

DTI measurements for Alzheimer's classification

Tommaso Maggipinto^{1,2}, Roberto Bellotti^{1,2},
Nicola Amoroso^{1,2,4}, Domenico Diacono², Giacinto Donvito²,
Eufemia Lella^{1,2}, Alfonso Monaco², Marzia Antonella
Scelsi³, Sabina Tangaro² and for the Alzheimer's Disease
Neuroimaging Initiative⁵

¹ Dipartimento Interateneo di Fisica 'M. Merlin', Università degli Studi di Bari
'A. Moro', Via Giovanni Amendola, 173, 70125 Bari, Italy

² Istituto Nazionale di Fisica Nucleare, Sezione di Bari, Via Orabona, 4, 70123, Bari,
Italy

³ Translational Imaging Group, Centre for Medical Image Computing,
University College London, Gower Street, London NW1 2HE, United Kingdom

E-mail: nicola.amoroso@ba.infn.it

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

⁴ Author to whom any correspondence should be addressed.

⁵ 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

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.

Keywords: Alzheimer's disease, DTI, random forests, feature selection

 Supplementary material for this article is available [online](#)

(Some figures may appear in colour only in the online journal)

1. 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 (Reitz and Mayeux 2014). Nevertheless, the investigation of novel biomarkers and strategies to predict and model its onset needs further investigation (Allen *et al* 2016). In particular, the investigation of biological markers aimed at diagnosing the disease promptly is crucial (Jongkreangkrai *et al* 2016). 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 (Tangaro *et al* 2014, Amoroso *et al* 2015, Jongkreangkrai *et al* 2016). 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 (Rose *et al* 2000, Head *et al* 2004, Wang *et al* 2016). 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 (Basser *et al* 1994, Huang-Jing *et al* 2015). Fractional anisotropy (FA) and mean diffusivity (MD) are among the invariants derived from the diffusion tensor that are closely related to white matter integrity (Le Bihan *et al* 2001). 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* (2012) adopted a support vector machine (SVM) classifier and a region of interest (ROI)-based approach; Dyrba *et al* (2015b) used a ROI-based approach and a multimodal SVM combining DTI indices with gray matter volume

derived from sMRI; Amoroso *et al* (2016) adopted topological measurements based on probabilistic tractography; Schouten *et al* (2016) used a ROI-based approach in combination with elastic net regression. For the classification HC/MCI, Cui *et al* (2012) used subcortical volumetric features extracted using a segmentation algorithm together with FA values obtained for white matter regions of interest. Dyrba *et al* (2015a) 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* (2010) and O'Dwyer *et al* (2012). In these works, a voxel-based approach is used considering as features the voxel intensities in the diffusion maps. However, as also remarked in O'Dwyer *et al* (2012), 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 (Singhi and Liu 2006, Kriegeskorte *et al* 2009). Although recommendations and best practices are available (Pereira *et al* 2009), 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 (Pereira *et al* 2009) or MEG (Olivetti *et al* 2010), and with data derived from structural T1-weighted MR imaging (cortical thickness) in Eskildsen *et al* (2013). 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 (Salas-Gonzalez *et al* 2010, Lebedev *et al* 2014), 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.

2. 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 ($b = 1000 \text{ s mm}^{-2}$); in addition to these, 5 images with negligible diffusion effects (b_0 images) were acquired as reference scans for subsequent analysis.

3. Methods

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

3.1. Image preprocessing

Diffusion-weighted images were preprocessed using the FMRIB Diffusion Toolbox, included in the FSL software (Jenkinson *et al* 2012). 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).

3.2. 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 $\lambda_1, \lambda_2, \lambda_3$ by Basser *et al* (1994) and Le Bihan *et al* (2001):

$$\text{FA} = \sqrt{\frac{1}{2} \frac{\sqrt{((\lambda_1 - \lambda_2)^2 + (\lambda_2 - \lambda_3)^2 + (\lambda_3 - \lambda_1)^2)}}{\sqrt{\lambda_1^2 + \lambda_2^2 + \lambda_3^2}}} \quad (1)$$

$$\text{MD} = \frac{\lambda_1 + \lambda_2 + \lambda_3}{3} \quad (2)$$

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 $\text{FA} = 0$, whereas a value of 1 indicates that diffusion is fully constrained along one direction. Water diffusion in a healthy axon or fiber bundle is highly anisotropic and

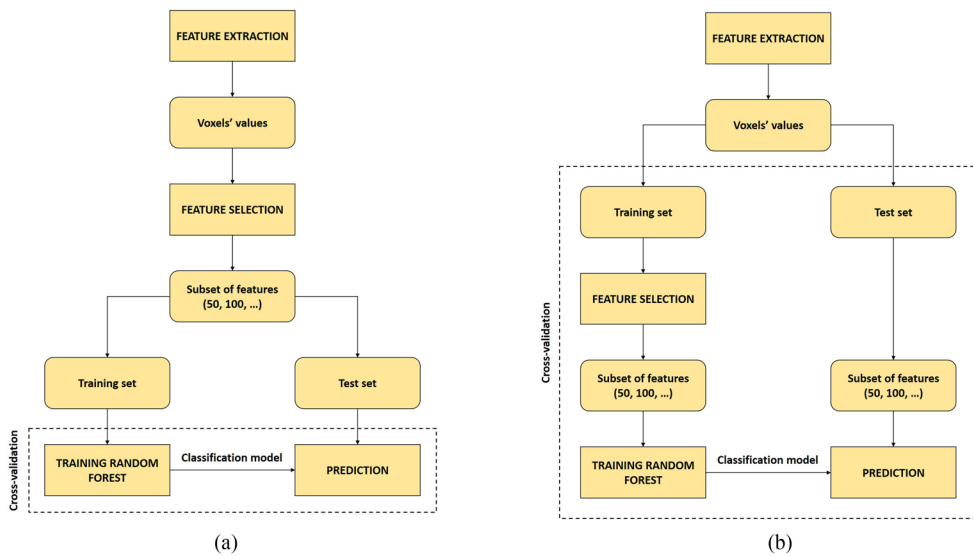


Figure 1. Flowcharts of the performed analyses: (a) non-nested feature selection and (b) nested feature selection. For readability, they only consider the steps following the feature extraction phase. (a) Non-nested approach. (b) Nested approach.

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 (Feldman *et al* 2010, Alexander *et al* 2011).

3.3. 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 (Smith *et al* 2006). 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 $1 \times 1 \times 1 \text{ mm}^3$ 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

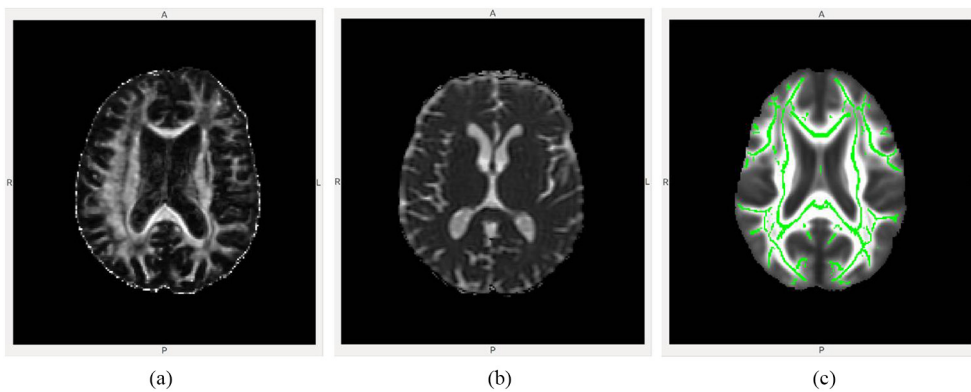


Figure 2. From left to right: (a) a fractional anisotropy (FA) map and (b) a mean diffusivity (MD) map. For all subsequent analyses both maps are projected onto the mean FA skeleton (c). (a) Example of FA map. (b) Example of MD map. (c) Mean FA skeleton.

from the outermost part of the cortex, which are zones of greater inter-subject variability. Figure 2 shows an example of FA map (figure 2(a)) and MD map (figure 2(b)), and the FA skeleton mask overlapped onto the mean FA map (figure 2(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.

3.4. 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 (Whitley and Ball 2002, Hollander *et al* 2013). Then voxels are ranked selected by thresholding on p -values. The basic principle of ReliefF (Kira and Rendell 1992, Kononenko *et al* 1997) 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, 1000, 2000 and 3000 voxels.

3.5. 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 (Breiman 2001). 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 $X = x_1, \dots, x_n$, with classes $Y = y_1, \dots, y_n$, the algorithm repeatedly (B times) selects a random sample with replacement of the training set and fits trees to these samples. More precisely, for $b = 1, \dots, B$:

- 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 (figure 1(a)). Conversely, in the nested approach (figure 1(b)), 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.

4. Results

4.1. 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 figure 3 with both feature selection techniques.

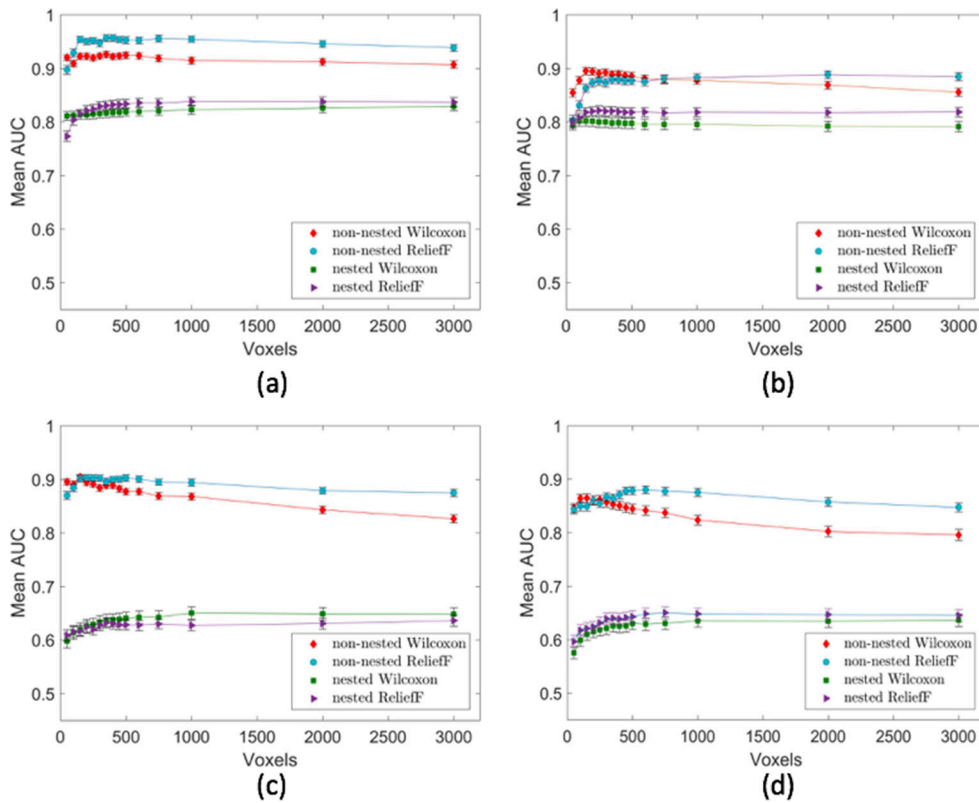


Figure 3. Mean AUCs obtained varying the number of voxels. (a) HC versus AD using FA. (b) HC versus AD using MD. (c) HC versus MCI using FA. (d) HC versus MCI using MD.

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 1.

The boxplot in figure 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.

Table 1. 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

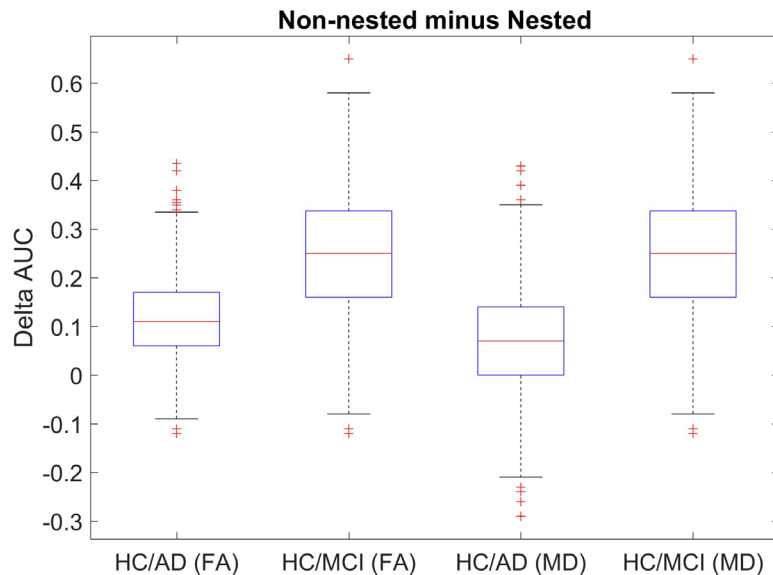


Figure 4. Distribution of the differences between the AUCs obtained in non-nested and nested best performances shows a consistent increment.

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 (Bengio and Grandvalet 2004). The dependence of the samples and the impossibility to get an unbiased estimation of the variance violate the

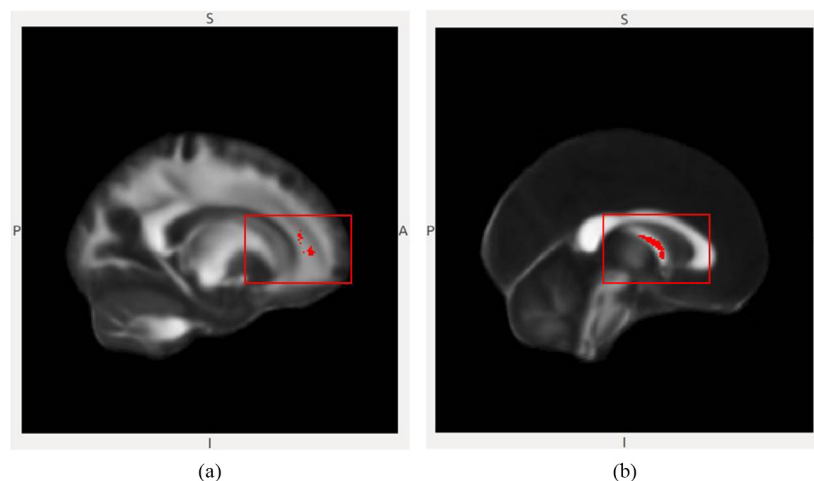


Figure 5. Clusters of voxels selected by ReliefF averaging all rounds of the nested feature selection (classification task HC/AD with FA): (a) voxels in the anterior corona radiata (left); (b) voxels in the fornix.

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.

4.2. 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 1000 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 figure 5.

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.

Table 2. 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

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 2. It can be noticed that they fall within one standard deviation of the corresponding mean value (second column).

5. 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.

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 versus AD using FA. For the latter, the best accuracy is higher than the accuracy achieved by Mesrob *et al* (2012) and slightly lower than the value obtained by Schouten *et al* (2016). Conversely, the AUC achieved is slightly higher than the one obtained

by Schouten *et al* (2016). 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* (2012); similarly, for the same classification task, the outcome is lower than the value obtained by Dyrba *et al* (2015a).

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 (Acosta-Cabronero and Nestor 2014). 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 Oishi and Lyketsos (2014) and Nowrangi and Rosenberg (2015). 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 Teipel *et al* (2007) and Agosta *et al* (2011). Additionally, the changes we observed in the Splenium of Corpus Callosum, when classifying HC versus AD using MD, have been reported in Stahl *et al* (2007) and Teipel *et al* (2007). The most consistent results with our findings are those reported in Stricker *et al* (2009), where significant changes have also been found in uncinate fasciculus, inferior longitudinal fasciculus, superior longitudinal fasciculus and forceps major, and in Sousa Alves *et al* (2012), which identified changes in anterior corona radiata, inferior fronto occipital fasciculus and forceps minor. Finally, we remark that Sousa Alves *et al* (2012) 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

- Acosta-Cabronero J and Nestor P J 2014 Diffusion tensor imaging in Alzheimer's disease: insights into the limbic-diencephalic network and methodological considerations *Frontiers Aging Neurosci.* **6**
- Agosta F, Pievani M, Sala S, Geroldi C, Galluzzi S, Frisoni G B and Filippi M 2011 White matter damage in Alzheimer disease and its relationship to gray matter atrophy *Radiology* **258** 853–63
- Alexander A L, Hurley S A, Samsonov A A, Adluru N, Hosseinbor A P, Mossahebi P, Tromp D P, Zakszewski E and Field A S 2011 Characterization of cerebral white matter properties using quantitative magnetic resonance imaging stains *Brain Connect.* **1** 423–46
- Allen G I et al 2016 Crowdsourced estimation of cognitive decline and resilience in Alzheimer's disease *Alzheimer's Dementia* **12** 645–53
- Amoroso N et al 2015 Hippocampal unified multi-atlas network (HUMAN): protocol and scale validation of a novel segmentation tool *Phys. Med. Biol.* **60** 8851
- Amoroso N, Monaco A and Tangaro S 2016 Topological measurements of DWI tractography for the Alzheimer's disease detection *Comput. Math. Methods Med.* (accepted)
- Basser P J, Mattiello J and LeBihan D 1994 MR diffusion tensor spectroscopy and imaging *Biophys. J.* **66** 259
- Bengio Y and Grandvalet Y 2004 No unbiased estimator of the variance of k-fold cross-validation *J. Mach. Learn. Res.* **5** 1089–105
- Breiman L 2001 Random forests *Mach. Learn.* **45** 5–32
- Cui Y et al 2012 Automated detection of amnesic mild cognitive impairment in community-dwelling elderly adults: a combined spatial atrophy and white matter alteration approach *Neuroimage* **59** 1209–17

- Dyrba M, Barkhof F, Fellgiebel A, Filippi M, Hausner L, Hauenstein K, Kirste T and Teipel S J 2015a 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 *J. Neuroimaging* **25** 738–47
- Dyrba M, Grothe M, Kirste T and Teipel S J 2015b Multimodal analysis of functional and structural disconnection in Alzheimer's disease using multiple kernel SVM *Hum. Brain Mapp.* **36** 2118–31
- Eskildsen S F et al 2013 Prediction of Alzheimer's disease in subjects with mild cognitive impairment from the ADNI cohort using patterns of cortical thinning *Neuroimage* **65** 511–21
- Feldman H M, Yeatman J D, Lee E S, Barde L H and Gaman-Bean S 2010 Diffusion tensor imaging: a review for pediatric researchers and clinicians *J. Dev. Behav. Pediatrics* **31** 346
- Haller S, Nguyen D, Rodriguez C, Emch J, Gold G, Bartsch A, Lovblad K O and Giannakopoulos P 2010 Individual prediction of cognitive decline in mild cognitive impairment using support vector machine-based analysis of diffusion tensor imaging data *J. Alzheimer's Disease* **22** 315–27
- Head D, Buckner R L, Shimony J S, Williams L E, Akbudak E, Conturo T E, McAvoy M, Morris J C and Snyder A Z 2004 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* **14** 410–23
- Hollander M, Wolfe D A and Chicken E 2013 *Nonparametric Statistical Methods* (New York: Wiley)
- Huang-Jing N, Lu-Ping Z, Peng Z, Xiao-Lin H, Hong-Xing L and Xin-Bao N 2015 Multifractal analysis of white matter structural changes on 3D magnetic resonance imaging between normal aging and early Alzheimer's disease *Chin. Phys. B* **24** 070502
- Jenkinson M, Beckmann C F, Behrens T E, Woolrich M W and Smith S M 2012 FSL *Neuroimage* **62** 782–90
- Jongkreangkrai C et al 2016 Computer-aided classification of Alzheimer's disease based on support vector machine with combination of cerebral image features in MRI *J. Phys.: Conf. Ser.* **694** 012036 (Bristol IOP)
- Kira K and Rendell L A 1992 The feature selection problem: traditional methods and a new algorithm *AAAI vol 2* pp 129–34
- Kononenko I, Simec E and Robnik-Sikonja M 1997 Overcoming the myopia of inductive learning algorithms with RELIEFF *Appl. Intell.* **7** 39–55
- Kriegeskorte N, Simmons W K, Bellgowan P S and Baker C I 2009 Circular analysis in systems neuroscience: the dangers of double dipping *Nat. Neurosci.* **12** 535–40
- Le Bihan D, Mangin J F, Poupon C, Clark C A, Pappata S, Molko N and Chabriat H 2001 Diffusion tensor imaging: concepts and applications *J. Magn. Reson. Imaging* **13** 534–46
- Lebedev A et al 2014 Random forest ensembles for detection and prediction of Alzheimer's disease with a good between-cohort robustness *Neuroimage: Clin.* **6** 115–25
- Mesrob L, Sarazin M, Hahn-Barma V, Souza L C D, Dubois B, Gallinari P and Kinkingnéhun S 2012 DTI and structural MRI classification in Alzheimer's disease *Adv. Mol. Imaging* **2** 12–20
- Nowrangi M A and Rosenberg P B 2015 The fornix in mild cognitive impairment and Alzheimer's disease *Frontiers Aging Neurosci.* **7** 1
- O'Dwyer L et al 2012 Using support vector machines with multiple indices of diffusion for automated classification of mild cognitive impairment *PLoS One* **7**
- Oishi K and Lyketos C G 2014 Alzheimer's disease and the fornix *Frontiers Aging Neurosci.* **6**
- Olivetti E, Mognon A, Greiner S and Avesani P 2010 Brain decoding: biases in error estimation 2010 *First Workshop on Brain Decoding: Pattern Recognition Challenges in Neuroimaging* pp 40–3
- Pereira F, Mitchell T and Botvinick M 2009 Machine learning classifiers and fMRI: a tutorial overview *Neuroimage* **45** S199–209
- Reitz C and Mayeux R 2014 Alzheimer disease: epidemiology, diagnostic criteria, risk factors and biomarkers *Biochem. Pharmacol.* **88** 640–51
- Rose S E, Chen F, Chalk J B, Zelaya F O, Strugnell W E, Benson M, Semple J and Doddrell D M 2000 Loss of connectivity in Alzheimer's disease: an evaluation of white matter tract integrity with colour coded MR diffusion tensor imaging *J. Neurol. Neurosurg. Psychiatry* **69** 528–30
- Salas-Gonzalez D, Górriz J, Ramírez J, López M, Alvarez I, Segovia F, Chaves R and Puntónet C 2010 Computer-aided diagnosis of Alzheimer's disease using support vector machines and classification trees *Phys. Med. Biol.* **55** 2807
- Schouten T M et al 2016 Combining anatomical, diffusion, and resting state functional magnetic resonance imaging for individual classification of mild and moderate Alzheimer's disease *Neuroimage: Clin.* **11** 46–51

- Singhi S K and Liu H 2006 Feature subset selection bias for classification learning *Proc. of the 23rd Int. Conf. on Machine Learning* pp 849–56
- Smith S M *et al* 2006 Tract-based spatial statistics: voxelwise analysis of multi-subject diffusion data *Neuroimage* **31** 1487–505
- Sousa Alves G *et al* 2012 Different patterns of white matter degeneration using multiple diffusion indices and volumetric data in mild cognitive impairment and Alzheimer patients *PLoS One* **7**
- Stahl R, Dietrich O, Teipel S J, Hampel H, Reiser M F and Schoenberg S O 2007 White matter damage in Alzheimer disease and mild cognitive impairment: assessment with diffusion-tensor MR imaging and parallel imaging techniques *Radiology* **243** 483–92
- Stricker N H, Schweinsburg B, Delano-Wood L, Wierenga C E, Bangen K J, Haaland K, Frank L R, Salmon D P and Bondi M W 2009 Decreased white matter integrity in late-myelinating fiber pathways in Alzheimer's disease supports retrogenesis *Neuroimage* **45** 10–6
- Tangaro S *et al* 2014 Automated voxel-by-voxel tissue classification for hippocampal segmentation: methods and validation *Phys. Med.* **30** 878–87
- Teipel S J, Stahl R, Dietrich O, Schoenberg S O, Perneczky R, Bokde A L, Reiser M F, Möller H J and Hampel H 2007 Multivariate network analysis of fiber tract integrity in Alzheimer's disease *Neuroimage* **34** 985–95
- Wang T, Shi F, Jin Y, Yap P T, Wee C Y, Zhang J, Yang C, Li X, Xiao S and Shen D 2016 Multilevel deficiency of white matter connectivity networks in Alzheimer's disease: a diffusion MRI study with DTI and HARDI models *Neural Plast.* **2016**
- Whitley E and Ball J 2002 Statistics review 6: nonparametric methods *Crit. Care* **6** 1