How Random is the Random Forest? Random Forest Algorithm on the Service of Structural Imaging Biomarkers for Alzheimer’s Disease: from Alzheimer’s disease neuroimaging initiative (ADNI) database

Dimitriadis S.I.¹, Liparas D.²

¹Neuroscience and Mental Health Research Institute, Cardiff University, Cardiff, UK; Cardiff University Brain Research Imaging Centre (CUBRIC), School of Psychology, Cardiff University, Cardiff, UK; MRC Centre for Neuropsychiatric Genetics and Genomics, Institute of Psychological Medicine and Clinical Neurosciences, Cardiff School of Medicine, Cardiff University, Cardiff, UK; Neuroinformatics Group, (CUBRIC), School of Psychology, Cardiff University, Cardiff, UK; School of Psychology, Cardiff University, Cardiff, UK; DimitriadisS@cardiff.ac.uk.

²High Performance Computing Center Stuttgart (HLRS), University of Stuttgart, Stuttgart, Germany; Department of Informatics, Aristotle University of Thessaloniki, Thessaloniki, Greece: liparas@hlrs.de.

Contact:
Email: DimitriadisS@cardiff.ac.uk, stidimitriadis@gmail.com; dimitrios.liparas@gmail.com

¹All the data used in preparation of this article were obtained from the Alzheimer’s Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf. The preprocessing of the T1-weighted Magnetic Resonance Images (MRI) was realized by the organizers of the competition where information can be found here: https://inclass.kaggle.com/c/mci-prediction
Abstract:

Neuroinformatics is a fascinating research field that applies computational models and analytical tools to high dimensional experimental neuroscience data for a better understanding of how the brain functions or dysfunctions in brain diseases. Neuroinformaticians work in the intersection of neuroscience and informatics supporting the integration of various sub-disciplines (behavioural neuroscience, genetics, cognitive psychology etc) working on brain research. Neuroinformaticians are the pathway of information exchange between informaticians and clinicians for a better understanding of the outcome of computational models and the clinical interpretation of the analysis.

Machine learning is one of the most significant computational developments in the last decade giving tools to neuroinformaticians and finally to radiologists and clinicians for an automatic and early diagnosis-prognosis of a brain disease. Random Forest (RF) algorithm has been successfully applied to high-dimensional neuroimaging data for feature reduction and also has been applied to classify the clinical label of a subject using single or multi-modal neuroimaging datasets. Our aim was to review the studies where RF was applied to correctly predict the Alzheimer’s disease (AD), the conversion from Mild Cognitive Impairment (MCI) and its robustness to overfitting, outliers and handling of non-linear data. Finally, we described our RF-based model that gave us the 1st position in an international challenge for automated prediction of MCI from MRI data.

Keywords: random forest, Alzheimer’s Disease, mild cognitive impairment, neuroimaging, classification, machine learning, biomarker, magnetic resonance imaging
1. Introduction

Machine learning techniques, including feature selection and classification, have been one of the most important development of computational science over the last years to satisfy the daily demands of clinicians for an accurate and automatic diagnosis and prognosis of various brain diseases and disorders (van Ginneken et al., 2011). Nowadays, the mental workload of radiologists has increased while the number of radiologists on national health systems (NHS) worldwide is still limited (Reiman, 2017). Simultaneously, the health care cost of imaging is rising very fast. We can clearly state that the inconsistency of the interpretation of the results among radiologists and the better performance of algorithms demand new approaches to handle neuroimaging data. Computer-aided diagnosis (CAD) may be the solution to speed up the diagnosis, increase the accuracy of the diagnosis, reduce the total cost of NHS and further improve any quantitative measurement related to the fast and accurate diagnosis and prognosis.

Alzheimer’s disease (AD) is a neurodegenerative disorder that mostly affects elderly individuals (Berchtold and Cotman, 1998). AD is the most common type of dementia and as the elderly populations grows, the number of AD patients will also rise. The prevalence of AD in Europe was 3.31% in men and 7.13% in women while the incidence was 7.02% per 1.000 person-years in men and 13.25% per person-years in women (Niu et al., 2016). In China, the burden of dementia seems to be increasing faster than in other countries and under the rules of international health community (Chan et al., 2013).

The characteristic features of AD is the decline of cognitive function, progressive loss of memory, language and also reasoning (Collie and Maruff, 2000). Mild cognitive impairment (MCI) is an intermediate stage between healthy aging and AD where the individual can handle its daily activities without any interference. The cognitive status of MCI remains constant for many years, while the incidence of progression from MCI to AD has been evaluated between 10%-15% per year (Palmqvist et al., 2012). Till now, there is no acceptable cure for AD but only several treatments that attempt to delay the decay of the progression of the disease. For that reason, it is extremely significant to define biomarkers that can detect accurately the MCI individuals that are at risk to convert to AD.

The diagnosis of AD over control healthy aging is based on a large set of potential features (variables and factors) like the genetic information, neuropsychological tests, demographics, brain imaging data and variables derived from cerebrospinal fluid (CSF). Especially for the risk of conversion from MCI to AD, the change of every individual variable and also the
alteration of a combination of a subset of those variables could weight their importance over that risk. Especially in neuroimaging, a large repertoire of technologies like magnetic resonance imaging (MRI), diffusion magnetic resonance imaging (dMRI), functional MRI (fMRI), diffusion tensor imaging (DTI) and positron emission tomography (PET) have been applied successfully for the study of MCI and AD (AcostaCabronero and Nestor, 2014). The choice of each neuroimaging modality could be altered, based on the severity of the disease and their sensitivity to detect alterations of the brain, both structurally and functionally. For example, fMRI and PET can detect metabolic abnormalities while DTI could investigate microstructural changes of the white matter and its properties (e.g. myelination) while d-MRI can give us a clear view of the myelination assessment.

The high dimensionality of all these features that are taken into consideration for the diagnosis of AD and for the progression from MCI to AD and also their complicated interactions makes the whole effort to select the best subset of them very difficult. CAD, in general, represents an automatic software system that supports the clinicians to manipulate in a fast and accurate way the large number of case studies. At first level, this CAD system should be taught by the clinicians in a supervised mode. Pattern Analysis and Machine Intelligence (PAMI) algorithms have been proven valuable for the classification of AD subjects versus healthy controls (HC) and also for the discrimination of stable MCI (sMCI) and progressive MCI (pMCI) that finally converted to AD (Trzepacz et al., 2014). Fig.1 illustrates the steps of the proposed CAD system for AD using MRI brain images. T1-MRI images were collected by the neuroimaging labs contributed to ADNI’s database. MRI images can be segmented according to anatomically oriented ROIs using an anatomical template. Then, morphological features can be extracted using various free academic softwares. Machine learning techniques can be applied afterward starting with feature selection and a classification approach on a training dataset with known labels (e.g. 0 for healthy control, 1 for AD). The performance can be evaluated by a radiologist/neurologist in a new sample.
The majority of machine learning neuroimaging studies relied on the Support Vector Machine (SVM), Linear Discriminant Analysis (LDA), or Naïve Bayes algorithms. In the last few years, ensemble algorithms proved to be an alternative pathway to single classifiers, based on a better performance than the latter, especially in the case where multi-modality features-variables were combined. Among all ensembles approaches, Random Forest (RF) (Breiman, 2001) produced the best classification accuracies in many scientific fields and in many neurological diseases apart from AD and only during the last few years researchers started paying attention to it (Sarica et al., 2017; Dimitriadis et al., 2017). In particular, RF demonstrated its advantage over other methodologies regarding their potentiality to manipulate non-linear variables, while it is a robust method to noise and can be easily tuned and processed in parallel (Caruana and Niculescu-Mizil, 2006). Additionally, RF introduces an initial feature selection step that can reduce the variable space by ranking the value of each feature.

Section 2 discusses in detail classification of neuroimaging data tailored to AD while section 3 is devoted generally to the latest review of Machine Learning Techniques in Neuroimaging Data. In section 4, we described all the necessary information for the International Challenge for Automated Prediction of MCI from MRI Data and also our approach that gave us the 1st position. Finally, section 5 describes briefly the methodology reported from the best team giving an explanation of why our approach gave the highest accuracy. Future directions of machine learning in neuroimaging for the designing of reliable biomarkers is given in the discussion part.
2. A review of Random forest algorithm for the classification of neuroimaging data in Alzheimer's Disease

Focusing on neuroimaging single or multi-modal studies tailored to AD that adapted RF in their analysis, we found heterogeneous analytic strategies. Particularly, for the classes diagnosis (healthy control, stable or progressive MCI, AD), two studies (Tripoliti et al., 2007; Lebedev et al., 2014) investigated the binary classification between Alzheimer’s patients (AD) and healthy controls (HC), four studies focused on AD, HC and MCI (Cabral et al., 2013; Sivapriya et al., 2015; Maggipinto et al., 2017; Son et al., 2017), two studies explored the classification on AD, HC, stable MCI (sMCI), and progressive MCI (pMCI, converted to AD) (Gray et al., 2013; Moradi et al., 2015), two had only sMCI and pMCI (Wang et al., 2016; Ardekani et al., 2017), one had HC and MCI (Lebedeva et al., 2017) and one last one had AD, HC, and Lewy-body dementia (LBD) patients (Oppedal et al., 2015).

Two studies integrated FDG-PET and DTI measurements in their analysis (Cabral et al., 2013; Maggipinto et al., 2017) while others investigated structural MRI as a single neuroimaging modality (Lebedev et al., 2014; Moradi et al., 2015; Ardekani et al., 2017; Lebedeva et al., 2017) or in combination with features from other modalities like FDG-PET (positron emission tomography) (Gray et al., 2013; Sivapriya et al., 2015), fMRI (Tripoliti et al., 2007; Son et al., 2017), florbetapir-PET (Wang et al., 2016) and FLAIR (fluid-attenuated inversion-recovery) (Oppedal et al., 2015).

Based on the aforementioned studies, two of them didn’t specify the number of trees in the RF model (Moradi et al., 2015; Son et al., 2017) while in the rest eight cased, they reported a feature selection strategy (Tripoliti et al., 2007; Cabral et al., 2013; Lebedev et al., 2014; Moradi et al., 2015; Sivapriya et al., 2015; Ardekani et al., 2017; Lebedeva et al., 2017; Maggipinto et al., 2017).

Only a small portion of neuroimaging studies applied a multi-class classification approach. Cabral et al. (2013) succeeded an accuracy of 64.63% for the discrimination of AD vs. MCI vs. HC based on FDG-PET. Oppedal et al. (2015) reported an accuracy of 85% for HC vs. AD vs. LBD using texture features extracted from the T1 images in the white matter lesions masks. Sivapriya et al. (2015) showed an accuracy of 96.3% for the ternary problem AD vs. MCI vs. HC using MRI and FDG-PET. Finally, Son et al. (2017) reported a low accuracy (53.33 %) for the problem of separating AD vs. MCI vs. HC using resting-state fMRI and 3T-MRI using RF classifier.
Of great interest is the study of Lebedev et al. (2014) where they reported robust classification accuracies between ADNI and ADDNEUROMED Consortium based on morphological features extracted from 1.5T MRI. Importantly, they succeeded to quantify the sensitivity of morphological features to predict the conversion of MCI to AD focusing on follow-up studies. Fig. 2 illustrates how RF works in a classification problem. RF are an ensemble of k untrained decision trees which are trees with only a root node with M bootstrap samples. RF are trained using a variant of the random subspace method, which is a method of training multiple RF models by randomly samples the initial feature space. The main reason of this ensemble learning method is to reduce the correlation between the estimators by training different RF with random samples of features.

The procedure for training a random forest is as follows:

1. At the current node, randomly select \( p \) features from available features \( D \). The number of features \( p \) is usually much smaller than the total number of features \( D \).

2. Compute the best split point for tree \( k \) using the specified splitting metric (Gini Impurity, Information Gain, etc.) and split the current node into daughter nodes and reduce the number of features \( D \) from this node on.

3. Repeat steps 1 to 2 until either a maximum tree depth \( l \) has been reached or the splitting metric reaches some extrema.

4. Repeat steps 1 to 3 for each tree \( k \) in the forest.

5. Vote or aggregate on the output of each tree in the forest.
3. A review of Machine Learning Techniques in Neuroimaging Data

PAMI methods can detect alterations of MR-protocol differences in disease groups compared to controls and also in intervention protocols after intense cognitive and physical training tasks (Rathore et al., 2017). Another relevant issue that PAMI algorithms stress and demands the decision of the neuroscientist is the analysis of the original high-dimensional neuroimaging dataspace versus neuro-anatomical parcellated atlases with brain areas defined be atlases like the AAL (Tzourio-Mazoyer et al., 2002), Oxford-Harvard atlas, etc or adaptive atlases (Lu et al., 2015). These ROI are separated according to histological and functional activation maps. Parcelled features have many advantages in terms of memory cost, computational time, preprocessing time and the derived results can be compared with many other existing studies. However, this type of analysis, fixed for every subject, in any disease state, condition, intervention protocol and across the lifespan introduced a significant bias. In contrast, the analysis of the original high-dimensional feature human brain space is unbiased but it is more difficult to handle using PAMI algorithms that fit to the problem like in other

---

2 A redesign of the original inspired figure found from the following website

https://www.linkedin.com/pulse/random-forest-algorithm-interactive-discussion-niraj-kumar/
problems, like computer vision, 3D video processing, etc. Another problem with the high-dimensional space is the case where the number of measurements (e.g. estimates within every voxel or ROIs) is much larger than the number of observations (e.g. number of subjects in a study) and it is often called as “curse of dimensionality” (Bellman, 1961). This term can be used in many events when manipulating high-dimensional features that always hampers the efficacy of the adopted model. For that reason, a preparatory step of feature selection and dimensionality reduction is more than significant.

Neuroimaging gave the opportunity to neuroscientists to quantify alterations of pathological brain in various diseases such as the Alzheimer’s disease (AD) (Rathore et al., 2017) and also in neuropsychiatric disorders (Liu et al., 2015). Especially the integration of neuroimaging with machine learning techniques improved our knowledge about the structural and functional changes in the pathological brain. A recent systematic review on classification of neuroimaging reported that there is no single neuroimaging modality that can reach alone to an absolute accuracy for an automated AD prediction but only the integration of the best features from different modalities can effectively transform a pipeline into a clinical reality (Rathore et al., 2017). A recent study discussed the promises and pitfalls of single-subject prediction in brain disorders in neuroimaging (Arbabshirani et al., 2017). By surveying over than 200 studies focusing on schizophrenia, mild cognitive impairment (MCI), Alzheimer's disease (AD), depressive disorders, autism spectrum disease (ASD) and attention-deficit hyperactivity disorder (ADHD), they found that the most common pitfall of the reported classification results were the procedure of feature selection, cross-validation and the distinction of training-testing dataset, presentation of classification performance, avoiding overfitting, optimizing parameters like in kernels in SVM etc. Additionally, the need of a higher number of subjects per class was also discussed which is a common drawback in many studies. This pitfall will be changed by the increased number of open multimodal neuroimaging databases like the ADNI (Petersen et al., 2010), ENIGMA (Spurdle et al., 2012), CAMCAN (Taylor et al., 2017) etc.

A recent review study tailored to AD and RF models supported the idea that there is a complementary information between the modalities that can boost the accuracies of prediction and this richness of features should be explored by combining alternative classifiers compared to a single one (Sarica et al., 2017).

RF’s feature selection capabilities can be considered as very effective, while alternative algorithms have been proposed for the reduction of the feature space and in some cases, further
improved the accuracy of the RF model (Tuv et al., 2009). Given in the previous section the effectiveness and promise of RF as a bagging ensemble model, we encouraged neuroscientists to compare and integrate this algorithm with alternative machine learning techniques like deep learning (Vieira et al., 2017).

In the future, the integration of multi-PAMI approaches (RF, Deep-learning and SVM), multimodal imaging-based features (MRI, DTI, PET; Wang et al., 2016) and multi-site data repositories (Abraham et al., 2017) would drastically increase the effectiveness and reliability of potential automated prediction neuroimaging pipelines of clinical reality.

4. International Challenge for Automated Prediction of MCI from MRI Data

In a recent international challenge for automated prediction of MCI from MRI data, we succeeded in getting the 1st place among 19 worldwide teams³. The feature input was morphometric measures extracted from 3D T1 brain MRI images for ADNI1 cohort, including 60 Healthy Controls (HC), 60 early mild cognitive impairment (MCI), 60 late MCI (cMCI) and 60 stable AD. This was the very first attempt to simultaneously classify the four groups using a single MRI modality. An extra blind dataset of 160 subjects (HC: 40, MCI: 40, cMCI: 40 and AD: 40) was used by the organizers of the competition to evaluate the proposed machine learning scheme and to rank the participating teams.

In the following sections, we describe how the organizers selected the datasets from the ADNI database, the pre-processing steps, the proposed RF model, the feature selection strategy, the final results and also the machine learning algorithms.

4.1 Materials and Methods

4.1.1 Participants

In particular, MRIs were selected from the Alzheimer’s disease Neuroimaging Initiative (ADNI). ADNI is an international project that collects and validates neurological data, such as MRI and PET images, genetics or cognitive tests.

Organizers randomly and automatically selected subjects by employing the data analytics platform Konstanz Information Miner (KNIME).

³ https://www.kaggle.com/c/mci-prediction
This dataset was obtained by grouping a balanced number of subjects (100) for each of the four classes (HC, AD, MCI, cMCI) by various diagnostic criteria.

Finally, the whole dataset of 400 subjects was split by the organizers into a training dataset of 240 subjects (60 subjects for each of the four groups) and a testing dataset of 160 subjects (40 subjects for each of the four groups).

Table 1 summarizes the demographics of the training and testing datasets, including the average age, the gender contribution and the average MMSE.

**Table 1a. Summary of demographics for the training dataset**

<table>
<thead>
<tr>
<th></th>
<th>Age</th>
<th>Gender (male/females)</th>
<th>MMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>HC (60)</td>
<td>72.3 ± 5.7</td>
<td>62/58</td>
<td>29.2 ± 1.1</td>
</tr>
<tr>
<td>MCI (60)</td>
<td>72.2 ± 7.5</td>
<td>66/54</td>
<td>28.3 ± 1.6</td>
</tr>
<tr>
<td>cMCI (60)</td>
<td>73.0 ± 7.3</td>
<td>60/60</td>
<td>27.2 ± 1.9</td>
</tr>
<tr>
<td>AD (60)</td>
<td>74.8 ± 7.4</td>
<td>56/64</td>
<td>23.4 ± 2.1</td>
</tr>
</tbody>
</table>

**Table 1b. Summary of demographics for the testing dataset**

<table>
<thead>
<tr>
<th></th>
<th>Age</th>
<th>Gender (male/females)</th>
<th>MMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>HC (40)</td>
<td>74.9 ± 5.6</td>
<td>18/22</td>
<td>29.00 ± 1.1</td>
</tr>
<tr>
<td>MCI (40)</td>
<td>72.4 ± 8.1</td>
<td>23/17</td>
<td>27.6 ± 1.9</td>
</tr>
<tr>
<td>cMCI (40)</td>
<td>71.7 ± 6.3</td>
<td>25/15</td>
<td>27.6 ± 1.8</td>
</tr>
<tr>
<td>AD (40)</td>
<td>73.1 ± 8.2</td>
<td>23/17</td>
<td>22.7 ± 2.0</td>
</tr>
</tbody>
</table>

4.1.2 MR Image Acquisition

All participants were scanned on a Philips 3 T Achieva MRI scanner. The MRI data acquisition protocol is described in ADNI’s official webpage[^4].

4.1.3 Freesurfer processing and Features Extraction

[^4]: [http://adni.loni.usc.edu/methods/mri-analysis/mri-acquisition/]
T1-weighted Magnetic Resonance Images (MRI) were pre-processed by the organizers of Neuroimaging Challenge/Competition for an automated classification of MCI. Further details of the adapted pipeline can be found here: https://inclass.kaggle.com/c/mci-prediction.

MRIs were pre-processed by Freesurfer (v5.3) with the standard pipeline (recon-all -hippo-subfields) on a GNU/Linux Ubuntu 14.04 with 16 CPUs and 16Gb RAM. They used the KNIME plugin K-Surfer (Sarica et al., 2014) for extracting numerical data produced by Freesurfer into a table format. Organizers of the competition then enriched this table with both demographical and clinical parameters.

The set of features employed for the training procedure are:

- MMSE_bl - Mini-mental state examination total score at the baseline of the subject
- Age
- (i) cortical thickness, (ii) cortical surface area, (iii) cortical curvature, (iv) grey matter density, (v) the volume of the cortical and subcortical structures, (vi) the shape of the hippocampus and (v) Hippocampal subfields volume

4.2. Problem formulation

The organizers of the International Challenge for Automated Prediction of MCI from MRI data generated an additional 340 artificial test observations that were joined with the real blind test set (4 × 40 = 160) to form a combined test set of 500 observations. This testing sample was used in the online Kaggle competition platform for the evaluation of the classification performance (Sarica et al., 2017). This set, which can be called an artificial – Challenge dataset, was split into a public and private test set. The competition started online on 21 December 2016 and finalized on 1 June 2017. Every team that participated in this neuroimaging competition had the option of one submission per day. After every submission, the organizers returned, via the kaggle web system, the accuracy estimated over 500 subjects, where only 160 subjects were the real blind dataset. The rest (340 subjects – dummy) were created via a model based on the features from the training dataset. By the end of the challenge on 1 June 2017, the best performance of each team was evaluated and selected based on the private test set. The final evaluation and the ranking of the teams in terms of the classification accuracy was realized based on the Challenge test set which contains the real test data. Finally, the labels of the Challenge real test data and the related confusion matrices were released to the participants and teams that were invited to contribute to a special issue in Journal of Neuroscience Methods,
dedicated to the international challenge for the automated prediction of MCI using MRI data. Our team won the 1st position in this neuroimaging challenge.

4.3 The proposed RF model from our team

Our best submission was built around an ensemble of five classification models. The construction of these models was based on the well-known Random Forests (RF) machine learning method and its operational capabilities. More specifically, in all models, we performed feature selection using the Gini impurity index, a type of feature importance measurement commonly used in RF. In addition, we employed early fusion, as well as weighted fusion by means of late fusion schemes based on internal mechanisms provided by RF, namely the out-of-bag error and proximity ratios.

In what follows, the theoretical background of the involved methodologies, as well as a description of each classification model that was utilized in our experiments, are provided.

Our study focused on different scenarios with respect to the analysis and ranking of the feature space. We finally used an ensemble of five classification models and the final prediction of the blinded dataset’s labels was estimated via a majority voting scheme. More specifically:

1. The first model included the training of a RF classifier using the whole feature set and a feature selection strategy based on the Gini importance measure (a feature importance measurement typically used in RF), which provided the selected features for the final retrained RF model.
2. In the second model, we decided to split the initial feature space into left and right hemispheres (step A). Then, we ranked the hemisphere-specific features using the Gini importance measure (step B), we retrained the two RF classifiers using the selected features (step C) and finally, we applied weighted fusion for the formulation of the final predictions from the two RF models (Step D). The proximity ratio late fusion strategy (Liparas et al., 2014), derived from an operational feature of RF, namely the proximity matrix, was utilized in the weighted fusion step. This matrix includes the proximities between all data cases and is constructed for the entire RF model. For computing the weights for each considered class and for each hemisphere – modality, the ratio values between the inner-class and the intra-class proximities (for each class) are used (Zhou et al., 2010). For more details on the proximity ratio late fusion strategy, we refer to (Dimitriadis et al., 2017).
3. In the third model, we adopted the same strategy as in the second model, but regarding the weighted fusion step, the out-of-bag (OOB) late fusion strategy (Liparas et al., 2014) was used instead of the proximity ratio scheme. This late fusion strategy is based on the OOB error estimate, another operational feature of RF. For the weight computation step, the OOB accuracy values are computed separately for each considered class. These values are normalized (by dividing them by their sum) and serve as weights for each hemisphere – modality. For more details on the OOB late fusion strategy, please see (Dimitriadis et al., 2017).

4. In the fourth model, instead of retraining RF classifiers for the two modalities (as in Step C – second model) with the use of the final feature subsets, we trained Support Vector Machine (SVM) classification models. Finally, as a fusion step (step D – second model), we averaged the probability scores provided by the SVM models.

5. In the fifth model, steps A and B from the second model were applied in the same way, with the only difference being the use of a different threshold for the Gini importance measure (step B). Then, early fusion (also called feature-level fusion) was applied to the resulting feature space, produced by the concatenation of the two feature subsets from the two hemispheric modalities and finally, a new RF model was trained with the use of this new feature vector.

In the final step, the labels of the unknown cases were predicted with the use of the outputs of the classification models in the ensemble, and more specifically, a majority voting scheme was adopted. Practically, the predicted label for each blind sample was the one receiving the highest number of votes from the five classification models. In the case of ties, the highest probability estimate, derived from any of the adopted models, was used for the final prediction.

Regarding the parameters used in the experiments for the RF models of the ensemble, the following values were utilized for the number of trees for each RF model, as well as for the number \(p\) of the subset of variables used to determine the best split for each node during the growing of a tree (for each RF model):

- First model: Number of trees = 2000, \(p = 53\)
- Second and third models: Number of trees = 2000, \(p = \sqrt{D}\) (where \(D\) is the total number of features)
- Fifth model: Number of trees = 1000, \(p = 9\)

In Figure 3, the graphical layout of the ensemble’s five models, along with the parameter values used for the RF models, are provided. Additionally, boxplots for 9 features (for each
class) that were selected as important in the overall classification process are depicted in Figure 4. The more significant features were: the mini-mental state examination score, the bilateral hippocampal volume, the age, the bilateral amygdala volume and the bilateral inferior lateral ventricle volume.

We finally achieved a remarkable 61.9% classification performance for the simultaneous discrimination of four groups (HC, MCI, cMCI and AD) in the second blind dataset. It is the very first time in the literature where classification is performed simultaneously in a four-class AD-based problem using a single modality, namely MRI. We can clearly state that this performance is closed to plateau for the four-class problem using morphometric features from the MRI modality. A possible increment of this classification performance could be achieved by a subject-specific parcellation scheme and also by the adaptation of other features from neuropsychological battery, cerebrospinal analysis and other modalities, including BOLD activity and brain connectivity at resting-state and in cognitive tasks.

![Figure 3. Graphical layout of the ensemble’s classification models](image-url)
Figure 4. Boxplots for 9 features (for each class), selected as important in the ensemble’s models

5. Comparison with the alternative models from the competitive teams

Salvatore and Castiglioni (2018) adapted a Fisher’s Discriminant Ratio (FDR) for the feature ranking and a wrapper-optimization procedure was applied in order to identify the optimal subset of features to be used for the classification. They used SVM (support vector machine) with a linear kernel and C hyperparameter equals to 1. They ran 100 times the 5-fold cross-validation scheme via a binary scheme. For each subject, the six labels derived from the six binary classifications of the four groups were combined via three alternative voting schemes. Finally, the voting scheme mainly based on the binary-classification performances on the different four groups is the best choice to model the multi-label decision function for AD.

Amoroso et al., (2017) adapted a RF feature selection approach while they performed 100 times a 5-fold cross-validation scheme. For each round, they selected the 20 most important features. As a proper classification scheme, they used a Deep Neural Network (DNN) while for comparison a fuzzy logic algorithm has been applied. For a better robustness of the DNN model, they performed 30 different initializations. Finally, the label with the highest score derived from the sum from each model is assigned to every subject.

Ramirez et al., proposed a novel scheme tailored to the competition. They standardized the features to zero mean and unit variance while they used, a one-vs.-rest ANOVA feature selection algorithm. To further reduce the feature space, a partial least square (PLS) model was fitted to the training set. The highest performance for this team was succeeded with a bagging-trained ensemble of one-vs.-rest multiclass classifiers using PLS scores as input features. They adapted a random forest classifier (Breiman, 2001) using bagging, or bootstrap aggregating, forming an ensemble of classification and regression tree like classifiers. The final outcome of the classifier was determined by the majority vote across the trees’ outcome.

Nanni et al., (2018) tested four different feature selection algorithms (KernelPLS (KPLS), Fisher score (FS), Lagrange Multipliers (LM) and Mutual Information (MI)) and for well-known classifiers (Support Vector Machine (SVM), Gaussian Process Classifier (GPC), Random Subspace of Adaboost (RS AB), Random Subspace of Rotation Boosting (RS RB)). To improve further the resulted classification performance, they designed an ensemble of
classifiers based on a variation of the Static Classifier Selection which succeeded to give their best performance.

Sørensen et al., (2018) reported its best performance with an ensemble of support vector machines (SVMs) that combined bagging without replacement and a feature selection strategy. They selected the best feature set via sequential forward feature selection method and using SVM as an evaluator. The design of their best approach has been inspired by RF algorithm and it contained a combination of data subsets and feature subsets in ensemble SVM construction.

The superiority of our approach compared to the competitor teams is due to different strategies. First of all, we adapted different models to count in the final majority vote on the Challenge test dataset. These five models included four RF approaches and also one with SVM. Secondly, we ranked our features using the whole set and also by splitting them into left and right hemisphere. Third, in our second model, features were ranked with the Gini importance measure, RF classifiers were retrained using the selected features and finally, a weighted fusion step based on the proximity ratio late fusion strategy, a feature based on the operational capabilities of RF, was applied for the final predictions. Fourth, in our third model, we applied the out-of-bag (OOB) late fusion strategy instead of the proximity ratio scheme. Fifth, in another model, we set a different threshold for the Gini importance measure while an early fusion has been applied to the selected features. All these approaches outperformed even groups, where also RF has been applied to the train set.

It is important to mention here that in all our experiments, we used the training set as a training set without splitting it into train and test for internal cross-validation. Table 2 summarizes the ranking of classifiers’ accuracies as released by the organizers after the end of the competition.

Table 2. Ranking of classifiers’ accuracies as calculated at the closing of the competition, not including the entire test set without the fake set. For each team, the best result between the automatically-selected submission and the chosen submission is reported.

<table>
<thead>
<tr>
<th>Team Final ranking</th>
<th>Team Final ranking (Automatically Selected or Chosen by Teams)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stavros Dimitriadis – Dimitris Liparas (Dimitriadis et al., 2018)</td>
<td>0.61875</td>
</tr>
<tr>
<td>SiPBA-UGR (Ramirez et al., 2018)</td>
<td>0.5625</td>
</tr>
</tbody>
</table>
Recent advances in machine learning in neuroimaging conclude that the AD pathological brain can be reliably detected (Perrin et al., 2009). The majority of neuroimaging studies based on AD/MCI classification and also prediction of AD conversion used various modalities such as structural MRI, functional MRI, DTI, and PET (Rathore et al., 2017). The second most frequent set of features are derived from genetics, cognitive scores, neuropsychological assessments and also cerebrospinal fluid biomarkers (Melah et al., 2016). The majority of multimodal neuroimaging studies tailored to the design of biomarkers for the detection of prodromal stages of AD aggregated features from all these modalities and with the available feature selection algorithms, they finally choose the most informative giving also a ranking of the modalities according to their contribution (Liu et al., 2013; Yu et al., 2016; Korolev et al., 2016).

Apart from univariate features extracted from structural MRI, functional MRI, DTI, and PET modalities, functional and structural connectivity have contributed also on this race of designing reliable biomarkers for prodromal stages of MCI. Jie et al., proposed the extraction of both local and global features from fMRI-based connectivity approach at resting-state condition and a multi-kernel SVM for MCI classification (Jie et al., 2014). Khazaee et al. (2015) estimated network metrics from fMRI-based functional brain networks at resting-state that quantify integration and segregation, and using Fisher score for feature selection and SVM for classification. They succeeded a 100% classification between healthy controls and AD patients. The functional connectivity-based neuroimaging methods performed very well in binary classification approaches (97.00% for AD/MCI (Challis et al., 2015) and 91.90% for MCI/CN (Jie et al., 2014)) but they have never be tested on multi-class approach like with structural MRI during the international competition.
Structural-based MRI-based studies tailored to the detection of best biomarkers for AD focused on the extraction of morphological features like volumes, thickness etc (Liu et al., 2013) and also on the estimation of density maps of white matter (WM), grey matter (GM), and cerebrospinal fluid using the well-known voxel-based morphometry (VBM) (Ashburner and Friston, 2000). Liu et al., (2015) succeeded an overall 79% for the detection of stable versus progressive mild cognitive impairment.

In the majority of the machine learning multimodal neuroimaging studies (Liu et al., 2013; Melah et al., 2016) and also the aforementioned here attempted to select the best features from each modality rather than to select the best modality among the available. The selection of the most informative modality could be more important than the set of feature selection and the classification algorithms (Sabuncu and Konukoglou, 2015).

The most common feature selection/reduction algorithm is linear discriminant analysis (LDA). Park et al. (2013) applied LDA to cortical features from MRI images and trained SVM on MCI, healthy controls and tested on subjects that converted to AD. They reported a remarkable 83% on the prediction of MCI to AD.

The most common classifier reported on neuroimaging machine learning studies for AD is SVM (Rathore et al., 2017). Recent studies also reported very good results with RF using single or multimodal features, binary and also three-class classification problems. We reported those studies in section 4. However, the major take home message from the international competition was the plateau of a single modality to simultaneously differentiate the four groups using RF. We totally agree with neuroinformaticians that our main goal should be to combine different models derived from various classifiers and also their modifications as we performed during the challenge. Moreover, the available modalities share complementary information and a sophisticated aggregation of the best features across all modalities can further enhance the reliability of the biomarkers (Sarica et al., 2017).

In order to design multi-site biomarkers for AD, its prodromal stages and also for the accurate prediction of the subjects converted from MCI to AD, we will need large opened shared multimodal databases Poline et al., 2012) e.g. ADNI (Petersen et al., 201). However, the last decade, hundreds of papers have been published with ADNI database that practically cannot be compared because of different non-shared analytic pipelines and also different subcohorts that are not public available. From the release of ADNI database, a large amount of studies revealed their results enhancing our knowledge about AD so far. However, it is very
difficult to compare all these studies because they used a different subset of subjects from the original cohort, different pre-processing pipelines with open or in-house software and also different features derived from different modalities and also from different anatomical atlases (Liu et al., 2015). For that reason, neuroimaging preprocessing approaches should be also released by the authors under a common software package (Gorgolewski et al., 2015; Savio et al., 2017). Finally, any methodological advance of machine learning applied to neuroimaging for the scope of the design of a biomarker with a clinical evaluation should be tested over multi-site across different neuroimaging labs with the same or different systems/equipments (Abraham et al., 2016).
References


Alzheimer’s disease with a good between-cohort robustness. Neuroimage Clin. 6, 115–125. doi: 10.1016/j.nicl.2014.08.023


**Acknowledgement**

SID was supported by MRC grant MR/K004360/1 (Behavioral and Neurophysiological Effects of Schizophrenia Risk Genes: A Multi-locus, Pathway Based Approach. SD is also supported by MARIE CURIE COFUND EU-UK Research Fellowship.