DTI measurements for Alzheimer's classification

Tommaso Maggipinto et al 2017 Phys. Med. Biol. 62 2361

DTI measurements for Alzheimer's classification

Tommaso Maggipinto[†¶], Roberto Bellotti[†¶], Nicola Amoroso[†¶], Domenico Diacono[¶], Giacinto Donvito[¶], Eufemia Lella[†¶], Alfonso Monaco[¶], Marzia Antonella Scelsi[‡] and Sabina Tangaro[¶], for the Alzheimer's Disease Neuroimaging Initiative*

[†]Dipartimento Interateneo di Fisica "M. Merlin", Università degli Studi di Bari "A. Moro", Via Giovanni Amendola, 173, 70125 Bari, Italia

[¶]Istituto Nazionale di Fisica Nucleare, Sezione di Bari, Via Orabona, 4, 70123, Bari, Italia

[‡]Translational Imaging Group, Centre for Medical Image Computing, University College London, Gower Street, London, NW1 2HE, UK

Abstract

Diffusion Tensor Imaging (DTI) is a promising imaging technique that provides insight into white matter microstructure integrity and it has greatly helped identifying white matter regions affected by Alzheimer's Disease (AD) in its early stages. DTI can therefore be a valuable source of information when designing machine-learning strategies to discriminate between healthy control (HC) subjects, AD patients and subjects with Mild Cognitive Impairment (MCI). Nonetheless, several studies have reported so far conflicting results, especially because of the adoption of biased feature selection strategies. In this paper we firstly analyzed DTI scans of 150 subjects from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database. We measured a significant effect of the feature selection bias on the classification performance (p-value < 0.01), leading to overoptimistic results (10% up to 30% relative increase in AUC). We observed that this effect is manifest regardless of the choice of diffusion index, specifically fractional anisotropy and mean diffusivity. Secondly, we performed a test on an independent mixed cohort consisting of 119 ADNI scans; thus, we evaluated the informative content provided by DTI measurements for AD classification. Classification performances and biological insight, concerning brain regions related to the disease, provided by cross-validation analysis were both confirmed on the independent test.

Index Terms

Alzheimer's disease, DTI, Random Forests, Feature selection.

I. Introduction

ALZHEIMER's Disease (AD) is the most common type of progressive neurodegenerative disorder, affecting millions of people worldwide. It is characterized by different stages, ranging from a pre-dementia phase to a final stage in which the patient is completely dependent from external assistance. Estimates indicate that 75% of dementia cases in the world, more than 25 million people, are of Alzheimer's type [1]. Nevertheless, the investigation of novel biomarkers and strategies to predict and model its onset needs further investigation [2]. In particular, the investigation of biological markers aimed at diagnosing the disease promptly is crucial [3]. Mild Cognitive Impairment (MCI) is an intermediate state between healthy aging and AD, which represents an early state of abnormal cognitive function and is thus considered a good target for this investigation.

Over the past twenty years, several studies based on structural magnetic resonance imaging (sMRI) highlighted the significant role played by brain atrophy in AD diagnosis [3], [4], [5]. Since 1980s it is also known that, besides a widespread gray matter atrophy, AD is characterized by a progressive disconnection of cortical and subcortical regions because of white matter (WM) injury [6], [7], [8]. However, conventional MRI is not able to highlight the structure of WM regions due to their homogeneous chemical composition.

Diffusion Tensor Imaging (DTI) is able to track and quantify water diffusion along white matter fiber bundles and can thus provide useful information regarding their integrity [9], [10]. Fractional anisotropy (FA) and mean diffusivity (MD) are among the invariants derived from the diffusion tensor that are closely related to white matter integrity [11]. Water diffusion along a healthy axon is highly anisotropic, being constrained almost completely to one direction, that is the fibre axis, and thus high values of FA and low values of MD describe a non-pathological scenario. FA and MD maps can be visualized as conventional gray-scale images and can be subsequently analyzed by means of classification tools. In recent years, DTI has revealed itself as a very promising imaging modality to discriminate between healthy control (HC) subjects, AD patients and subjects with MCI. An analysis approach commonly found in literature consists in the computation of FA and MD maps (or other diffusion indices), followed by the identification of the most representative voxels; these voxels are then fed into machine-learning algorithms to automate the classification.

For the discrimination HC/AD, Mesrob et al. [12] adopted a Support Vector Machine (SVM) classifier and a region of interest (ROI)-based approach; Dyrba et al. [13] used a ROI-based approach and a multimodal SVM combining DTI indices

*Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf with gray matter volume derived from sMRI; Amoroso et al. [14] adopted topological measurements based on probabilistic tractography; Schouten et al. [15] used a ROI-based approach in combination with Elastic Net Regression. For the classification HC/MCI, Cui et al. [16] used subcortical volumetric features extracted using a segmentation algorithm together with FA values obtained for white matter regions of interest. Dyrba et al., in [17], used a ROI-based approach and SVMs on a multicentric dataset and apply variance reduction methods.

The best performances in literature for the HC/MCI classification, using a single DTI modality, can be found in Haller et al. [18] and O'Dwyer et al. [19]. In these works, a voxel-based approach is used considering as features the voxel intensities in the diffusion maps. However, as also remarked in [19], in each of the above mentioned work, the methodological procedure relies on an *a priori* feature selection performed on the entire dataset to be analyzed. This procedure, also known as non-nested feature selection, circular analysis, or double dipping, chooses the most discriminative voxels by using also the test set, thus introducing a bias in the classification model. A non-nested feature selection necessarily leads to overestimate the numerical values of accuracy and area under the ROC curve (AUC). On the contrary, a nested feature selection is obtained when the selection procedure is performed blind to the test set.

The practice of double dipping and its dangers are well known to the statistics and computer science community, and have been extensively described in the literature [20], [21]. Although recommendations and best practices are available [22], the field of neuroimaging is still widely populated by studies that noticeably perform non-nested feature selection, claiming classification performances close to perfect accuracy. The effects of double dipping on classification performances in neuroimaging studies have been quantitatively assessed when dealing with functional brain data, such as fMRI [22] or MEG [23], and with data derived from structural T1-weighted MR imaging (cortical thickness) in [24]. However, some of the image classification studies involving DTI cited above seem to be affected by such feature selection bias, and to date no study has yet investigated to which extent the reported performances are inflated by its presence.

In this work we used DTI images for classification tasks in AD; considering the profitability of using classification trees in the context of machine learning techniques applied to AD [25], [26], we used a Random Forest approach. The main aim of this work is to perform a comparative study between nested and non-nested feature selection on the same data set. To the best of our knowledge, this is the first study attempting to measure the bias introduced by non-nested feature selection, from now onward feature selection bias (FSB), in the classification of DTI images with a fair comparison, i.e., measuring the effect on the same fixed data set. We finally confirmed on an independent test set how the FSB impacts the reliability of estimated classification performances.

II. Materials

Data used in preparation of this article 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, positron emission tomography, other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of Mild Cognitive Impairment and early Alzheimer's Disease.

The images analyzed for this study are diffusion-weighted scans of 150 subjects (50 HC, 50 AD patients and 50 MCI), both males and females, aged 55 to 90, from the ADNI-GO and ADNI-2 phases. Scans were randomly selected from baseline and follow-up study visits. HC subjects show no signs of depression, mild cognitive impairment or dementia; participants with AD are those who meet the NINCDS/ADRDA criteria for probable AD; MCI subjects have reported a subjective memory concern, but without any significant impairment in other cognitive domains: they substantially preserved everyday activities with no signs of dementia. Two MCI levels (early or late) are usually distinguished according to the Wechsler Memory Scale Logical Memory II. For this study, we used a balanced group of 25 early and 25 late MCI, but these labels were not taken into account in the classification tasks. Further details about diagnostic criteria for ADNI study participants can be found at http://adni.loni.usc.edu/study-design/background-rationale/.

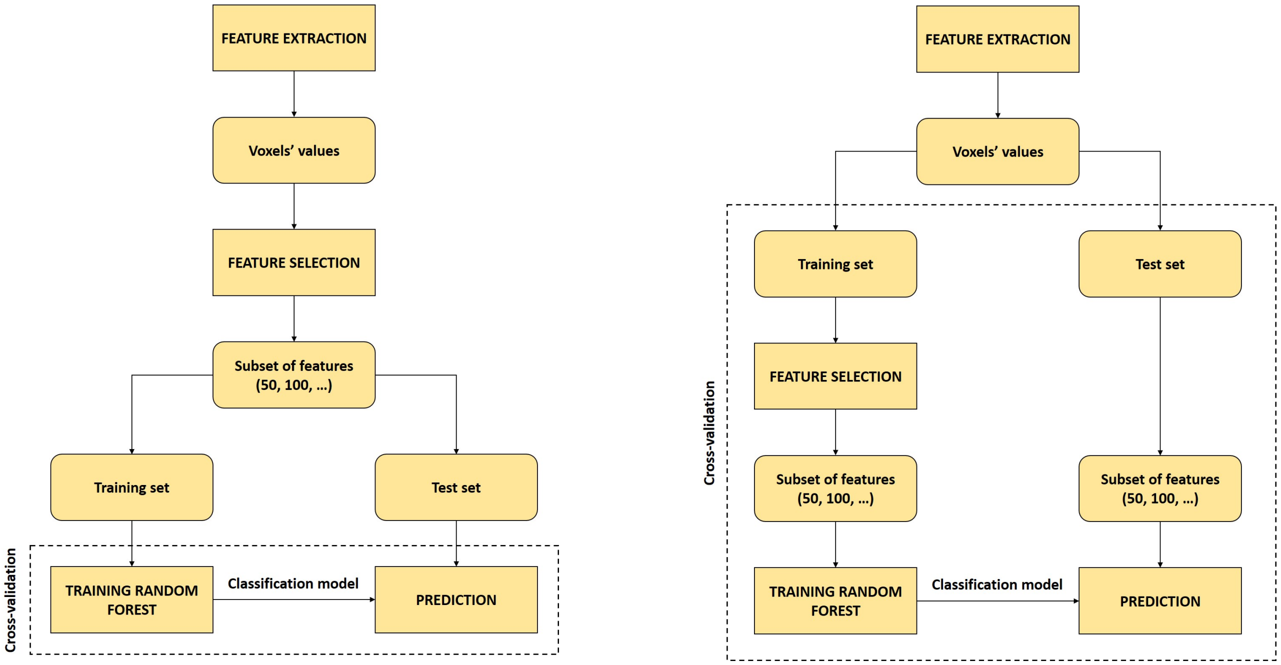
In order to evaluate the proposed algorithm on an independent test set, a second different set of scans from the ADNI database was also considered, consisting of 40 HC, 40 MCI (22 early and 18 late) and 39 AD. This second test set included both male and female subjects, and was age-matched with the training sample. Diffusion-weighted scans were acquired using a 3 T GE Medical Systems scanner with 41 gradient directions (b = 1000 s/mm^2); in addition to these, 5 images with negligible diffusion effects (b_0 images) were acquired as reference scans for subsequent analysis.

III. Methods

The main steps of our analysis are outlined in the flowcharts in Fig. 1a and Fig. 1b.

A. Image preprocessing

Diffusion-weighted images were preprocessed using the FMRIB Diffusion Toolbox, included in the FSL software [27]. Preprocessing comprised: (i) conversion to Nifti format; (ii) extraction of gradient directions and b-values; (iii) correction for eddy currents and head motion; (iv) skull-stripping using the Brain Extraction Tool (BET).



(a) Non-nested approach

(b) Nested approach

Fig. 1: Flowcharts of the performed analyses: (a) non-nested feature selection and (b) nested feature selection. For readability, they only consider the steps following the feature extraction phase.

B. *Diffusion tensor fitting*

After preprocessing, a single diffusion tensor was fitted at each voxel in the image, using DTIfit. From the diffusion tensor, fractional anisotropy (FA) and mean diffusivity (MD) were then calculated. By definition, these two invariants are related to the eigenvalues of the diffusion tensor λ_1, λ_2, λ_3 by [9], [11]:

$$FA = \sqrt{\frac{1}{2}} \frac{\sqrt{\left((\lambda_1 - \lambda_2)^2 + (\lambda_2 - \lambda_3)^2 + (\lambda_3 - \lambda_1)^2\right)}}{\sqrt{\lambda_1^2 + \lambda_2^2 + \lambda_3^2}} \tag{1}$$

$$MD = \frac{\lambda_1 + \lambda_2 + \lambda_3}{3} \tag{2}$$

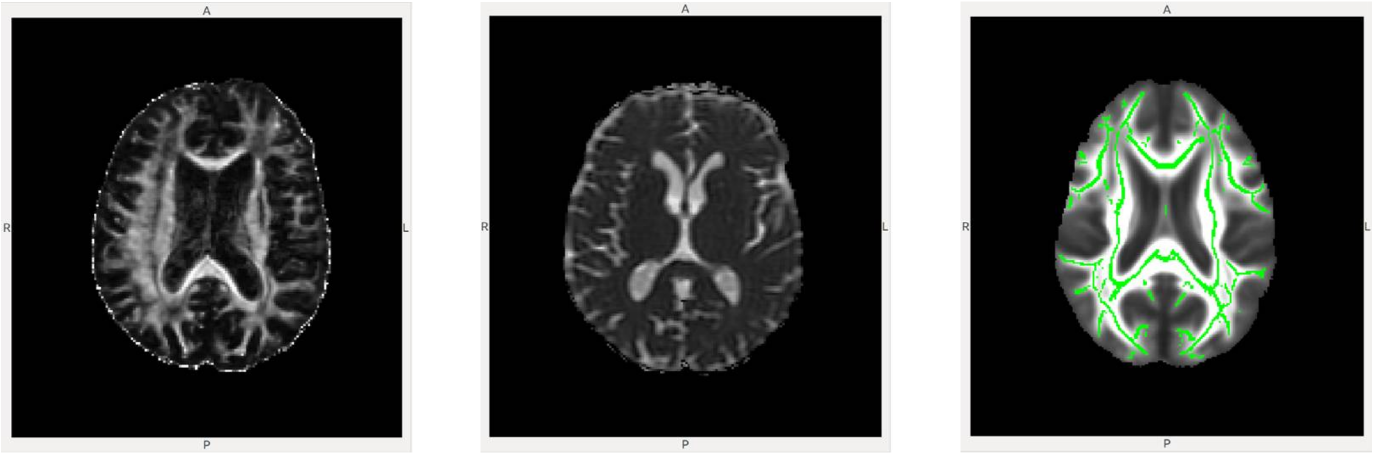
FA and MD maps were computed for each subject in the study. FA quantifies the degree of anisotropy of any diffusion process, taking values in the range $[0, 1]$. Diffusion is said to be isotropic for FA $= 0$, whereas a value of 1 indicates that diffusion is fully constrained along one direction. Water diffusion in an healthy axon or fiber bundle is highly anisotropic and constrained almost exclusively to the fiber direction, due to the presence of the surrounding myelin sheath. FA is typically higher in white matter than in grey matter or cerebrospinal fluid (CSF), and is an established marker of microstructural fibre integrity, in the sense that its value decreases in presence of axonal degeneration or demyelination. MD instead relates to the mean free path of water molecules in all directions. It is typically of the same order of magnitude in gray and white matter, while being consistently higher in the CSF, and can be regarded as an inverse measure of membrane density. Increases in MD in white matter areas are therefore indicative of myelin disruption or loss [28], [29].

C. *Tract-Based Spatial Statistics*

After diffusion tensor fitting, FA and MD maps need to be carefully aligned to a group-wise space before any voxel-wise statistical analysis is carried out; in addition to this, it is desirable to restrict the analysis only to voxels belonging to white matter fiber bundles. All this was achieved by means of the Tract-Based Spatial Statistics (TBSS) algorithm implemented in FSL [30]. TBSS performs the following steps:

- Identify a common registration target (it can be either a mean FA template provided with the software or the most representative subject of the cohort) and apply nonlinear registration to align all subjects FA maps to the selected target. The chosen target was the FMRIB58_FA standard-space FA template, generated by averaging 58 FA images from diffusion MRI data, in MNI152 space.

- After the nonlinear registration, the entire aligned dataset undergoes an affine transformation to bring it into $1 \times 1 \times 1$ mm^3 MNI152 space. Then, a mean FA image is created, averaging all the FA maps in the dataset, and the result is used to generate a mean FA skeleton of white matter fibre tracts common to all subjects. The mean skeleton is thresholded to exclude voxels belonging to gray matter or cerebrospinal fluid, as well as voxels from the outermost part of the cortex, which are zones of greater inter-subject variability. Fig. 2 shows an example of FA map (2a) and MD map (2b), and the FA skeleton mask overlapped onto the mean FA map (2c).



(a) Example of FA map

(b) Example of MD map

(c) Mean FA skeleton

Fig. 2: From left to right: (a) a fractional anisotropy (FA) map and (b) a mean diffusivity (MD) map. For all subsequent analyses both maps are projected onto the mean FA skeleton (c).

- Finally, all subjects FA images are projected onto the mean FA skeleton, achieving an alignment between subjects in the direction orthogonal to the fibre bundle orientation.

TBSS was performed also on MD maps. After applying TBSS, each subject's map comprised about 7×10^6 nonzero voxels.

D. Feature selection

As a result of TBSS, the skeleton of main white matter fibre tracts was extracted from each subject, together with the corresponding values of FA and MD at each voxel in the skeleton. Approximately $120'000$ voxels for each subject map were projected onto the skeleton.

The following stage aimed at assessing which voxels are most significant for the purpose of discriminating HC from AD and MCI. It is important to note that it is not possible to rely on any assumption about the distribution of the test statistic under the null hypothesis; this implies that any statistical test has to be non-parametric. Wilcoxon rank sum test and the ReliefF algorithm were used both within a non-nested and nested approach. A Wilcoxon test compares the medians of the groups of data to determine if the samples come from the same population, and returns a p-value for the null hypothesis that samples are drawn from the same population [31], [32]. Then voxels are ranked selected by thresholding on p-values. The basic principle of ReliefF [33], [34] is to estimate features according to how well their values distinguish among data instances close to each other. Features are then ranked and sorted in order of decreasing importance.

For each classification task, fifteen reduced datasets were created by selecting an increasing number of most discriminating voxels, depending on the feature selection's output: 50, 100, 150, 200, 250, 300, 350, 400, 450, 500, 600, 750, $1'000$, $2'000$ and $3'000$ voxels.

E. Classification

In the present work, the learning and classification phase was accomplished by Random Forests. They constitute an ensemble learning method for classification and are known for producing highly accurate classifiers and for running efficiently on large datasets [35]. Random Forests operate by building a multitude of decision trees at training time and outputting the class that is the mode of the classes predicted by the individual trees at evaluation time. The training algorithm for Random Forests applies the general technique of *bootstrap aggregating*, or *bagging*, to tree learners. Given a training set $X = x_1, ..., x_n$, with classes $Y = y_1, ..., y_n$, the algorithm repeatedly (B times) selects a random sample with replacement of the training set and fits trees to these samples. More precisely, for $b = 1, ..., B$:

- n training examples are sampled with replacement from X, obtaining X_b.
- A subset of features is randomly chosen. Typically, for classification problems with m features, $\sqrt{m}$ features are chosen. The reason for doing this is to reduce the high correlation of the trees obtained in an ordinary bagging.

- A decision tree is trained on X_b.

It is worth noting that B (i.e., the number of samples/trees) is a free parameter. Since a few hundreds of samples represent the typical size of the forest, in this study a value equal to 300 for B was chosen. After training, predictions for unseen samples are made by taking the majority vote of all the predictions obtained by each individual tree. To perform the classification tasks, the implementation of Random Forests in MATLAB was used.

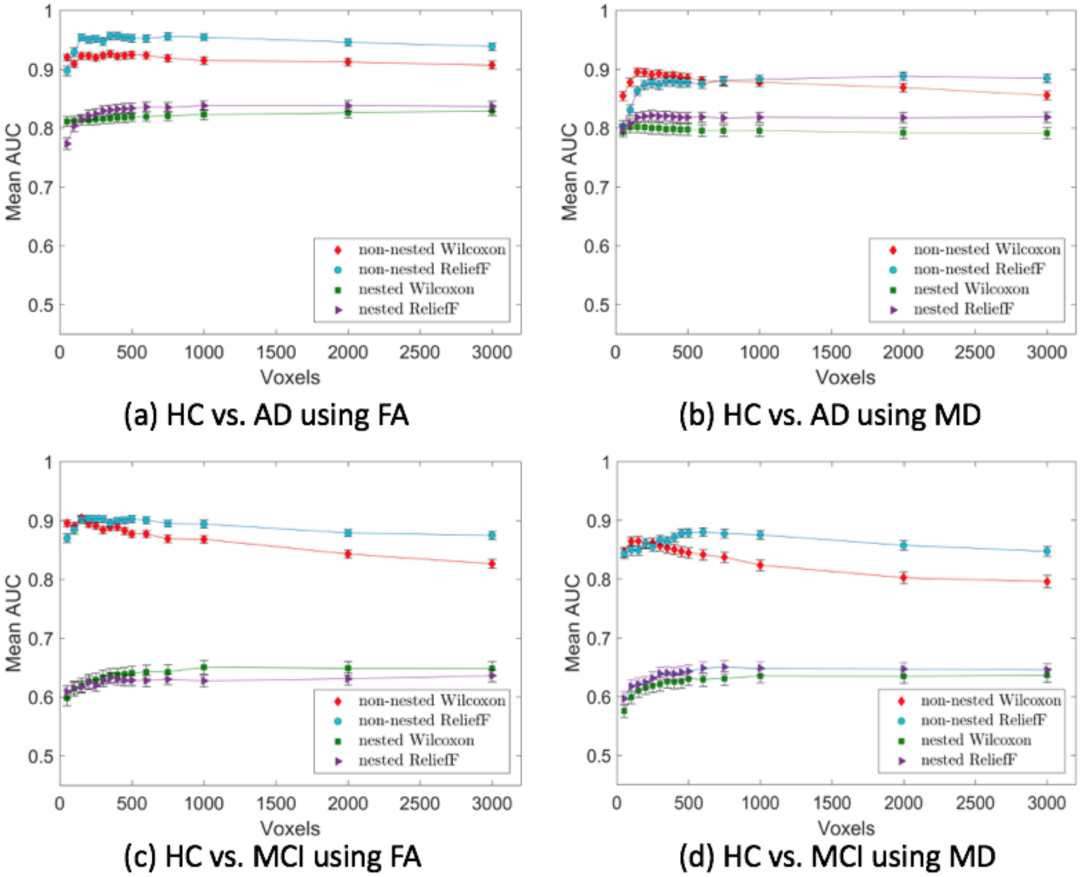
To determine the classification performance of the Random Forests classifier, a 100 times repeated 5-fold cross-validation for each reduced dataset was adopted. More precisely, every subject was shuffled into one of five folds from which one fold was selected as the test set, while the remaining folds form the training set. The subjects were stratified by diagnosis, such that each fold contained the same number of subjects from each diagnostic group. The classification process was repeated until each of the five folds was used as test set once. Finally, the full cross-validation procedure was repeated 100 times, using different permutations, to shuffle the subjects into the folds for a more general approximation of the performance.

It is worth noting that the non-nested approach employed a feature selection on the entire dataset before the dataset was split (Fig. 1a). Conversely, in the nested approach (Fig. 1b), for each cross-validation round, the dataset was split into a training and test set, then the feature selection was applied on the training set blind to the test set. As measures of performance, the widely used accuracy and AUC were calculated.

IV. Results

A. The feature selection bias effect

A primary question about the effects of excluding the feature selection from cross-validation procedures is whether or not the induced FSB is affected by the different kind of information employed, specifically FA and MD. Another question concerns the size of this effect. Besides, we also investigated whether or not the FSB was associated with the diagnosis, thus we separately studied the binary classification of HC/AD and HC/MCI. Finally, we included in our investigation two different feature selection techniques to assess whether the FSB effect could in some way depend on the methodology adopted to select the features. Mean AUCs for the classification involving both FA and MD measurements are plotted in Fig. 3 with both feature selection techniques.



(a) HC vs. AD using FA

(b) HC vs. AD using MD

(c) HC vs. MCI using FA

(d) HC vs. MCI using MD

Fig. 3: Mean AUCs obtained varying the number of voxels.

It can be observed that switching from non-nested to nested feature selection, for the classification between HC and AD, accuracy considerably decreases from a maximum mean value of 0.87 to a maximum value of 0.75, while the best AUC drops from 0.96 to 0.84. It is worth noting that the best performance is obtained using ReliefF, but for both feature selection techniques a significant drop in performance is consistently seen. The performance decrease switching from non-nested to nested approach is more evident for the classification between HC and MCI: the best classification performance changes from 0.81 to 0.59 concerning accuracy, and from 0.90 to 0.65 concerning AUC.

The same procedure was applied using MD. It is worth noting that moving from non-nested to nested feature selection, for the classification between HC and AD, best mean accuracy and AUC decrease respectively from 0.83 to 0.76 and from 0.90 to 0.82. For the discrimination HC/MCI the best accuracy falls from 0.79 to 0.60, while AUC decreases from 0.88 to 0.65. Again in this case, ReliefF performed better and the same performance deterioration detected for FA is clearly recognizable.

For each classification task and for each feature selection technique, the best performances in terms of mean accuracy and mean AUC are summarized in Table I.

TABLE I: The first column refers to the classification task. Best average performances in terms of accuracy (Acc) and Area Under the Curve (AUC) obtained in cross-validation with non-nested and nested feature selection are respectively reported in the second and third column; values are affected by a standard error of the mean approximately equal to 0.01 and a standard deviation approximately equal to 0.10. Non-nested feature selection always yields higher performances.

Classification	**Non-nested**	**Nested**
HC/AD with FA	Acc = 0.87	Acc = 0.75
	AUC = 0.96	AUC = 0.84
HC/MCI with FA	Acc = 0.81	Acc = 0.59
	AUC = 0.9	AUC = 0.65
HC/AD with MD	Acc = 0.83	Acc = 0.76
	AUC = 0.9	AUC = 0.82
HC/MCI with MD	Acc = 0.79	Acc = 0.6
	AUC = 0.88	AUC = 0.65

The Boxplot in Fig. 4 shows the distributions of the differences between the AUC values obtained in non-nested and nested best cases. It can be noticed that the FSB effect occurs regardless of the diffusion index (FA or MD) used for the classification and that this effect is more pronounced in the HC/MCI classification task.

A Wilcoxon rank sum test was performed to assess differences between the performance distributions with the nested and non-nested approach in a non-parametric fashion. Statistically significant differences ($p < 0.01$) were found between the median best performance obtained in the two cases (nested and non-nested) for all classification tasks and for both FA and MD. However, it must be noted that, for a given diffusion index (FA or MD), classification task (HC/AD or HC/MCI) and approach (nested or non-nested), the 100 measured performance metrics are not independent samples: all the 100 repetitions make use of the same images, and within each repetition there is substantial overlap among the training folds used for the cross-validation. It has been shown that, in cases like the present one, no unbiased estimator exists for the variance of the k-fold cross-validation [36]. The dependence of the samples and the impossibility to get an unbiased estimation of the variance violate the main assumption behind the use of standard parametric and non-parametric hypothesis tests. Therefore, we acknowledge the violation of the main assumption of hypothesis testing, and we warn the reader to use caution when interpreting the reported p-values.

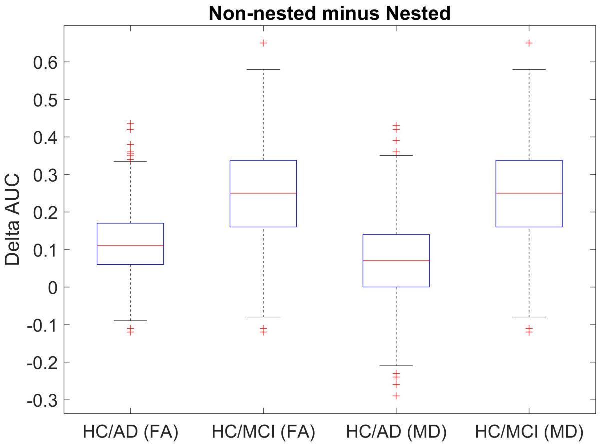
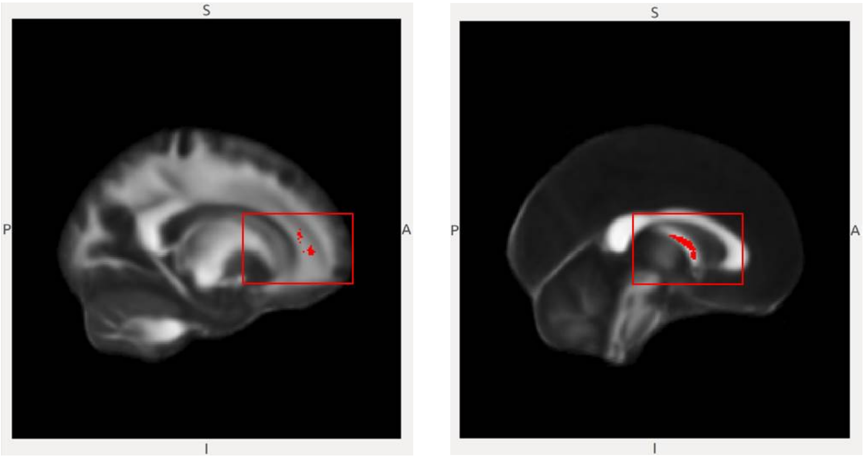


Fig. 4: Distribution of the differences between the AUCs obtained in non-nested and nested best performances shows a consistent increment.

B. *DTI measurements: evaluation on an independent test set*

It is worth noting that the information coming from the voxel selection can be used to identify the most disease-related brain regions concerning the fiber integrity. Therefore, in the present study, it was also investigated whether the voxels selected during the feature selection were localized in specific regions of interest of the brain.

For each classification task (HC/AD and HC/MCI) and for each feature selection technique (Wilcoxon and ReliefF), we considered the $1'000$ most discriminative voxels selected by the averaged nested feature-selection. They are “averaged” in the sense that they are the voxels that were more frequently selected throughout all the 500 rounds of the entire nested cross-validation procedure. Two selected clusters of FA voxels are shown as an example in Fig. 5.



(a) (b)

Fig. 5: Clusters of voxels selected by ReliefF averaging all rounds of the nested feature selection (classification task HC/AD with FA): (a) voxels in the Anterior Corona Radiata (left); (b) voxels in the Fornix.

The position of the voxels derived from the average cross validation was then investigated. In order to carry out the disease-related-regions analysis, a combination of three atlases (HarvardOxford-Subcortical, JHU-ICBM-labels, JHU-ICBM-tracts) was used. More precisely, using the voxels selected from the FA maps, the comparison of HC and AD reveals differences predominantly in the Anterior Corona Radiata (bilateral but more widespread in the left hemisphere) but also in the Superior Longitudinal Fasciculus (more widespread in the left hemisphere), Fornix, Cingulum (Hippocampus), Forceps Major and Minor, Inferior Fronto Occipital Fasciculus (right), Cortospinal Tract, Anterior Thalamic Radiation, Uncinate Fasciculus (right, only with Wilcoxon), Superior Corona Radiata and External Capsule (only with ReliefF). In the comparison between HC and MCI the FA changes are predominantly located in Forceps Minor, Superior Longitudinal Fasciculus, External Capsule (left) and, to a minor extent, in Inferior Fronto Occipital Fasciculus, Anterior Thalamic Radiation, Inferior Longitudinal Fasciculus, Cortical Spinal Tract, Fornix, Forceps Minor, Anterior Limb of Internal Capsule, Left Cerebral Cortex.

Concerning the voxels selected from the MD maps, comparing HC and AD, the predominant changes are localized in Fornix, Superior Longitudinal Fasciculus (more widespread in the left hemisphere in the case of Wilcoxon), Anterior Thalamic Radiation, Splenium and Body of Corpus Callosum, Inferior Longitudinal Fasciculus, Anterior Corona Radiata, Superior Corona Radiata (left). In the case of HC versus MCI, the MD differences are predominantly in Anterior Thalamic Radiation, Inferior Fronto Occipital Fasciculus (right), Forceps Major, Superior Longitudinal Fasciculus, Posterior Thalamic Radiation (right), Inferior Longitudinal Fasciculus, Fornix, Forceps Minor.

The effectiveness of the voxels selected by the nested cross-validation in discriminating the diagnostic groups was then evaluated on a second independent set of images from the ADNI database, consisting of new scans of 40 HC, 40 MCI and 39 AD. We considered the classification tasks HC/AD and HC/MCI with FA and MD and adopted the classification tool obtained at the end of the training phase. In particular, we considered only those models constructed on the reduced sets of voxels corresponding to the best classification performance and by fixing the feature selection technique adopted, i.e. ReliefF.

In order to evaluate the classification performances on the new data set, we calculated the mean scores, indicating the average predicted class posterior probabilities obtained by all models; then we calculated accuracy and AUC accordingly. The results obtained are reported in the third column of Table II. It can be noticed that they fall within one standard deviation of the corresponding mean value (second column).

V. Discussion and Conclusion

In this study we show that: (i) the use of non-nested feature selection techniques leads to overoptimistic classification performance; (ii) the FSB is manifest both for FA and MD, thus it does not depend on the features adopted; (iii) the FSB effect is more evident for the HC/MCI classification tasks.

TABLE II: Comparison between best average performances, both in terms of accuracy (Acc) and Area Under the Curve (AUC), on the training sample with nested feature selection and on the independent test sample. Independent test results (third column) are in good agreement with those obtained on the training set (training performances in the second column are affected by a standard deviation approximately equal to 0.10).

Classification	**Nested**	**Test (nested)**
HC/AD with FA	Acc = 0.75	Acc = 0.80
	AUC = 0.84	AUC = 0.91
HC/MCI with FA	Acc = 0.59	Acc = 0.56
	AUC = 0.65	AUC = 0.58
HC/AD with MD	Acc = 0.76	Acc = 0.73
	AUC = 0.82	AUC = 0.86
HC/MCI with MD	Acc = 0.6	Acc = 0.54
	AUC = 0.65	AUC = 0.60

The results obtained show that the voxel-based approach adopted in this study, without the bias introduced by the a priori feature selection, does not improve the classification performance obtained with other methodological procedures, except for the AUC achieved in the discrimination of HC vs. AD using FA. For the latter, the best accuracy is higher than the accuracy achieved by Mesrob et al. [12] and slightly lower than the value obtained by Schouten et al. [15]. Conversely, the AUC achieved is slightly higher than the one obtained by Schouten et al. [15]. For the classification HC/MCI it can be noticed that the accuracy and the AUC achieved with nested feature selection is lower than the one obtained in Cui et al. [16]; similarly, for the same classification task, the outcome is lower than the value obtained by Dyrba et al. [17].

If such detrimental effects on performance were somehow expected, it is worth noting that, as far as we know, no other study has measured this effect in the field of machine learning techniques applied to diffusion tensor imaging for AD. Furthermore, our findings regarding the significant regions for AD are consistent with several studies involving DTI, also when using other datasets than ADNI/ICBM, thus reassuring about the informative content of the voxel-based approach from the clinical point of view. Therefore the presence of the FSB in some studies using this approach is not detrimental to the anatomical and biological plausibility of the findings. In general, the existing literature provides evidence about the vulnerability of Fornix, Corpus Callosum and Cingulum to the early disease process involved in AD [37]. In particular, the white matter changes we found in the Fornix in all classification tasks (to a minor extent in the discrimination between HC and MCI using FA) have been reported in [38] and [39]. Indeed, FA reduction in the Fornix has been identified in the majority of whole-brain-TBSS studies applied to AD. Similarly, the predominant differences we observed in Cingulum, in the classification HC/AD using FA, are confirmed by looking, for example, at [40] and [41]. Additionally, the changes we observed in the Splenium of Corpus Callosum, when classifying HC vs. AD using MD, have been reported in [42] and [40]. The most consistent results with our findings are those reported in [43], where significant changes have also been found in Uncinate Fasciculus, Inferior Longitudinal Fasciculus, Superior Longitudinal Fasciculus and Forceps Major, and in [44], which identified changes in Anterior Corona Radiata, Inferior Fronto Occipital Fasciculus and Forceps Minor. Finally, we remark that [44] also confirms the predominance of differences in the left hemisphere we found in our analysis.

Acknowledgments

M. A. Scelsi acknowledges financial support by the EPSRC-funded UCL Centre for Doctoral Training in Medical Imaging (EP/L016478/1).

Data used in the preparation of this article was obtained from the the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI was launched in 2003 by the National Institute on Aging (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the Food and Drug Administration (FDA), private pharmaceutical companies and non-profit organizations, as a 60 million, 5-year public-private partnership. The primary goal of ADNI is to test whether serial MRI, PET, other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of MCI and early AD. Determination of sensitive and specific markers of very early AD progression is intended to aid researchers and clinicians to develop new treatments and monitor their effectiveness, as well as lessen the time and cost of clinical trials.

The Principal Investigator of this initiative is Michael W. Weiner, MD, VA Medical Center and the University of California, San Francisco. ADNI is the result of efforts of many co-investigators from a broad range of academic institutions and private corporations, and subjects have been recruited from over 50 sites across the U.S. and Canada. The initial goal of ADNI was to recruit 800 subjects; however, ADNI has been followed by ADNI-GO and ADNI-2. To date these three protocols have recruited over 1500 adults, ages 55 to 90, to participate in the research, consisting of cognitively normal older individuals, individuals with early or late MCI, and those with early AD. The follow up duration of each group is specified in the protocols for ADNI-1, ADNI-2, and ADNI-GO. Subjects originally recruited for ADNI-1 and ADNI-GO had the option to be followed in ADNI-2. For up-to-date information, see www.adni-info.org. Data collection and sharing for this project was funded by the ADNI (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: Alzheimer's Association; Alzheimer's Drug Discovery Foundation; BioClinica, Inc.; Biogen Idec Inc.; Bristol-Myers Squibb Company; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; GE Healthcare; Innogenetics, N.V.; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Medpace, Inc.; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Synarc Inc.; and Takeda Pharmaceutical Company. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

All authors disclose any actual or potential conflicts of interest, including any financial, personal, or other relationships with other people or organizations that could inappropriately influence their work. All experiments were performed with the informed consent of each participant or caregiver in line with the Code of Ethics of the World Medical Association (Declaration of Helsinki). Local institutional ethics committees approved the study.

REFERENCES

[1] C. Reitz and R. Mayeux, "Alzheimer disease: epidemiology, diagnostic criteria, risk factors and biomarkers," *Biochemical pharmacology*, vol. 88, no. 4, pp. 640–651, 2014.

[2] G. I. Allen, N. Amoroso, C. Anghel, V. Balagurusamy, C. J. Bare, D. Beaton, R. Bellotti, D. A. Bennett, K. L. Boehme, P. C. Boutros *et al.*, "Crowdsourced estimation of cognitive decline and resilience in alzheimer's disease," *Alzheimer's & Dementia*, vol. 12, no. 6, pp. 645–653, 2016.

[3] C. Jongkreangkrai, Y. Vichianin, C. Tocharoenchai, H. Arimura, A. D. N. Initiative *et al.*, "Computer-aided classification of alzheimer's disease based on support vector machine with combination of cerebral image features in mri," in *Journal of Physics: Conference Series*, vol. 694, no. 1. IOP Publishing, 2016, p. 012036.

[4] S. Tangaro, N. Amoroso, M. Boccardi, S. Bruno, A. Chincarini, G. Ferraro, G. Frisoni, R. Maglietta, A. Redolfi, L. Rei *et al.*, "Automated voxel-by-voxel tissue classification for hippocampal segmentation: methods and validation," *Physica Medica*, vol. 30, no. 8, pp. 878–887, 2014.

[5] N. Amoroso, R. Errico, S. Bruno, A. Chincarini, E. Garuccio, F. Sensi, S. Tangaro, A. Tateo, R. Bellotti, A. D. N. Initiative *et al.*, "Hippocampal unified multi-atlas network (HUMAN): protocol and scale validation of a novel segmentation tool," *Physics in medicine and biology*, vol. 60, no. 22, p. 8851, 2015.

[6] S. E. Rose, F. Chen, J. B. Chalk, F. O. Zelaya, W. E. Strugnell, M. Benson, J. Semple, and D. M. Doddrell, "Loss of connectivity in Alzheimer's disease: an evaluation of white matter tract integrity with colour coded MR diffusion tensor imaging," *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 69, no. 4, pp. 528–530, 2000.

[7] D. Head, R. L. Buckner, J. S. Shimony, L. E. Williams, E. Akbudak, T. E. Conturo, M. McAvoy, J. C. Morris, and A. Z. Snyder, "Differential vulnerability of anterior white matter in nondemented aging with minimal acceleration in dementia of the Alzheimer type: evidence from diffusion tensor imaging," *Cerebral Cortex*, vol. 14, no. 4, pp. 410–423, 2004.

[8] T. Wang, F. Shi, Y. Jin, P.-T. Yap, C.-Y. Wee, J. Zhang, C. Yang, X. Li, S. Xiao, and D. Shen, "Multilevel deficiency of white matter connectivity networks in Alzheimers disease: a diffusion MRI study with DTI and HARDI models," *Neural plasticity*, vol. 2016, 2016.

[9] P. J. Basser, J. Mattiello, and D. LeBihan, "MR diffusion tensor spectroscopy and imaging," *Biophysical journal*, vol. 66, no. 1, p. 259, 1994.

[10] N. Huang-Jing, Z. Lu-Ping, Z. Peng, H. Xiao-Lin, L. Hong-Xing, and N. Xin-Bao, "Multifractal analysis of white matter structural changes on 3D magnetic resonance imaging between normal aging and early Alzheimer's disease," *Chinese Physics B*, vol. 24, no. 7, p. 070502, 2015.

[11] D. Le Bihan, J. F. Mangin, C. Poupon, C. A. Clark, S. Pappata, N. Molko, and H. Chabriat, "Diffusion tensor imaging: concepts and applications," *Journal of Magnetic Resonance Imaging*, vol. 13, no. 4, pp. 534–546, 2001.

[12] L. Mesrob, M. Sarazin, V. Hahn-Barma, L. C. D. Souza, B. Dubois, P. Gallinari, and S. Kinkingnhun, "DTI and structural MRI classification in Alzheimer's disease," *Adv. Mol. Imaging*, vol. 2, pp. 12–20, 2012.

[13] M. Dyrba, M. Grothe, T. Kirste, and S. J. Teipel, "Multimodal analysis of functional and structural disconnection in Alzheimer's disease using multiple kernel SVM," *Human brain mapping*, vol. 36, no. 6, pp. 2118–2131, 2015.

[14] N. Amoroso, A. Monaco, and S. Tangaro, "Topological measurements of DWI tractography for the Alzheimer's disease detection," *Computational and Mathematical Methods in Medicine*, p. in press, 2016.

[15] T. M. Schouten, M. Koini, F. de Vos, S. Seiler, J. van der Grond, A. Lechner, A. Hafkemeijer, C. Möller, R. Schmidt, M. de Rooij, and S. A. R. B. Rombouts, "Combining anatomical, diffusion, and resting state functional magnetic resonance imaging for individual classification of mild and moderate Alzheimer's disease," *NeuroImage:clinical*, 2016.

[16] Y. Cui, W. Wen, D. M. Lipnicki, M. F. Beg, J. S. Jin, S. Luo, W. Zhu, N. A. Kochan, S. Reppermund, L. Zhuang *et al.*, "Automated detection of amnestic mild cognitive impairment in community-dwelling elderly adults: a combined spatial atrophy and white matter alteration approach," *Neuroimage*, vol. 59, no. 2, pp. 1209–1217, 2012.

[17] M. Dyrba, F. Barkhof, A. Fellgiebel, M. Filippi, L. Hausner, K. Hauenstein, T. Kirste, and S. J. Teipel, "Predicting prodromal Alzheimer's disease in subjects with mild cognitive impairment using machine learning classification of multimodal multicenter diffusion-tensor and magnetic resonance imaging data," *Journal of Neuroimaging*, vol. 25, no. 5, pp. 738–747, 2015.

[18] S. Haller, D. Nguyen, C. Rodriguez, J. Emch, G. Gold, A. Bartsch, K. O. Lovblad, and P. Giannakopoulos, "Individual prediction of cognitive decline in mild cognitive impairment using support vector machine-based analysis of diffusion tensor imaging data," *Journal of Alzheimer's Disease*, vol. 22, no. 1, pp. 315–327, 2010.

[19] L. O'Dwyer, F. Lamberton, A. L. W. Bokde, M. Ewers, Y. O. Faluyi, C. Tanner, B. Mazoyer, D. O'Neill, M. Bartley, D. R. Collins, T. Coughlan, D. Prvulovic, and H. Hampel, "Using support vector machines with multiple indices of diffusion for automated classification of mild cognitive impairment," *PLoS ONE*, vol. 7, no. 2, 2012.

[20] N. Kriegeskorte, W. K. Simmons, P. S. Bellgowan, and C. I. Baker, "Circular analysis in systems neuroscience: the dangers of double dipping," *Nature neuroscience*, vol. 12, no. 5, pp. 535–540, 2009.

[21] S. K. Singhi and H. Liu, "Feature subset selection bias for classification learning," in *Proceedings of the 23rd international conference on Machine learning*. ACM, 2006, pp. 849–856.

[22] F. Pereira, T. Mitchell, and M. Botvinick, "Machine learning classifiers and fMRI: a tutorial overview," *Neuroimage*, vol. 45, no. 1, pp. S199–S209, 2009.
[23] E. Olivetti, A. Mognon, S. Greiner, and P. Avesani, "Brain decoding: biases in error estimation," in *Brain Decoding: Pattern Recognition Challenges in Neuroimaging (WBD), 2010 First Workshop on*. IEEE, 2010, pp. 40–43.
[24] S. F. Eskildsen, P. Coupé, D. García-Lorenzo, V. Fonov, J. C. Pruessner, D. L. Collins, A. D. N. Initiative *et al.*, "Prediction of Alzheimer's disease in subjects with mild cognitive impairment from the ADNI cohort using patterns of cortical thinning," *Neuroimage*, vol. 65, pp. 511–521, 2013.
[25] D. Salas-Gonzalez, J. Górriz, J. Ramírez, M. López, I. Alvarez, F. Segovia, R. Chaves, and C. Puntonet, "Computer-aided diagnosis of alzheimer's disease using support vector machines and classification trees," *Physics in Medicine and Biology*, vol. 55, no. 10, p. 2807, 2010.
[26] A. Lebedev, E. Westman, G. Van Westen, M. Kramberger, A. Lundervold, D. Aarsland, H. Soininen, I. Kłoszewska, P. Mecocci, M. Tsolaki *et al.*, "Random forest ensembles for detection and prediction of alzheimer's disease with a good between-cohort robustness," *NeuroImage: Clinical*, vol. 6, pp. 115–125, 2014.
[27] M. Jenkinson, C. F. Beckmann, T. E. Behrens, M. W. Woolrich, and S. M. Smith, "FSL," *NeuroImage*, vol. 62, no. 2, pp. 782–790, 2012.
[28] A. L. Alexander, S. A. Hurley, A. A. Samsonov, N. Adluru, A. P. Hosseinbor, P. Mossahebi, D. P. Tromp, E. Zakszewski, and A. S. Field, "Characterization of cerebral white matter properties using quantitative magnetic resonance imaging stains," *Brain connectivity*, vol. 1, no. 6, pp. 423–446, 2011.
[29] H. M. Feldman, J. D. Yeatman, E. S. Lee, L. H. Barde, and S. Gaman-Bean, "Diffusion tensor imaging: a review for pediatric researchers and clinicians," *Journal of developmental and behavioral pediatrics: JDBP*, vol. 31, no. 4, p. 346, 2010.
[30] S. M. Smith, M. Jenkinson, H. Johansen-Berg, D. Rueckert, T. E. Nichols, C. E. Mackay, K. E. Watkins, O. Ciccarelli, M. Z. Cader, P. M. Matthews, and T. E. Behrens, "Tract-based spatial statistics: Voxelwise analysis of multi-subject diffusion data," *NeuroImage*, vol. 31, no. 4, pp. 1487–1505, 2006.
[31] M. Hollander, D. A. Wolfe, and E. Chicken, *Nonparametric statistical methods*. John Wiley & Sons, 2013.
[32] E. Whitley and J. Ball, "Statistics review 6: Nonparametric methods," *Critical Care*, vol. 6, no. 6, p. 1, 2002.
[33] K. Kira and L. A. Rendell, "The feature selection problem: Traditional methods and a new algorithm," in *AAAI*, vol. 2, 1992, pp. 129–134.
[34] I. Kononenko, E. Simec, and M. Robnik-Sikonja, "Overcoming the myopia of inductive learning algorithms with ReliefF," *Applied Intelligence*, vol. 7, pp. 39–55, 1997.
[35] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.
[36] Y. Bengio and Y. Grandvalet, "No unbiased estimator of the variance of k-fold cross-validation," *Journal of Machine Learning Research*, vol. 5, no. Sep, pp. 1089–1105, 2004.
[37] J. Acosta-Cabronero and P. J. Nestor, "Diffusion tensor imaging in Alzheimer's disease: insights into the limbic-diencephalic network and methodological considerations," *Frontiers in aging neuroscience*, vol. 6, 2014.
[38] K. Oishi and C. G. Lyketsos, "Alzheimer's disease and the fornix," *Frontiers in aging neuroscience*, vol. 6, 2014.
[39] M. A. Nowrangi and P. B. Rosenberg, "The fornix in mild cognitive impairment and Alzheimer's disease," *Frontiers in aging neuroscience*, vol. 7, p. 1, 2015.
[40] S. J. Teipel, R. Stahl, O. Dietrich, S. O. Schoenberg, R. Perneczky, A. L. Bokde, M. F. Reiser, H.-J. Möller, and H. Hampel, "Multivariate network analysis of fiber tract integrity in Alzheimer's disease," *Neuroimage*, vol. 34, no. 3, pp. 985–995, 2007.
[41] F. Agosta, M. Pievani, S. Sala, C. Geroldi, S. Galluzzi, G. B. Frisoni, and M. Filippi, "White matter damage in Alzheimer disease and its relationship to gray matter atrophy," *Radiology*, vol. 258, no. 3, pp. 853–863, 2011.
[42] R. Stahl, O. Dietrich, S. J. Teipel, H. Hampel, M. F. Reiser, and S. O. Schoenberg, "White Matter Damage in Alzheimer Disease and Mild Cognitive Impairment: Assessment with Diffusion-Tensor MR Imaging and Parallel Imaging Techniques," *Radiology*, vol. 243, no. 2, pp. 483–492, 2007.
[43] N. H. Stricker, B. Schweinsburg, L. Delano-Wood, C. E. Wierenga, K. J. Bangen, K. Haaland, L. R. Frank, D. P. Salmon, and M. W. Bondi, "Decreased white matter integrity in late-myelinating fiber pathways in Alzheimer's disease supports retrogenesis," *Neuroimage*, vol. 45, no. 1, pp. 10–16, 2009.
[44] G. Sousa Alves, L. O'Dwyer, A. Jurcoane, V. Oertel-Knöche, C. Knöchel, D. Prvulovic, F. Sudo, C. E. Alves, L. Valente, D. Moreira, F. Fußer, T. Karakaya, J. Pantel, E. Engelhardt, and J. Laks, "Different patterns of white matter degeneration using multiple diffusion indices and volumetric data in Mild Cognitive Impairment and Alzheimer patients," *PLoS ONE*, vol. 7, no. 12, 2012.
[45] S. Tangaro, N. Amoroso, M. Antonacci, M. Boccardi, M. Bocchetta, A. Chincarini, D. Diacono, G. Donvito, R. Errico, G. Frisoni *et al.*, "MRI analysis for hippocampus segmentation on a distributed infrastructure," in *2016 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*. IEEE, 2016, pp. 1–6.
[46] S. Vicario, B. Balech, G. Donvito, P. Notarangelo, and G. Pesole, "The biovel project: Robust phylogenetic workflows running on the grid," *EMBnet. journal*, vol. 18, no. B, pp. pp–77, 2012.

Supplementary Data

All the elaborations and analyses presented in this work require huge computational resources, with preprocessing and diffusion tensor fitting time of about one hour per subject. The present study was carried out on the distributed computing infrastructure ReCaS-Bari computing farm[1]. This data center has been built by the ReCaS project, funded by the Italian Research Ministry of Education, University and Research to the University of Bari and INFN (National Institute for Nuclear Physics), whose goal is to empower preexisting computing infrastructures located in Catania, Cosenza, Napoli and Bari. In particular the data center offers 128 servers, 64 cores per server, for a total amount of $8'192$ new cores, reaching $12'000$ cores with the old computing farm. Each new server hosts 256GB RAM, 4GB RAM per core per server. Additionally, it offers about 3.5PB of disk space and 2.5PB of tape space.

To implement our analysis on distribute infrastructure we used LONI Pipeline, one of the most used workflow manager for medical image processing developed by the Laboratory of Neuro Imaging[2]. The LONI Pipeline (LP) is widely used by the scientific community since it has proved to be a convenient and powerful tool. In particular XML resource description facilitates the integration of disparate resources and provides a natural and comprehensive mechanism to support data provenance. It also enables the broad dissemination of resource metadata descriptions via web-services and the constructive utilization of multidisciplinary expertise by experts, novice users and trainees. We have developed a general approach to submit and monitor LP workflows on distributed infrastructures [45]. This framework is based on a meta-scheduler, the Job Submission Tool (JST) [46], that is able to submit jobs to different computing architectures, exposing to the end users only a simple Web Service interface based on the Representational State Transfer (REST) protocol.

[1] https://www.recas-bari.it/index.php/it/
[2] http://pipeline.loni.usc.edu/