of data analyzed needs to be increased, the system is promising and it could be directly incorporated into broader diagnostic support systems.

Keywords: image segmentation, k-means algorithm, machine learning, magnetic resonance imaging, mid-sagittal plane

1 Introduction

The use of magnetic resonance imaging (MRI) can provide valuable information in the detection of degenerative diseases, not just from a qualitative point of view, but even to measure 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 concerns the identification of the optimal slice on which to make these measurements. In this framework, the identification of the Mid Sagittal Plane (MSP) in MRI brain scans is crucial for detecting many of the most important neurodegenerative diseases such as Parkinson Disease (PD) (Nigro et al., 2014), Huntington's disease (HD) (Di Paola et al., 2010, 2012), Multiple Sclerosis (MS) (Cerasa et al., 2012, Bilotta et al., 2010, 2012), Alzheimer disease (AD) (Di Paola et al., 2015). In this paper, we present a fully automated method to identify the midsagittal plane, in MRI scans of healthy, Progressive Supranuclear Palsy (PSP), Parkinson's, Alzheimer and Multiple Sclerosis diseases subjects.

Parkinson's disease, for example, comes in a variety of neurological malfunctions determining pyramidal cerebellar, vegetative, and cognitive degeneration. The disease causes the nigro-striatal pyramidal worsening, involving the cerebellum and deep cerebral structures, but also neuronal degeneration of the neo-striatum. An accurate early diagnosis of PD is important both for therapeutic purposes – to target therapy more precisely to the various symptoms – and in terms of the prognosis. However, although advanced diagnostic techniques have recently been developed for PD (Oba et al., 2005; Quattrone et al., 2008), this disease suffers from a lack of universally accepted diagnostic criteria, making it difficult to distinguish it, and therefore characterized by a high rate of misdiagnosis (Litvan et al., 1996). Structural MRI is routinely used to detect early signs of PD, from a hyper-intensity of the lateral edge of the putamen and an atrophy of the brainstem; a 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 instead used to measure the arrangement of basal ganglia. It is particularly worth noting that MRI morphometry (Oba et al. 2005) has allowed the onset of 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 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. Finding the exact location of this plane has many applications. However, there is no universal agreement about the identification of MSP, as many times the dividing plane between the brain hemispheres does not correspond with the symmetry plane of the head. Instead, in patients with Alzheimer's dementia, alterations of the neurotransmitter systems and the signal transduction mechanism are very frequent, altering the cholinergic signaling system, and the production of the neurotransmitter acetylcholine (Crews & Masliah, 2010). Moreover, we can observe other cerebral alterations both macroscopic (decrease in weight and volume of the brain, due to cortical atrophy and ventricular dilatation) and microscopic (neuronal loss, glial and astroglial reaction, micro vessel alteration). The consequence of these brain modifications is the impossibility for the neurons to transmit nerve impulses, and therefore the death of the same, with consequent progressive atrophy of the brain as a whole (Crews & Masliah, 2010).

Multiple sclerosis, characterized by inflammation which results in multifocal demyelinating lesions and degeneration, with diffuse axonal loss leading to brain atrophy in the central nervous system, is a complex neurodegenerative disease. Given its relapsing/remitting cyclical behavior, MRI technique 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, usually time-consuming procedure if they are not automatized by particular segmentation algorithms (Bilotta et al., 2011; Cerasa et al., 2012).

Moreover, the improvement of MRI techniques suits to the novel field of Network Physiology, in particular in order to reach the main goal to build a first complex atlas of dynamic interactions between different brain locations and organ systems (Bartsch et al., 2015). The human organism is constituted by a complex and integrated network of different organ systems, each with its own regulatory dynamic mechanism and with dynamics of interactions between each other that define different physiological states (Ivanov & Bartsch, 2014). Changes in these networks of interactions indicate not only the change 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 a total failure of the organism, therefore mapping and studying changes in the network of interactions could help to early diagnose neurodegenerative diseases that involve other organ systems, such as Parkinson and multiple sclerosis.

The problem of identifying the mid sagittal (and in general the morphometric measurements of the brain) is because measurements are made in an environment whose variability characteristics are relevant. Consider, for example, the difference in resolution of the brain scans, depending on the employed brain scan, the time taken for the shooting, the variability of the morphology of each individual patient, the multiplicity of motion artefacts, due to technical problems or casual movements of the skull during records. Moreover, the most used method 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 this new data set. But very often, this approach is not useful because it reveals that interpolation techniques modify original data, alter brain images, making subsequent measurements unsuitable for the correct identification of the proper diseases' markers. This happens when, for instance, the entire set of MRI scans is tilted in order to have the scan plane parallel to sagittal plane. In this case, a rigid rotation is the first step in the pre-processing of the MRI set scans; then, an interpolation is required in order to represent the new MRI set as an imaginary cube, the 3D reconstructed image of whole brain, and the same is made for each voxel. In this last operation some information is lost, e.g. the information about the ratio between the different dimensions of some brain areas: the distances and volumes change, especially in the mid-brain, area of interest in the diagnostic process of the previously mentioned diseases.

To meet these needs and to support medical diagnosis, we have implemented a method for the automatic mid-sagittal identification from the raw data. Developed in Matlab, it uses a classic k-means method to identify the slice of the DICOM file containing the mid-sagittal plane. To advance the method, we have compared its performance to a gold standard gained from manual measurements, conducted by expert raters, who carefully analyzed healthy, PSP, PD, AD and MS subjects MRI brain scans. Moreover, the scans have been performed by machines having different resolutions (from 1.5 to 3T) in order to verify the independence of the method with respect to the data acquisition systems. On the same data, we trained a machine learning system in order to improve much more the system's performance. The idea is to develop fully automated systems with the capability to recognize patters relevant to medical diagnosis.

The paper is organized as follows. After this introduction, the used data sets and the methods are outlined. Results follows together with the main conclusions that can be outlined to develop cognitive systems to automatically analyze complex visual data.

2 Materials and Methods

2.1 Data set

The data set consists of 109 MRI scans, grouped as follows:

37 subject brain scans, 14 Healthy Control subjects (mean age: 52, 6 female, 8 male), 13 PD subjects (mean age: 69, 4 female, 9 male), and 10 Progressive Supranuclear Palsy (PSP) subjects (mean age: 71, 3 female, 7 male) were provided by the CNR Catanzaro (CZ, Italy) and were acquired using a 3.0 T magnetic resonance (MR) scanner (GE MEDICAL SYSTEMS DISCOVERY MR750). We used 3D T1-weighted sequence (Acquisition Plane = SAGITTAL, Inversion Time = 650 ms, Repetition Time (TR) = 9.15 ms, Echo Time (TE) = 3.67 ms, Slice Thickness = 1.0 mm, Resolution 256 x 256 pixels, Voxel Size 1.0 x 1.0 x 0.5 mm). This set of data has been used in a previous paper of some of the authors (Nigro et al., 2014).

To test if our methods are independent in respect to the specific MRI scanner used and from the Parkinson disease and its variants, 15 MS Subjects (mean age: 45, 12 female, 3 male). Data related to MS have been collected at the Neurodiagnostic Unit of the Hospital of Cetraro (CS), in compliance with the Privacy Act and current legislation (Declaration of Helsinki), 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 x 336 pixels, Voxel Size 0.762 x 0.762 x 1.0 mm, TR = 7.0286 ms, TE = 3.178 ms. Subjects data were treated according to the current privacy rights protection laws. The Ethics Committee of the Cetraro Hospital has approved the research.

Furthermore, 57 AD Subjects (mean age: 75, 29 female, 28 male) were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (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 magnetic resonance imaging (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 Alzheimer's disease (AD). For up-to- date information, see www.adni-info.org. 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 x 256 pixels, Slice Thickness = 1.0 mm, Voxel Size 1.0 x 1.0 x 1.0 mm).

In synthesis, for each group of subjects there are different technical specifications related to the type of brain scan performed and the physical characteristics of the used system. For testing our methods, we used 3D T1-weighted sequences (Acquisition Plane = SAGITTAL), restricting our interest on 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 do not weigh down computationally the software. Technical features of the datasets are summarized in Table I. All data scans have been anonymized, in obedience with the current Ethical laws.

TABLE I
TECHNICAL FEATURES OF DATASETS USED IN THIS WORK

Dataset	Subsets	Scanner Machine	Slice Resolution [pixels]	Voxel Size [mm]	Slice Thickness [mm]	Slices total number	Range considered [slice number]
CNR Catanzaro	Healthy	GE MEDICAL SYSTEMS DISCOVERY MR750 3.0 T	256 x 256	1.0 x 1.0 x 0.5	1.0	367 / 368	133-233 / 134-233
	PD		256 x 256	1.0 x 1.0 x 0.5	1.0	367 / 368	133-233 / 134-233
	PSP		256 x 256	1.0 x 1.0 x 0.5	1.0	367 / 368	133-233 / 134-233
Hospital of Cetraro	MS	Philips Achieva Rev R5 v3-rev.00 1.5 T	336 x 336	0.762 x 0.762 x 1.0	1.0	210	55-154
ADNI	AD	SIEMENS 3.0 T	240 x 256	1.0 x 1.0 x 1.0	1.0	176 / 208	38-137 / 54-153

Table I: Technical features of the datasets used in this work

2.2 K-means

K-means is part of the exclusive or partitioning-type algorithms. Given a set of n objects D and the number k of clusters, it organizes objects into separate partitions k $(k \leq n)$, where each one represents a cluster (Mac Queen, 1967). Clusters are used in order to optimize a grouping criterion, generally a function based on the distance; 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 the center of gravity.

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 k of clusters in which to partition the set, k elements $\underline{c_j} \,\epsilon\, S\ , j \,\epsilon\, \{1, \dots, k\}$, settled in a random manner, will be at first the centroids of the clusters C_j. Then, another element $\underline{x_i} \,\epsilon\, S$, that will be associated with the cluster whose centroid is closer, in according to the metric d, is randomly chosen as follows:

$$\underline{x_i} \,\epsilon\, \underline{C_j}_0 \qquad so\ that \qquad d\left(\underline{x_i}, \underline{c_j}_0\right) = \ min_{1\leq j\leq k}\, d(\underline{x_i}, \underline{c_j})$$

Cluster C_{j_0} will have a new centroid, calculated considering both $\underline{c_j}_0$ and $\underline{x_i}$. This is repeated by identifying the cluster to which another random point belongs. The process ends when the whole set has been partitioned as follows:

$$\forall\ i \,\epsilon\, \{1, \dots, n\}\ \exists\ j\ \in\ \{1, \dots, k\}\ \ such\ that\ \ \underline{x_i} \in\ \underline{c_j}$$

For the aim of this paper, the brains of the experimental subjects have been partitioned into slices on the sagittal plane and, using the k-means algorithm, we got a subdivision into 4 clusters (Fig.1).

Since we want to locate the Mid-Sagittal reference slice, we can only consider the 100/101 central slices that we will place in an INPUT folder, repeating the procedure for each subject. The goal is to identify the slice in which the difference between the different brain tissues is more marked. In our algorithm we used two main MatLab codes: *k_mean.m* and *peaks.m*.

1. *k_mean.m*
   We can divide this code into three sub-parts:
   a. Iteration of k-means method to all DICOM files in the INPUT folder of each subject. In this case:

♦ the set S is an image 2D (the slice on sagittal plane)
♦ the elements x_i are the points correspondents to the pixels
♦ the metric d is defined by

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

where $v(x_i)$ is the value in grey scale corresponding to point x_i
♦ $k = 4$
We chose a clusters number equal to 4 because experimentally we saw that in this way we can obtain the best slice partition. In fact, this choice makes it possible to distinguish quite well the different areas of the brain and in particular the mid-brain that is the region of greatest interest in defining the mid-sagittal (Fig.1). Next, the clusters are sorted in ascending order depending on the number of pixels they contain. This means that cluster 1 will contain the points corresponding to pixels representing the cerebrospinal fluid, cluster 2 representing gray matter, cluster 3 representing white matter, and finally cluster 4 corresponding to the pixels representing the background. Determining this order is possible because at the variation of the slice, at least for the central ones, the ratio between the quantity of pixels present in the various regions is always the same. The process is repeated iteratively for all files in the folder for each subject.

b. Graph creation.
The code manages to create graphs representing the amount of pixels contained in each cluster by varying the slice number, always considering only the 100/101 central slices. Note the clusters for each slice are sorted in an increasing order. This is because the choice of the number, associated with each cluster, occurs randomly when choosing the centroids between all points in the image. Consequently, once number 1 can be associated with the cluster containing the points representing the cerebrospinal

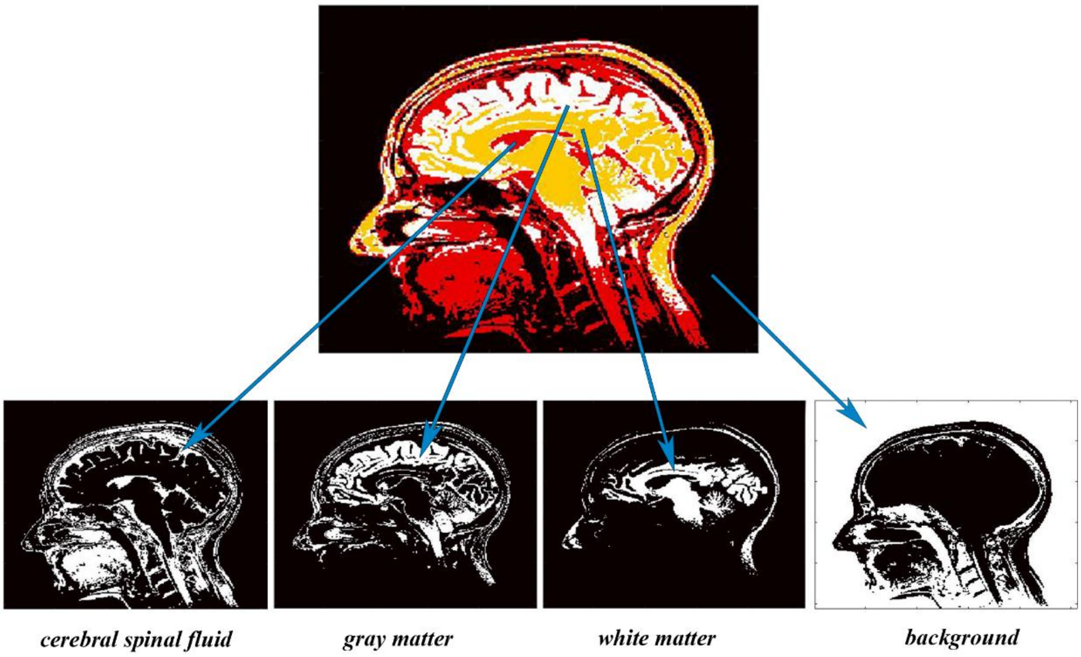


Fig. 1 First step of the code. The 2D MRIs are divided into four clusters according to the values of the gray scale. Each cluster corresponds, approximately, to a different brain tissue.

fluid, another time, 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 we can establish a two-way correspondence between the cluster identification number and the cerebral tissue (Fig.2a).

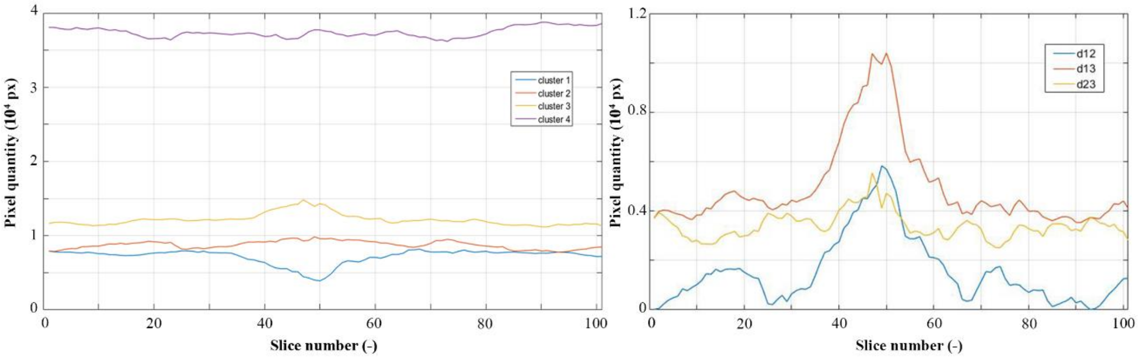


Fig. 2 (a) Second step of the code. The graph shows the amount of pixels present in each cluster by varying the slice number. (b) Third step of the code. Every curve shows the difference of quantities in pixels between two different clusters.

c. Finally, the code creates a graph representing the differences of the amount of pixel between different tissues, by varying the slice number (Fig.2b). Note that pixel quantity differences between region 4 and the other ones are not considered because cluster 4 contains pixels corresponding to background and there are, of course, no substantial differences in the number of pixels representing the background between one slice and the another, at least for the central slices.

2. *peaks.m*

This code makes it possible 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 value that the function takes at the critical point and the average of the values that the function assumes elsewhere is quite relevant. In particular, we suppose to analyze a difference curve that we call v. Now we can consider two cases:

a. the peak of the curve is situated among the 40 central slices

In this case are considered only the 40 central slices to establish if the curve has a maximum or a minimum, in formulas

- ♦ 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 $(j) = min_{30\leq i\leq 70}\, v(i)$
- ♦ if $\left| min_{30\leq i\leq 70}\, v(i) - \frac{v(30)+v(70)}{2} \right| < | max_{30\leq i\leq 70}\, v(i) - \frac{v(30)+v(70)}{2} |$, then v has an absolute maximum in j such that $v(j) = max_{30\leq i\leq 70}\, v(i)$**.**

We can note that is not considered the average of the values that v assumes elsewhere to make the code faster and because experimentally is seen that are not substantial changes between $v(30)$ and $v(i), i \in \{1, \dots, 30\}$, the same thing is valid for the values assumed in the last 30 central slices.

b. the peak of the curve is situated outside the 40 central slices or close enough to the edges (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 central 40 slices range) to have a bifurcation point in the evaluation of the criticality o that point in the curve.

In this case we consider the whole interval of the 100/101 central slices. The minimum or the maximum of the curve is chosen in the same manner, but the edges considered are the first slice and last one.

We indicated the difference curve between clusters i and j as d_{ij}, $i, j \in \{1,2,3\}$.

Finally, we computed the arithmetic average between the number of the slice corresponding to the peak of each curve. The output of the previous code is a vector $p = (p_1, p_2, p_3)$, where $p_{h,h \in \{1,2,3\}}$ is the number of the slice correspondent to the maximum or minimum of the difference curve. 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.

2.3 Implementation of the system

The block diagram of our proposed algorithm is represented in Fig.3.

Regardless of the brain scans' resolution, the central 100/101 slices for each subject were provided as input to the developed system. The code works on each slice improving k-means method, shown from second cycle. Then code creates a quantity cluster graphic and computes the difference between quantities in pixels of different clusters. Finally, the average number of slices, which correspond to critical point of difference functions, detects the number of reference slice of Mid-Sagittal.

2.4 Inter raters reliability and gold standard definition

To obtain a gold standard on which compare the performance of the developed algorithm, 4 expert raters manually segmented the mid-sagittal plane for each subject of the considered sample.

In order to arrive to a perfect agreement by mathematical algorithms in delineating the mid-sagittal slice that could be used as standard to compare the performance of the algorithm, we used a statistical-mathematical model, which allows to outline either an evaluation of the performance of the individual rater, and an analysis of 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* are assigned to each response of raters, obtaining, then, a proportional rating. The results gave us a percentage of agreement among raters. This means that if, for example, raters agree in 61% of the cases, out of the 109 cases considered, they do not agree in the remaining 39% of the considered sample. 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 a so low percentage of success, it could not be considered an optimal gold standard on which test the performance of our automatic algorithm.

The central problem the raters found in the measurement of the mid-sagittal, was that there is no one 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 proper mid-sagittal configuration. Therefore, the choice could be among the slices, which belong to previously defined sets of slices, to which we assigned the values for the rating categories or levels. For satisfying this need, we implemented the polychromous rating, giving a score of 2, 3 and 4, for difference among raters of 2, 3 and 4 slices. This method, while showing improved agreement percentages, exhibited many downsides as well: we can have a percentage of agreement high enough only when the raters individuate exactly the same slice.

To establish the agreement between the raters on individuation of the mid-sagittal for each subject the following procedure was applied.

- In the case of the presence of a majority of agreement among the raters upon a slice as mid-sagittal reference, that slice was chosen as gold standard
- Otherwise, a random way to choose the mid-sagittal reference slice between the different slices individuated by raters was employed

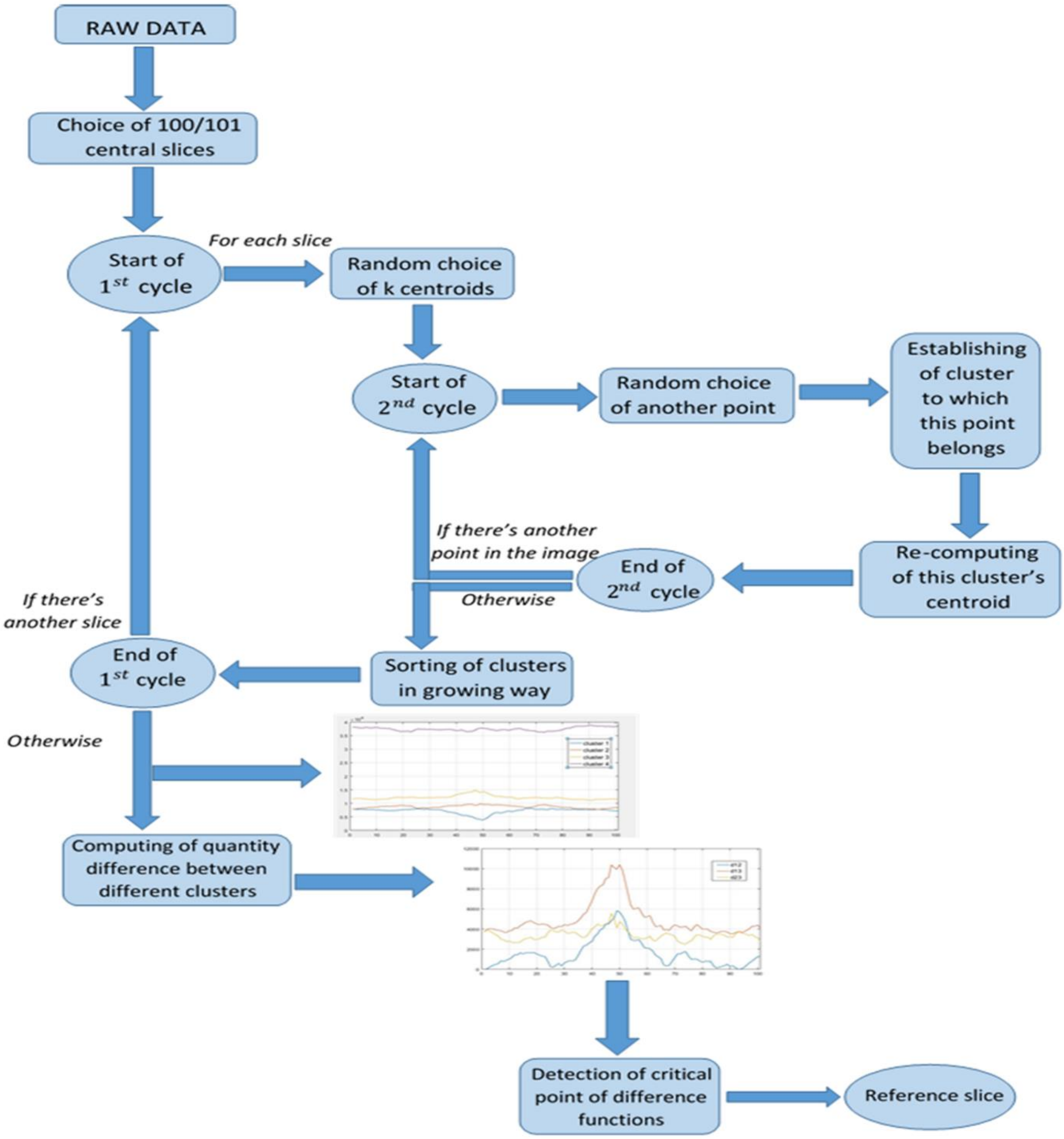

RAW DATA
Choice of 100/101 central slices
Start of 1st cycle
For each slice
Random choice of k centroids
Start of 2nd cycle
Random choice of another point
Establishing of cluster to which this point belongs
Re-computing of this cluster's centroid
End of 2nd cycle
If there's another point in the image
Otherwise
Sorting of clusters in growing way
End of 1st cycle
If there's another slice
Otherwise
Computing of quantity difference between different clusters
Detection of critical point of difference functions
Reference slice

Fig. 3 The diagram of the implemented algorithm.

Then the gold standard was compared with the results obtained from the 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 between the slices corresponding to the peak of all curves, in order to study the reliability of each curve and the average to the gold standard.

2.5 Machine learning approach

Finally, to try to further improve these results, we built a machine learning tool that can be used to find a better algorithm than the arithmetic average.

In this case, the three results of the algorithm are entered as input and the target value represented by the gold standard is the output. The method used is the Random Forest, chosen automatically by Mathematica®.

The machine learning allowed for the automatic forecasting of the mid-sagittal, with any new sample of data, as all the procedures have been embodied into the system and they automatically use the stored data as a computational benchmark.

3 Results

3.1 The algorithm performance

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

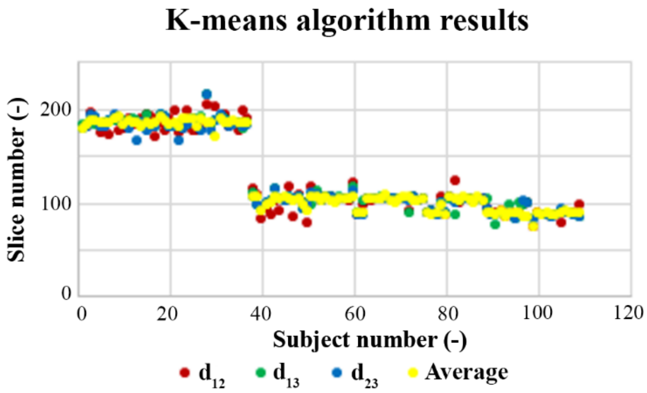


Fig. 4. Results of the automatic k-means algorithm for the entire dataset of   subjects analyzed in this study. R_{ij} , i, j ∈ {1,2,3} indicates the slice number result coming from the critical point of the corresponding difference curve d_{ij} .

Observing the first results of the algorithm, we can notice that the first 37 subjects have a MRI set of 367/368 slices, thus the automatic method returns a value corresponding to the MSP around the 186^{th}, whereas the rest of the subjects, who have a MRI set of 176/210 slices, show a value around the 98^{th} for the MSP.

3.2 Inter-raters agreement and gold standard emergence

To improve the results, we thought to use the expert manual identification in order to arrive to an inter-raters reliability that can be considered a gold standard on which compare the performance of the algorithm. Results for the manual segmentation of the mid-sagittal performed by expert raters in the different data subsets are reported in Fig.5. The agreement between raters has a standard deviation SD = 1.85549 for Healthy Subject, SD = 1.277988 for PD subjects, SD = 1.919522 for PSP subjects, SD = 2.320752 for MS Subjects, SD = 1.400329 for AD Subjects.

We can observe that raters differ among themselves in a relevant way, if we consider a dichotomous approach, whereby they can reach the identification target or not. To start optimizing the performance of our system, we calculated the arithmetic average between the evaluations of the raters and evaluated the performance of the raters on the average (Fig.6). In this way, we obtained the distribution of the performances of the raters on the arithmetic average. We realized that each rater has a different performance which can diverge very much on the frequency of right evaluation.

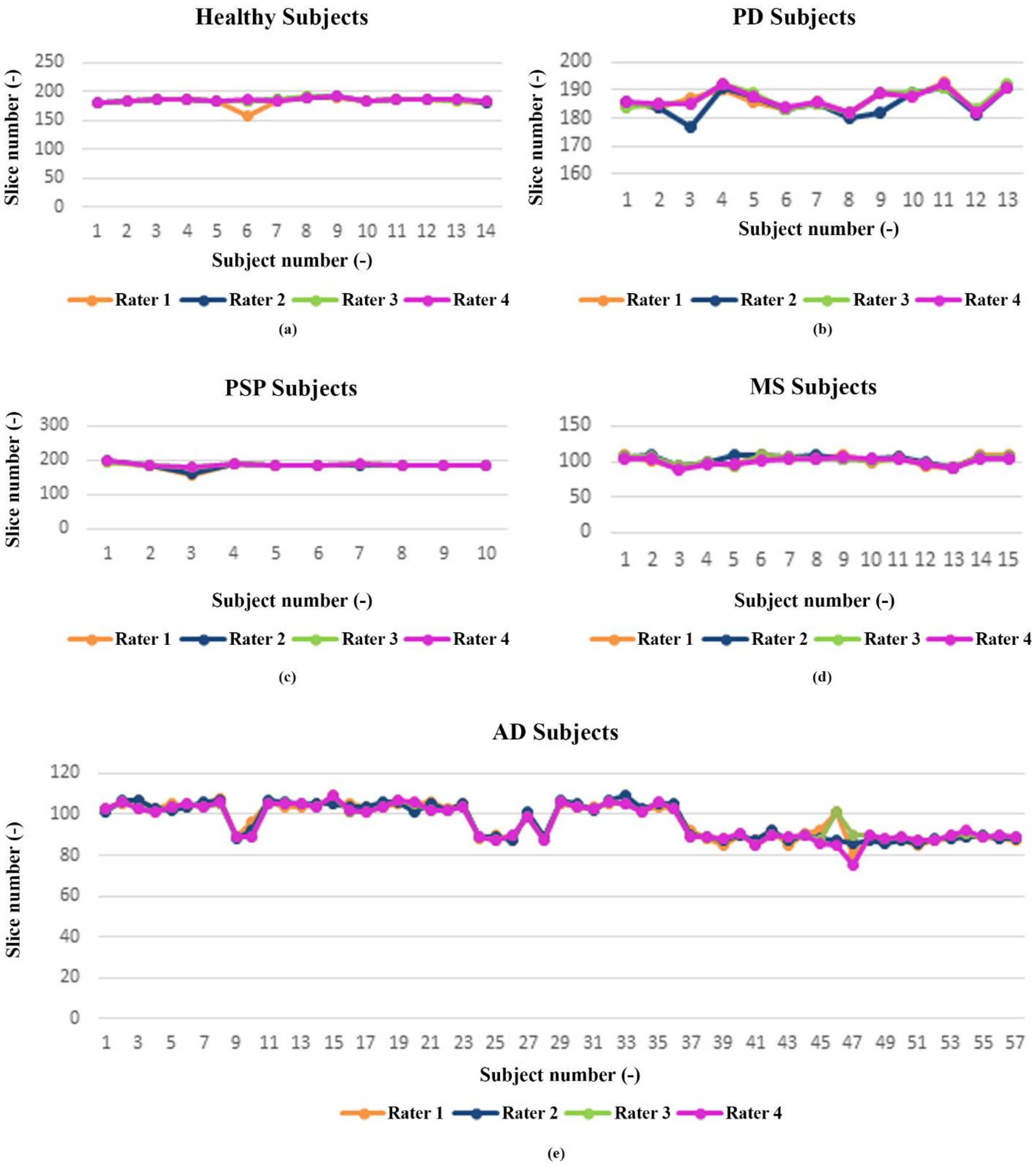


Fig. 5 Medical expert identification of the mid-sagittal – Agreement between raters.

For the gold standard ratings, as said before, we used different steps. The first step foresaw the use of the dichotomous model. With this function we have obtained results of manual segmentation agreement among raters, which is more than 0.61 in a range [0,1], where 0 is complete disagreement and 1 is complete agreement between raters. This means that the mathematical model implemented is built on a function that evaluates agreements or non-agreements among raters. This basic model considered an agreement where the variation is “yes, it is the same”, “no, it is different”.

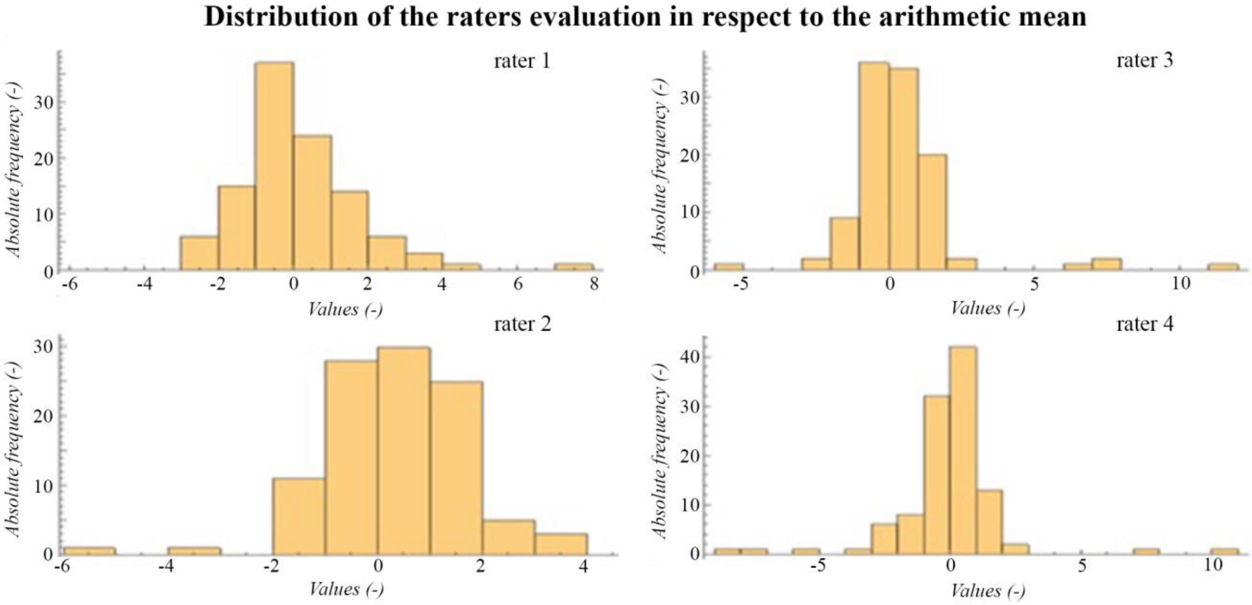


Fig. 6 Distribution of the raters performance on the arithmetic mean.

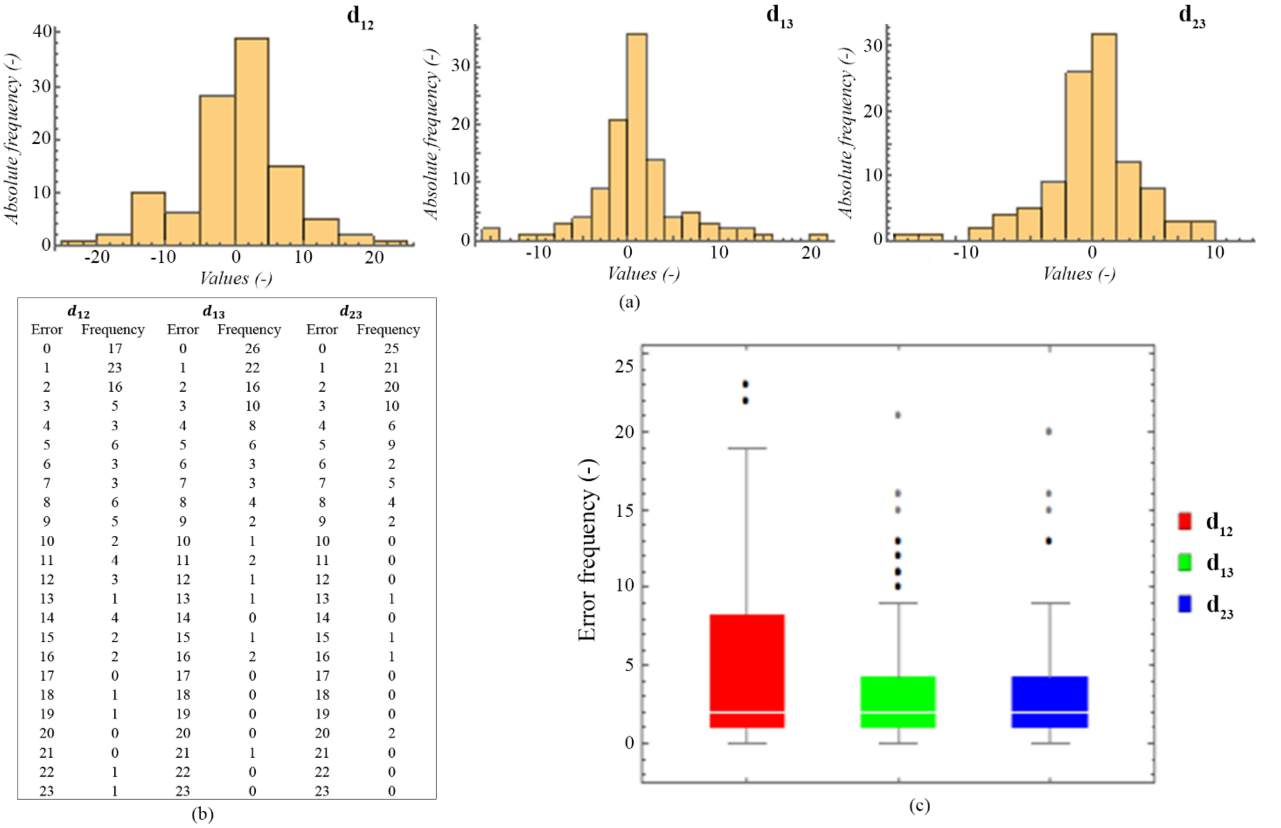
3.3 Going further to create a cognitive decisional algorithm

The previous algorithm is not completely convincing for the determination of the gold standard, even because it could be useful if we could assign a weight to some of the raters, but obviously all the raters are all experts as well so their opinions have the same weight in the manual identification of the mid-sagittal. Besides, a gold standard based on the arithmetic average between the slices individuated by raters as mid-sagittal is not much reliable. For example, if we have three raters that have indicated the same value x and only one that has indicated the value y, it seems reasonable that the correct value is x and not the average of $\{x, x, x, y\}$. We have, therefore, created the "gold standard" sequence, incorporating a decision-making process and creating a cognitive algorithm to support the choice. The decision-making process is as follows:

1. If three or four raters agree on the x value, choose that value;
2. If 2 raters agree on the value x and the other two indicate a different value y and z with y and z different, choose x;
3. If two raters agree on x and the other two on y, choose one of the two random values;
4. If all the four raters indicate different values, choose one randomly.

3.4 Comparison gold standard - algorithm

Fig.7 shows results about slices distributions individuated as peak of the three curves compared to the gold standard, shown in the histograms (Fig.7a), distribution of the results grouped according to the error's frequency (Fig.7b), and the corresponding distribution in quartiles of the results of the algorithms (Fig.7c). From the tables in Fig.7b 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. 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 can notice that the third algorithm has the least dispersion, as is also clear from the observation of Fig.7c.



(a)

d_{12}		d_{13}		d_{23}	
Error	Frequency	Error	Frequency	Error	Frequency
0	17	0	26	0	25
1	23	1	22	1	21
2	16	2	16	2	20
3	5	3	10	3	10
4	3	4	8	4	6
5	6	5	6	5	9
6	3	6	3	6	2
7	3	7	3	7	5
8	6	8	4	8	4
9	5	9	2	9	2
10	2	10	1	10	0
11	4	11	2	11	0
12	3	12	1	12	0
13	1	13	1	13	1
14	4	14	0	14	0
15	2	15	1	15	1
16	2	16	2	16	1
17	0	17	0	17	0
18	1	18	0	18	0
19	1	19	0	19	0
20	0	20	0	20	2
21	0	21	1	21	0
22	1	22	0	22	0
23	1	23	0	23	0

(b)

(c)

Fig. 7 Results about slices distributions compared to the gold standard, shown in the histograms (a), distribution of the results grouped according to the error's frequency (b), and the distribution in quartiles of the results of the algorithms (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 coming from the three curves as the opinion of three different persons 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 to 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 average absolute error, it is more logical to compute the relative error among the whole set of slices. Therefore, we must divide the absolute error by the total number of slices for each subject and compute the average across the subjects. The relative errors averaged across all subjects are equal to e_{12} = 0.02025, e_{13} = 0.01439, e_{23} = 0.01195 respectively, thus the error drops to about 1.20% in the case of the best algorithm, while it is at 2.03%, in the case of the worst algorithm.

Since we want to localize the reference slice where the differences between brain tissues is more marked and there is no reason to assign a larger weight to a difference curve than another, we computed an arithmetic average of results extracted from the three curves. The relative error averaged across the subjects is $e_{ave} = 0.00961$, obtaining a better result even than the best algorithm d_{23}. The standard deviation is SD_{ave} = 4.23703, lower than SD_{23} = 4.49718. A comparison with the other distributions (considering the absolute errors on the slices) is shown in Fig.8.

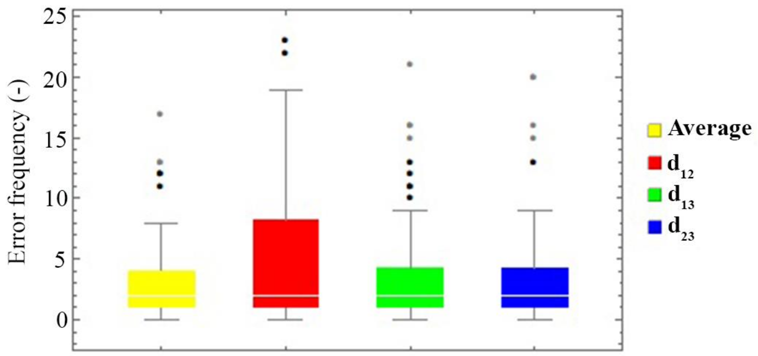


Fig. 8. Comparison with the distributions of the three algorithms and their average, considering the absolute error on the slices.

For whom to concern the machine learning approach, we employed the first step in the application of machine learning techniques. We trained the machine with the dataset of 109 subjects and we tested it on the same dataset. Comparing the results with the gold standard, we obtained an average relative error e_{mlr} = 0.00654 and an average absolute error e_{mla} = 1.9633, a value that is very close to the absolute error of the worst rater ($e_1 = 1.77982$). Instead, the standard deviation is SD_{ml}= 2.06347, lower than the standard deviation of the best rater ($SD_3 = 2.41502$). All these results show that the developed system is already comparable to the performance of the raters and less dispersive.

4 Conclusions

From the obtained results, we demonstrated how the system improved its performance, increasing its sensitivity and accuracy, 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 to detect this variability, in turn they have many drawbacks in the visual rendering of data. So, the problem we faced is highly sensitive to the initial data. Consequently, each subject of the sample has been carefully analyzed, adopting the technique of polychromous ratings, which by enlarging the intervals, specified better the sensitivity and accuracy of the developed tool. Moreover, the machine learning developed allows the forecasting of the mid-sagittal 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, which are used as a computational benchmark, we can forecast any set of data, independently of the MRI systems and neurodegenerative diseases. Continuing along this path can provide excellent results, optimizing the system to make it as sensitive and reliable as a human expert. Indeed, the next step in this framework is collecting a larger dataset and test the trained machine learning on it. Cognitive systems are very important and potentially they can be embodied into the same MRI machine, as can be done for cognitive function developed for smart robots (Bertacchini et al., 2017), for improving visual systems (Abdechiri et al., 2017), or to mathematically model specific patterns of multiple sclerosis (Lombardo et al., 2017). Besides, this method could find applications in the individuation of different brain locations, key points in the understanding of brain network interactions and in the connections with other organ systems. Indeed, detecting particular brain areas responsible of the strongest connection with a particular organ system during a particular physiological state could help to discriminate a pathological picture in the whole organism from a physiological one (Bashan et al., 2012).

5 Acknowledgements

Regarding 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 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: 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 (www.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.

References

[1] S. Nigro, A. Cerasa, G. Zito, P. Perrotta, F. Chiaravalloti, G. Donzuso, F. Fera, E. Bilotta, P. Pantano, A. Quattrone, the Alzheimer's Disease Neuroimaging Initiative, "Fully automated segmentation of the pons and midbrain using human T1 MR brain images", *PloS One*, vol. 9, no. 1, pp. e856182014, Jan. 2014.

[2] M. Di Paola, E. Luders, F. Di Iulio, A. Cherubini, D. Passafiume, P.M. Thompson, C. Caltagirone, A. W. Toga, G. Spalletta, "Callosal atrophy in mild cognitive impairment and Alzheimer's disease: different effects in different stages", *Neuroimage*. vol. 49, no. 1, pp. 141-149, Jan. 2010.

[3] M. Di Paola, E. Luders, A. Cherubini, C. Sanchez-Castaneda, P.M. Thompson, A. W. Toga, C. Caltagirone, S. Orobello, F. Elifani F. Squitieri, U. Sabatini, "Multimodal MRI analysis of the corpus callosum reveals white matter differences in presymptomatic and early Huntington's disease", *Cereb Cortex*, vol. 22, no. 12, pp. 2858-2866, Jan. 2012.

[4] A.Cerasa,E.Bilotta,A.Augimeri,A.Cherubini,P.Pantano,G.Zito,P.Lanza,P.Valentino,M.C.Gioia,A.Quattrone,"A cellular neural network methodology for the automated segmentation of multiple sclerosis lesions",*J Neurosci Methods*,vol.2013,no.1,pp.193-199,Jan.2012.

[5] E. Bilotta, A. Cerasa, P. Pantano, A. Quattrone, A. Staino, F. Stramandinoli, "A CNN based algorithm for the automated segmentation of multiple sclerosis lesions", in *EvoApplications 2010*, Berlin, Heidelberg, 2010 pp. 211-220.

[6] E. Bilotta, A. Cerasa, P. Pantano, A. Quattrone, A. Staino, F. Stramandinoli, "Evolving cellular neural networks for the automated segmentation of multiple sclerosis lesions", in *Variants of Evolutionary Algorithms for Real-World Applications*, Berlin, Heidelberg, Germany: Springer, 2012, pp. 377-412.

[7] M. Di Paola, O. Phillips, M. D. Orfei, F. Piras, C. Cacciari, C. Caltagirone, G. Spalletta, "Corpus callosum structure is topographically correlated with the early course of cognition and depression in Alzheimer's disease", *J Alzheimers Dis*, vol. 45, no. 4, pp. 1097-1108, Apr. 2015.

[8] H. Oba, A. Yagishita, H. Terada, A. J. Barkovich, K. Kutomi, T. Yamauchi, S. Furui, T. Shimizu, M. Uchigata, K. Matsumura, M. Sonoo, M. Sakai, K. Takada, A. Harasawa, K. Takeshita, H. Kohtake, H. Tanaka, S. Suzuki, "New and reliable MRI diagnosis for progressive supranuclear palsy", *Neurology*, vol. 64, no. 12, pp. 2050-2055, Jun. 2005.

[9] A.Quattrone, G.Nicoletti, D.Messina, F.Fera, F.Condino, P. Pugliese, P.Lanza, P.Barone, L.Morgante, M.Zappia, U.Aguglia, O.Gallo, "MR imaging index for differentiation of progressive supranuclear palsy

from Parkinson disease and the Parkinson variant of multiple system atrophy”,*Radiology*, vol.246, no.1, pp.214-221, Jan.2008.

[10] I. Litvan, Y. Agid, D. Calne, G. Campbell, B. Dubois, R. C. Duvoisin, C. G. Goetz, L. I. Golbe, J. Grafman, J. H. Growdon, M. Hallett, J. Jankovic, N. P. Quinn, E. Tolosa, D. S. Zee, “Clinical research criteria for the diagnosis of progressive supranuclear palsy (Steele-Richardson-Olszewski syndrome)”, *Neurology*, vol. 47, no. 1, pp. 1-9, Jul 1996.

[11] K. Bhattacharya, D. Saadia, B. Eisenkraft, Y. Melvin, O. Warren, D. Burton, H. Kaufmann, “Brain magnetic resonance imaging in multiple-system atrophy and Parkinson disease: a diagnostic algorithm”, *Arch Neurol*, vol. 59, no. 5, pp. 835–842, Jan. 2002.

[12] G. C. S. Ruppert, L. Teverovskiy, C. Yu, A. X. Falcão and Y. Liu, "A new symmetry-based method for mid-sagittal plane extraction in neuroimages”, in *ISBI 2011*, Chicago, IL, USA, 2011, pp. 285-288.

[13] L. Crews, E. Masliah, “Molecular mechanisms of neurodegeneration in Alzheimer's disease”, *Human Molecular Genetics*, vol. 19, no. R1, pp. R12–R20, Apr. 2010.

[14] J. MacQueen, “Some methods for classification and analysis of multivariate observations”, in *Proc. Fifth Berkeley Symp. on Math. Statist. and Prob.*, Los Angeles, CA, USA, 1965, pp. 281-297.

[15] F. M. Lord, M. R. Novick, *Statistical theories of mental test scores*. Reading: Addison-Wesley, New York, USA: ETS, 1968.

[16] F. Bertacchini, E. Bilotta, P. Pantano, “Shopping with a robotic companion”, *Comput Human Behav*, vol. 77, pp.382-395, Dec. 2017.

[17] M. Abdechiri, K. Faez, H. Amindavar, E. Bilotta, “The chaotic dynamics of high-dimensional systems”, *Nonlinear Dyn*, vol. 87, no. 4, pp. 2597-2610, Mar. 2017.

[18] M. C. Lombardo, R. Barresi, E. Bilotta, F. Gargano, P. Pantano, M. Sammartino, “Demyelination patterns in a mathematical model of multiple sclerosis”, *J Math Biol*, vol.75, no.2, pp.373-417, Aug.2017.

[19] R. P. Bartsch, K. K. Liu, A. Bashan, P. C. Ivanov, “Network physiology: how organ systems dynamically interact.” *PLOS ONE*, vol.10, no.11, pp. e0142143, November 2015.

[20] P. C. Ivanov & R. P. Bartsch, “Network Physiology: Mapping Interactions Between Networks of Physiologic Networks.”, in *Networks of Networks: the last Frontier of Complexity*, edited by G. D'Agostino and A. Scala, Series: Understanding Complex Systems Springer Complexity, pp. 203-222, 2014.

[21] A. Bashan, R. P. Bartsch, J. W. Kantelhardt, S. Havlin, P. C. Ivanov, “Network physiology reveals relations between network topology and physiological function.” *Nature Communications*, vol. 3, no. 702, February 2012.

[22] P. C. Ivanov, K. K. Liu, R. P. Bartsch, “Focus on the emerging new fields of network physiology and network medicine”, *New Journal of Physics* vol. 18, no. 10, pp. 100-201, October 2016.

[23] S. Wolfram. (2017, July). Mathematica. The world’s definitive system for modern technical computing. [software]. Available: http://www.wolfram.com/mathematica