Electronic Letters on Computer Vision and Image Analysis 12(1):42-56, 2013

Alzheimer's disease early detection from sparse data using brain importance maps

Andreas Kodewitz, Sylvie Lelandais, Christophe Montagne and Vincent Vigneron

IBISC Laboratory, University of Evry, 40 rue du Pelvoux, 91020 Evry cedex, France

Received 31th Jan 2013; accepted 27th May 2013

Abstract

Statistical methods are increasingly used in the analysis of FDG-PET images for the early diagnosis of Alzheimer's disease. We will present a method to extract information about the location of metabolic changes induced by Alzheimer's disease based on a machine learning approach that directly links features and brain areas to search for regions of interest (ROIs). This approach has the advantage over voxel-wise statistics to also consider the interactions between the features/voxels. We produce "*maps*" to visualize the most informative regions of the brain and compare the maps created by our approach with voxel-wise statistics. In classification experiments, using the extracted map, we achieved classification rates of up to 95.5%.

Key Words: Nuclear Imaging, Brain, Computer-aided diagnosis, Machine learning, Alzheimer's disease

1 Introduction

In 2010 there are 35.6 million people estimated to be affected by dementia worldwide. With the growth of the elder population this number is expected to rise to about 115 million by 2050 while there is no reliable early diagnosis method and effective treatment at hand. The most common cause of dementia is Alzheimer's disease (AD). Dementias do not only affect the patient himself but also his social environment. In many cases relatives cannot handle daily life without external help. Hence, dementias also charge health care systems with huge amounts of money (estimated to 604 billion US dollars in 2010) ([1]).

Positron emission tomography (PET) nuclear medicine provides a non-invasive, three-dimensional functional imaging method that measures the metabolism in the brain. AD typically causes bilateral hypo-metabolism in the temporal and parietal lobes, posterior cingulate gyri and precunei, as well as the frontal cortex that is imaged by the PET scan. Hypo-metabolism has been correlated to dementia severity by comparing measures of cognitive function, such as Mini-Mental State Exam (MMSE) and Clinical Dementia Rating (CDR), and cerebral metabolic rate of glucose ([2],[3]).

The American Alzheimer's Disease Neuroimaging Initiative (ADNI) collects data of patients affected by AD, mild cognitive impairment (MCI) and normal control (NC). This includes diagnosis based on mental state exams (e.g. MMSE), biomarkers, magnetic resonance imaging (MRI) and PET scans. With a collection of PET

Correspondence to: <kodewitz@iup.univ-evry.fr>

Recommended for acceptance by <Angel Sappa>
ELCVIA ISSN:1577-5097
Published by Computer Vision Center / Universitat Autònoma de Barcelona, Barcelona, Spain

scans, acquired with the tracer ^{18}F fluorodeoxyglucose (^{18}F-FDG), of more than 200 AD patients, 200 controls and 400 MCI patients, the ADNI database provides a fairly large data set for investigations. It is engaged in the development of early diagnosis and treatment of dementias, especially AD.

A huge amount of literature about AD suggests that first manifestations of AD occur years before symptoms are clinically detectable. This makes nuclear medicine a promising method for early diagnosis and its predictive abilities have been widely studied. Visual assessment by medical experts has been studied by several others, see e.g. [3, 4, 5] for a short review. As visual reading is very time consuming there is a growing interest in computer assisted/based PET analysis. Even well-trained experts may have a high inter-observer variation rate, therefore, computer-aided diagnosis (CAD) is needed to help neurologists in AD detection and classification. Recently, several CAD analysis have been studied to increase the diagnostic sensitivity and specificity: Sun et al. [6] performed a covariance estimation of anatomical volumes of interest (VOIs); Ishii et al. [7], Drzezga et al. [8], Scarmeas et al. [9] and several other groups used voxel-wise statistical methods such as statistical parametric mapping (SPM) and NEUROSTAT & 3D-SSP to find a feature set of AD lesions that can accurately distinguish normal/AD patients. Matrix factorization based analysis methods were proposed in [10, 11, 12, 13], but captured rather large scales and were not able to detect very localized changes. Finally, machine learning techniques such as linear discriminant analysis (LDA) or support vector machine (SVM) have shown to be also appropriate methods for classification of large brain data volume [14, 15, 16], but they generally did not help to explain the information content. The common character of these previous computer based works is that they do not incorporate information about the location of discriminant information that is already available. In addition, most of the CAD systems need a large number of samples to construct the models or rules.

This work will focus on the aspect of retrieval of information about the *localization* of early stage AD related metabolic changes in the patient brain. In such a framework, only some texture features obtained directly from the images or ROIs are used as inputs of classifiers. The advantages of such CAD system are its simple structure, the drastic data reduction ($\sim 95\%$) combined with an outstanding classification performance and its fast processing speed.

The remainder of this article is structured as follows. In section 2 we introduce the image database and our notations. Section 4 presents our approach to learn about the regions in the brain that are most informative in the classification of AD. Section 5 presents the results of both map extraction and classification experiments. In section 6 we discuss our findings and the article finishes with the conclusion in section 7.

2 Materials

Notations

Let us establish some definitions. Though brain scans are usually displayed as 2D slices through the brain for preprocessing and computer based analysis, it is more convenient to consider the scans as a 3D matrix with a three coordinates vector (j, k, ℓ), where $1 \leq j \leq J$, $1 \leq k \leq K$ and $1 \leq \ell \leq L$. J, K, L are put for the number of voxels in the three dimension. In this notation L corresponds to the number of transversal slices. Each point in the image $\mathtt{I}$ is called a voxel with intensity $\mathtt{I}(j, k, \ell)$. In some cases we prefer a vectorized notation. $\mathtt{I}(j, k, \ell)$ can be written as $\mathtt{I}(m) = \mathtt{I}(j + (k-1)J + (\ell-1)JK)$, with $1 \leq m \leq M$ and $M = JKL$ the total number of voxels in the brain volume. Note that physically $\mathtt{I}(j, k, \ell)$ counts the number of photons emitted by the radioactive marker at coordinate position (j, k, ℓ). Throughout, $X = (X^1, \ldots, X^p)$ is a vector of p feature components. The domain of X is the set of possible observations, denoted by $\mathcal{X}$, typically $\mathbb{R}^{+p}$. In this work, we study the relation between X and a discrete target variable Y that refers to the patient class, y a realization of Y. In this case we will use $y = \{+1, -1\}$ as it is a mathematically convenient notation; of course, the particular values of y only serve as indicators and have no meaning *per se*).

2.1 Image database

All data used in the preparation of this article were obtained from the ADNI database. 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 initial goal of the ADNI was to recruit 800 adults, ages 55 to 90, to participate in the survey, about 200 cognitively normal elder individuals to be followed for 3 years, 400 people with MCI to be followed for 3 years and 200 people with early AD to be followed for 2 years. 50% of the survey participants are selected for the acquisition of PET scans (see also table 1a). Our analysis is based on the ^{18}F-FDG PET data from ADNI participants, acquired with Siemens, General Electric and Philips PET scanners, that were collected from the web site of the Laboratory on NeuroImaging (LONI, University of California, Los Angeles) ([17]). The tracer ^{18}F-FDG accumulates in the brain cells according to their metabolic rate. Therefore, ^{18}F-FDG PET visualized the metabolism in the brain. In the preparation of our data set we used the meta information that is also available via the ADNI database.

	NC	MCI	AD	
Patients/baseline	102	207	95	404
Scans	433	956	313	1702

(a) ADNI FDG-PET preprocessed scans

	NC	AD	
Patients/baseline	84	82	166
MMSE	28 ± 2	23 ± 2	
Age	76 ± 5	75 ± 7	

(b) Selected scans

Table 1: Numbers of scans per patient selected with our criteria.

2.2 Patient and scan selection

In the creation of the dataset for a statistical analysis it is absolute necessary to restrict the selection to just one scan per patient as in the contrary case a patient recognition effect could influence the disease recognition process, most probably resulting in a positive biased classification result. We chose to select the first scan that was acquired for each patient, called baseline scan, as this represents best the real life case: the examiner aims to determine whether a patient is suffering from AD or not based on the PET scan to confirm clinical exams and exclude other diseases.
As confidence in class labels is of high importance in a supervised classification task and follow-up scans will not be used in this analysis, we applied more restrictive criteria for the attribution of class labels as in the ADNI database:

- Patients that were examined only once are excluded.
- For the AD group: CDR of 1 or 0.5 and MMSE 20-26 in all examinations.
- For the NC group: CDR of 0 and MMSE > 24 in all examinations.

This selection criteria result in a NC group of 84 patients with mean MMSE score of (28 ± 2) and an early AD group of 82 patients with mean MMSE score of (23 ± 2) (see Tab. 1b). The two groups remain age matched as initially in the ADNI survey. The age of NC patients is 76 ± 5 years and the age of the AD patients is 75 ± 7 at the moment of acquisition of the PET scan.
Despite the high interest in MCI, we preferred a complete exclusion of MCI images for two reasons. First to avoid derogation of the results by the influence of possible other diseases or statistical bias. Second, and most important, the AD subjects in our data set already present very mild to mild AD cases (CDR $\leq$ 1, MMSE 23 ± 2). During the review of meta-data we found approx. 85% of MCI patients to be stable MCI as expected according to a MCI to AD conversion rate in the range of 10-30% per year like stated in [4]. Excluding MCI

non-converters from our analysis as probability for actual AD is rather low, this leaves us a remainder of 24 MCI due to AD patients. Disposing of more than 80 patients in NC and AD group including a 24 patients MCI group would leave us with a rather unbalanced data set.
We registered the selected scans to the Montral Neurological Institute (MNI) PET brain template to achieve voxel-to-voxel comparability between all scans. To do so we used the Matlab® SPM toolbox (for details on the software see [18]). After this spatial preprocessing step the images have a bounding box of $91 \times 109 \times 91$ voxels. The number of voxels that is covered by the brain is about $650,000$, roughly 75% of the whole image.

3 Related Work

The principle of random forest (RF) is to combine many binary decision trees – so-called CART models – built using several bootstrap samples coming from the learning sample $L = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ made of n iid observations and choosing randomly at each node a subset of input features $\{X^{\epsilon(1)}, \ldots, X^{\epsilon(q)}\}, q \ll p$, where $\epsilon(\cdot)$ is a permutation operator. Each of these CARTs (see Breiman *et al*. [22]) for itself has a "weak" predictive power but the whole ensemble builds a fast classifier with high predictive power [20] even in a high dimensional space.

To obtain a RF of T trees with n examples and p features we proceed the following way:

Require: number of trees T, $q \ll p$ the number of features that has to be tested at each node to choose the best split.

1: **repeat**
2: Draw a random sample of examples of size n with replacement.
3: Grow a classification and regression tree (CART) with this sample of examples and calculate the best split between the features with a random selection of m features.
4: **until** T trees have been grown.

The classification for a new example is obtained by presenting the example to *all* trees in the forest. Each tree votes for the class of the new example and the random forest assigns the class that got the most votes (majority vote).

By drawing the sample for each tree at random with replacement while training the RF, several examples are drawn multiple times and others not at all. For a sample size of $n > 40$ the probability that an example appears in the sample is $1 - (1 - \frac{1}{n})^n \approx 0.632$. With a fraction of about 63.2% of examples to be used in the growing of a tree there is a remaining fraction of $100 - 63.2 = 36.8\%$ examples that were not used to grow the tree. These examples are called out-of-bag (OOB) examples. Presenting each tree in a sufficiently large forest its OOB examples we obtain enough votes for each example to obtain a class prediction by majority vote as in the classification of new examples. The prediction error of the OOB examples is a good estimate for the predictive power of the RF.

4 Methods

Generally the CAD systems for AD diagnosis involve 3 stages: image preprocessing, feature extraction and selection and classification. Our CAD system does not have image preprocessing component because only some features obtained directly from the images or ROIs are used as inputs of classifiers. Clinical experience tells us that AD does not affect all the brain evenly and that there is a large inter-patient variance where hypometabolism is visible in the PET images.

4.1 Finding all relevant features with the Random Forest

An optimum input dataset should have effective and discriminating features, which reduce the redundancy of feature space to avoid the "curse of dimensionality" problem. The "curse of dimensionality" problem is

based on the sampling density of the training data being too low to allow a meaningful estimation of a high dimensional classification function with the available finite number of training data.

Intuitively, it seems unreasonable to automatically use all the available voxels because the magnitude of var($\hat{y}$) is influenced by the number of input variables. Hence, we divided the image matrix into cubic patches and used the mean intensity of these patches as features. Assuming that metabolic changes are smaller than 15mm, we will only use patch sizes of $7\times7\times7$ voxels and smaller. A special treatment of the image extremities is superfluous as with patches of this size, the space not covered by the brain is at least 9 voxels. For each cubic sub-volume of the $i-$th 3D scan, the mean grey level was computed and merged together to form a column vector $X_i = (X_i^1, X_i^2, \ldots, X_i^p)^T$, with p the number of patches. Given the size of the brain scans, this feature vector is computationally more tractable, but still very demanding. Note that this rather simple feature, the mean intensity of each patch, was chosen to keep complexity low. However, more sophisticated features will also be examined in the future such as 3D local binary pattern (LBP) textural features (see [19]).

The *feature selection* process – also called the *wrapper* approach – is a type of inference that attempts to determine which of the features $X^i, 1 \leq i \leq p$, are "related" to the target variable Y. Here, the classifier is "wrapped" by a loop that monitors its performance as features are added or deleted. The wrapper approach is appealing because feature selection and supervised learning are essentially optimized together, as a complete learning system, rather than sub-optimized individually. In this paper a wrapper approach is proposed for using with RF, an ensemble learning machine recently introduced by Breiman [20]* and related to ensemble methods as proposed by [21].

4.2 Construction of a brain importance map

The quantification of the feature importance (abbreviated FI) is a key issue not only for ranking the features – from most to least important – before a stepwise estimation model but also to interpret data and understand the underlying phenomenons. The most widely used score for the importance of a given feature is the increasing in mean of the misclassification rate for a tree in the forest when the observed values of this feature are randomly permuted in the OOB samples [23].

For each tree t we consider the corresponding OOB sample and permute at random the values of the i-th feature of the sample. Then we compute the OOB error of t with these modified OOB data. The feature importance of the i-th feature is defined as the increase of OOB error after permutation. The more the increase of OOB error is, the more important is the feature.

The importance of a feature X_i in tree t is:

$$\xi_i^{*(t)} = \frac{\sum_{\alpha\in \mathrm{OOB}^{*(t)}} \delta(y_\alpha, \hat{y}_\alpha^{*(t)})}{q_t} - \frac{\sum_{\alpha\in \mathrm{OOB}^{*(t)}} \delta(y_\alpha, \hat{y}_{\alpha,\pi_i}^{*(t)})}{q_t}, \tag{1}$$

with $\delta(\cdot,\cdot)$ the Dirac delta-function which is 1 if its arguments are equal and zero in the other case, $\hat{y}_\alpha^{*(t)}$ the predicted classes for example α before, $\hat{y}_{\alpha,\pi_i}^{*(t)}$ the predicted classes for observation α after permuting the values of feature i and q_t the number of OOB cases in tree t. The overall importance of feature i is calculated from (1) as the mean over all trees:

$$\xi_i = \sum_{t=1}^{T} \xi_i^{*(t)}/T. \tag{2}$$

This measure is based on the assumption that the permutation of a predictor variable X_i with high importance will lead to a higher loss in classification accuracy than a less important predictor variable.

The patch importance gained in this manner can then be transformed to a voxel importance for display and further use on the grid of the original images just by normalizing the feature importance ξ_i onto the interval $[0, 1]$ and mapping the vector back to the original image grid. As each feature of the RF corresponds to a patch,

* but in principle it could be applied to any learning machine that produces a measure of feature importance.

the importance ξ_i of the i-th feature is equivalent to the importance $\xi(p)$ of patch p. To obtain the importance per voxel $\xi(m)$ we attribute to each voxel m part of patch p the importance $\xi(p)$.

To obtain a brain map as independent as possible from the used image set we bootstrapped the importance extraction, i.e. we repeated the calculation of feature importance with 50 different random samples of images out of the whole image set. The maps shown in Fig. 2 are the mean of those 50 maps. This measure of feature importance can be used as a map for the importance of brain regions

Further experiments (see section 5) show whether the feature selection and the distribution of importance over the brain are consistent for different patch sizes and whether there is a preferred patch size.

If a common map for all patch sizes is desired we propose to create a combined map using a weighted mean with the corresponding classification rate as weight. Hence, for a voxel m, the combined importance is given by:

$$\overline{\xi_m} = \sum_{\alpha \in \mathcal{S}} w_\alpha \xi_m^\alpha / \sum_{\alpha \in \mathcal{S}} w_\alpha \tag{3}$$

with $\mathcal{S}$ the set of maps –or patch sizes–, α the index on the number of maps to be combined ($1 \leq \alpha \leq |\mathcal{S}|$) and w_α the classification rate of the RF that computed the importance ξ_m^α.

4.3 Voxel-wise statistics and feature importance

We will compare the extracted importance maps with the ROIs provided by physicians and maps obtained by voxel-wise statistics.

In ^{18}F-FDG PET imaging, the diagnosis of AD is primarily based on the finding of bilateral temporo-parietal hypo-metabolism where all other patterns of hypo-metabolism are suspected to have an other cause than AD [3]. Besides temporal and parietal lobe regions of the brain that are often examined are hippocampus, posterior cingulate, and precuneus [24, 25]. We compare all these regions to the RF importance map. To gain more detail we divided the parietal lobe in inferior and superior parietal lobes and we divided the temporal lobe in inferior, middle and superior temporal lobes. This results in a total of 8 ROIs. To be able to calculate a statistics for the ROI information we generated a map composed of the ROIs using the Matlab® toolbox WFU PickAtlas created by [26]. This software is using the Talairach Daemon software [27] to translate semantic description of brain areas into a voxel-wise mask of the ROIs. The created map is shown in Figure 1. This map allows us to evaluate in which brain area the most important voxels lie and whether importance is locally related to a specific ROI or wide spread over the ROIs.

We use voxel-wise statistics in the spirit of analysis of variance to identify the set of voxels whose patterns clearly separate classes and sub-classes of voxels according to the category of patients. Let μ_{AD}^m and μ_{NC}^m be the estimated mean of the m-th voxel for the AD and NC classes, respectively; and let ${\sigma^2}_{AD}^m$ and ${\sigma^2}_{NC}^m$ be the estimated variance of the m-th voxel respectively, for $m = 1, 2, \ldots, p$. These values can be easily estimated from the feature database as follows:

$$\mu_{AD}^m = \frac{1}{N_{AD}} \sum_{j=1}^{N_{AD}} x_{j,AD}^m, \tag{4}$$

$$\mu_{NC}^m = \frac{1}{N_{NC}} \sum_{j=1}^{N_{NC}} x_{j,NC}^m, \tag{5}$$

$${\sigma^2}_{AD}^m = \frac{1}{N_{AD}} \sum_{j=1}^{N_{AD}} [x_{j,AD}^m - \mu_{AD}^m]^2, \tag{6}$$

$${\sigma^2}_{NC}^m = \frac{1}{N_{NC}} \sum_{j=1}^{N_{NC}} [x_{j,NC}^m - \mu_{NC}^m]^2, \tag{7}$$

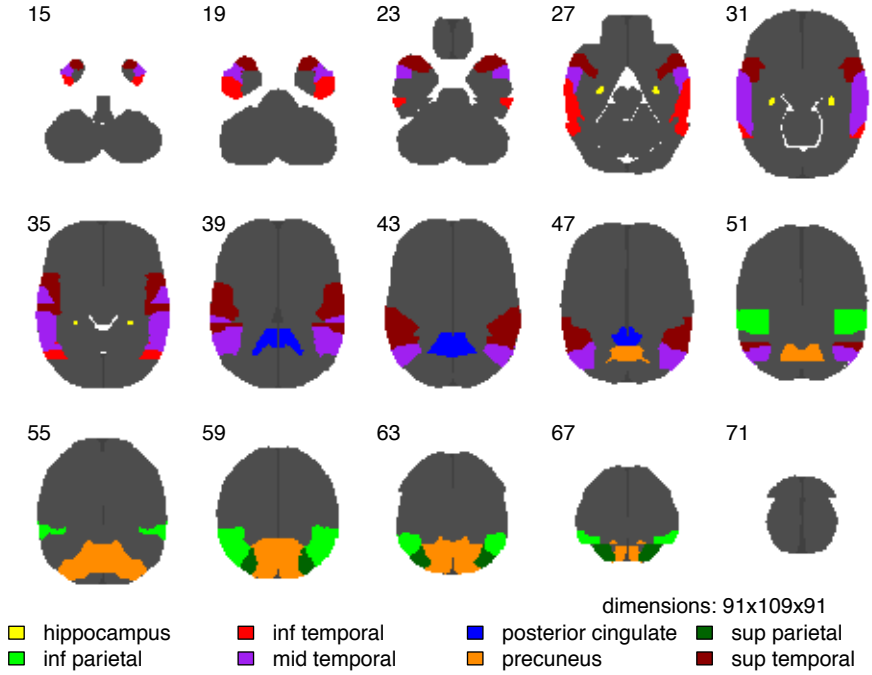


Figure 1: ROIs most commonly analyzed in AD detection.

where N_{AD} and N_{NC} is the total number of scans in the database labelled AD and NC. The class separation distance between the AD and NC classes for the m-th voxel is defined as

$$D_m = \frac{|\mu^m_{AD} - \mu^m_{NC}|}{\sqrt{1 + \sigma^{2m}_{AD} + \sigma^{2m}_{NC}}}. \tag{8}$$

and therefore points zones of potential interest for image analysis. The potential for discrimination capabilities of each voxel increases with D_m. In the performed experiments, the features with larger D_m correspond to large inertia. The remaining voxels led to poorer discriminative performance. The poor performance of the small inertia voxels originates in the AD affected areas being spread sparsely in the brain. The fact that the inertia presents some of the smaller class separation distances illustrates an important and very common problem in multiscale analysis: although regions may be composed of several structures of different scales, it is important to attempt to identify the *good* analyzing scales. In our specific case we have two indicators for a good analyzing scale available. From medical literature we know that we are searching for metabolic changes that are smaller than 14mm, i.e. 7 voxels. Focusing our attention on early stage AD we expect, priori, the metabolic changes to have a size of 4-8mm. A posteriori, we obtain a feedback from the RF classifier, whose classification performance depends on the analyzing scale.

Finally, we compare the distribution of our RF importance map with a map of the inter-class separation D_m (see the resulting map Fig. 2b): we show good congruence between voxels ranked by RF and voxels ordered by D_m. The two maps to compare however comprise different scales. The inter-class D_m map has in the case of the ADNI images a range of 0 to 0.15 approx. and the RF importance map is normalized to the interval of 0 to 1. Therefore we use for each map a separate threshold to select brain regions that are supposed to be of high interest in a AD classification task and compare the results via the number of selected patches. This allows to assess which map selects better features for classification at same number of features. As there is by design a direct relation between features and brain patches, this experiment can be equally interpreted as a ROI selection experiment. We used a SVM classifier with 100-fold bootstrap cross-validation to obtain an accurate estimate of prediction error for unknown examples. This avoids also possible effects of a double usage of the RF as map extraction and classification method [28].

5 Results

In results tables we will display the classification performance by giving the OOB error rate (OOB. err.)

$$\text{OOB. err.} = 1 - \frac{\sum_{\alpha \in \text{OOB}} \delta(y_\alpha, \hat{y}^*_\alpha)}{q}, \tag{9}$$

where $\hat{y}^*_\alpha$ is the predicted class label of example α and y_α the actual class label, the false positive rate (FPR) corresponding to the OOB error rate of the NC examples, i.e. 1-specificity, and the false negative rate (FNR) corresponding to the OOB error rate of the AD examples, i.e. 1-sensitivity. We display all values in percent and with their standard deviation.

5.1 RF Classification performance

In the training of the RFs used to calculate the feature importance we vary the patch size from $2 \times 2 \times 2$ to $7 \times 7 \times 7$. This range of patch sizes corresponds to a range of 2535 ($7 \times 7 \times 7$ patch) to 109350 ($2 \times 2 \times 2$ patch) features. Classification rates of the trained RFs declined gradually with augmenting patch size and the FNR was consistently 1-2% higher than the FPR, as shown Table 2. The RF used to calculate the feature importance was trained on the whole set of 84 NC and 82 AD patients. The RFs for the classification task had a size of 500 trees and at each split $m = \sqrt{p}$, with p the number of features, were tried to determine the best split. The variance between the 50 calculated importance maps was of the order of 10^{-2}. The maps were normalized to the interval $[0, 1]$.
Figure 2 visualizes the mean decrease in accuracy mapped back to the original image size as a colored overlay on our indexed map to facilitate interpretation of the gained importance map. The feature importance had for all patch sizes very similar distribution over the brain. This is why we fusion of all the maps in a single one (Fig. 2a). As classification rates obtained in the measurement of the importance showed significant differences for the different patch sizes (see table 2), we combine the single maps, after mapping back to the original image size, as proposed in section 4.2.
Total classification rate, classification rate for NC group (specificity) and for AD group (sensitivity) achieved by the RF classifier used to extract the feature importance are listed in Table 2. Best classification results are obtained with smallest patch sizes.

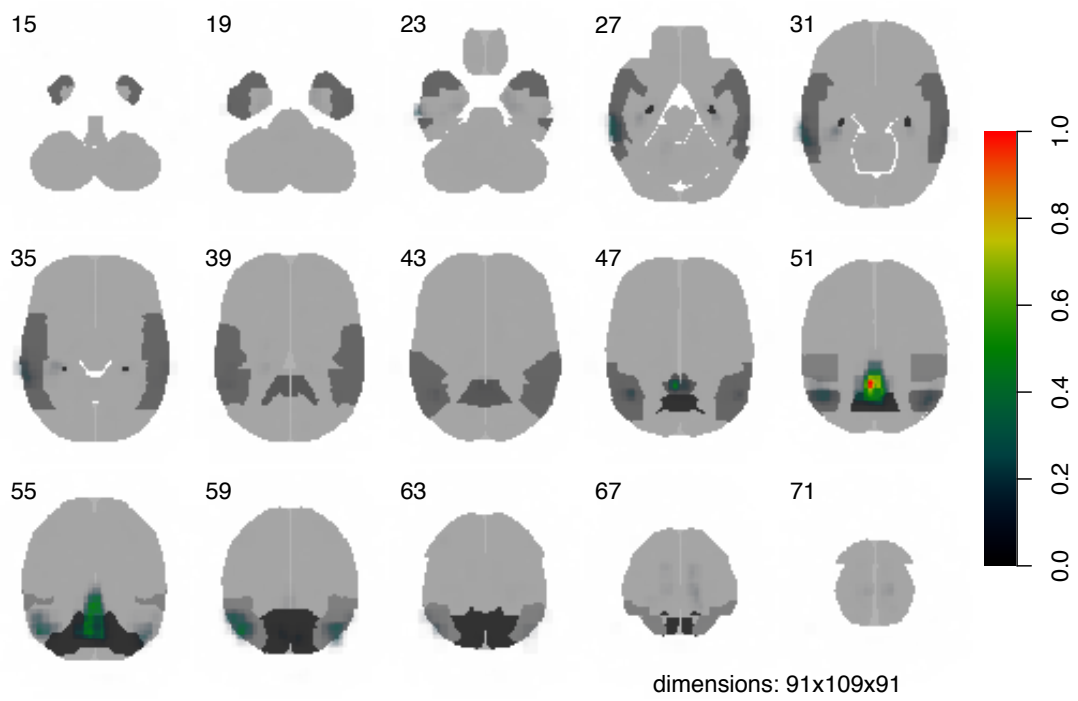
cube size	oob. err. [%]	FPR [%]	FNR [%]
$2 \times 2 \times 2$	17.0 ± 0.9	16.1 ± 1.4	17.9 ± 1.4
$3 \times 3 \times 3$	17.8 ± 1.1	16.9 ± 1.3	18.7 ± 1.8
$4 \times 4 \times 4$	18.5 ± 1.0	17.6 ± 1.2	19.4 ± 1.9
$5 \times 5 \times 5$	18.8 ± 1.0	17.6 ± 1.2	20.1 ± 1.8
$6 \times 6 \times 6$	20.5 ± 1.2	19.4 ± 1.6	21.8 ± 1.8
$7 \times 7 \times 7$	21.0 ± 1.1	20.2 ± 1.4	21.9 ± 1.8

Table 2: Error rates of the RFs used for importance measurement.

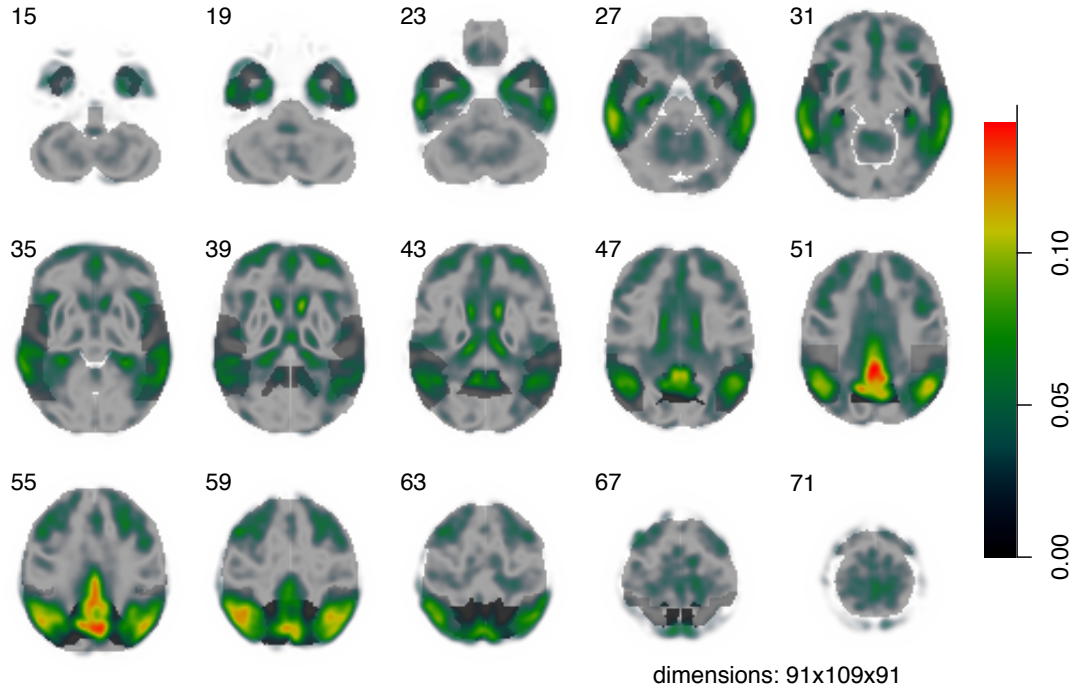
These results justify the confidence in the measured importance and for comparison. Note that there was no data reduction, in the sense that all voxels where used to obtain the features, in this experiment. The results obtained by map guided classification, using only a fraction of the whole data, are presented in the following section.

5.2 Comparison of RF importance map with ROI map and Class-separation distance

Comparing class-separation distance map and the RF importance map visually it is immediately apparent that the RF importance map points much smaller ROIs, i.e. regions of high importance (compare Figures 2a and



(a) weighted mean of importance images of patch size $2 \times 2 \times 2$ to $7 \times 7 \times 7$; a detail figure of slices 48 to 55 can be found in Figure S.1 (supplementary material)



(b) class-separation distance map

Figure 2: Overlay visualization of (a) RF feature importance and (b) class-separation distance in AD vs. NC classification obtained; red indicates highest importance.

2b). To assess the distribution of important voxels over the ROIs defined in Fig. 1 we plotted an importance histogram for each ROI. These histograms, shown in Fig. 3, confirm the visual impression. All histograms besides those of inferior parietal lobe, posterior cingulate and precuneus have a very heavy lower tail. It means most of the voxels forming the ROI are not important at all. In Tables 3 and 4, summarizing the key figures of the histograms, it becomes even more visible that both maps concurrently point posterior cingulate and precuneus as ROIs of high interest in the classification of AD vs. NC. As visible in Table 3, all voxels exceeding 75% of maximum importance are situated in posterior cingulate, precuneus or outside the defined ROIs. The amount of voxels of high importance not contained in a ROI is at least a magnitude lower than the amount of voxels in posterior cingulate or precuneus. A specific look-up of those voxels not contained in a ROI revealed that those voxels are located in close vicinity to posterior cingulate and/or precuneus. Superior temporal lobe, superior parietal lobe and hippocampus do not contain any voxels exceeding a RF importance of 0.25.

On the contrary, the class-separation distance has less voxels with very low class-separation distance. The percentage of voxels between .25 and .5 of maximum class-separation distance in Table 4 is much more important than the corresponding group of voxels of the RF importance map. The class-separation distance points posterior cingulate and precuneus as important regions in the brain as the RF importance map, but also inferior parietal lobe. The percentage of important voxels in this lobe is even higher than in posterior cingulate. In the RF importance map this area is of no special importance. The important areas obtained with this measure are much larger than those obtained with the RF method.

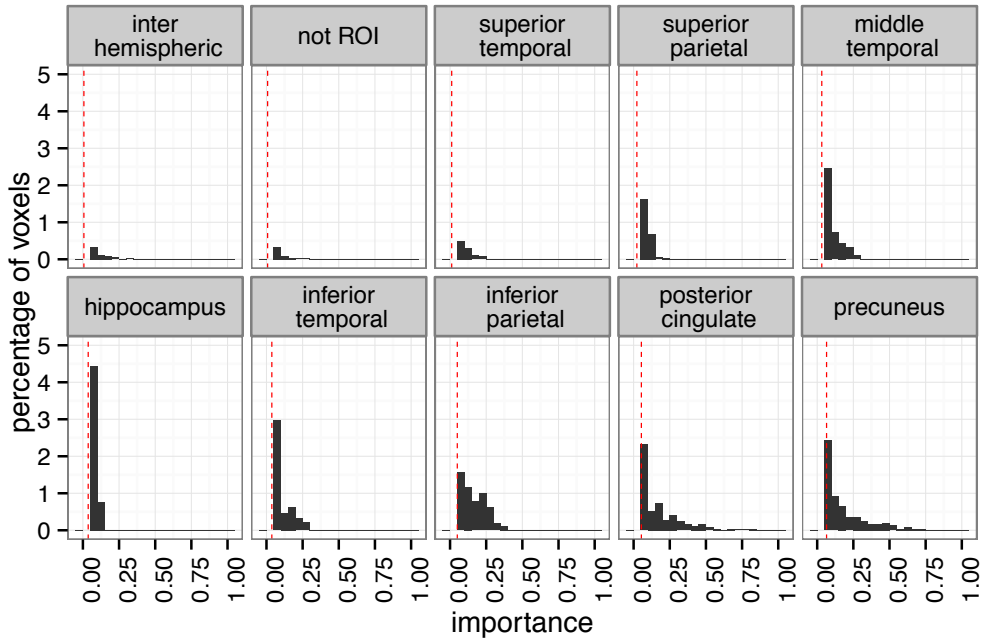


Figure 3: Distributions of voxel importance in the ROIs defined in figure 1. The red dashed lines indicate the average importance of the region. Numeric results see Table 3.

5.3 RF importance in feature selection

The importance maps extracted from the ADNI data in the preceding section are now used in a feature selection experiment. We tested the map fusioned by the weighted mean and the map extracted with patch size $2 \times 2 \times 2$ as this patch size yielded the best classification result (see Table 2). For comparison we used a map of class-separation distance (see figure 2b). A threshold is applied to the maps and only voxels whose importance, respectively class-separation distance, is exceeding the threshold are used in the classification. This threshold

	importance		% voxels in bin		
	mean	max	.25 to .5	.5 to .75	$> .75$
background voxels	1.13e-03	0.31	0.0025	0	0
not ROI	7.96e-03	1.00	0.24	0.078	0.032
inter hemispheric	6.87e-03	0.50	0.16	0	0
hippocampus	3.66e-02	0.13	0	0	0
inferior parietal	5.00e-02	0.40	4.6	0	0
inferior temporal	3.71e-02	0.28	1	0	0
middle temporal	3.18e-02	0.31	0.6	0	0
posterior cingulate	5.35e-02	**0.83**	**5.4**	**0.8**	**0.3**
precuneus	6.49e-02	**0.92**	**5.8**	**1.6**	**0.14**
superior parietal	2.24e-02	0.21	0	0	0
superior temporal	1.14e-02	0.24	0	0	0

Table 3: Comparison of the importance map obtained by RF (see fig. 2a) and the map of ROIs (see fig. 1).

	class-sep. dist.		% voxels in bin		
	mean	max	.25 to .5	.5 to .75	$> .75$
background voxels	1.27e-03	0.10	0.45	0.049	0
not ROI	2.19e-02	0.15	16	1.3	0.27
inter hemispheric	2.73e-02	0.13	25	3.1	0.64
hippocampus	4.91e-02	0.08	73	3.8	0
inferior parietal	4.75e-02	0.12	23	**24**	**4.4**
inferior temporal	4.94e-02	0.11	59	13	0
middle temporal	4.42e-02	0.11	51	10	0.01
posterior cingulate	4.69e-02	**0.13**	**39**	**17**	**2.7**
precuneus	5.20e-02	**0.14**	**36**	**20**	**7.1**
superior parietal	4.16e-02	0.11	40	14	0
superior temporal	3.11e-02	0.11	28	3.9	0.038

Table 4: Comparison of the class-separation distance map and the map of ROIs.

is swept over the whole scale as there is no analytical method to determine a good level for this value. All other voxels are dropped. As class-separation distance and feature importance of RF inhibit different scales the results are compared via the selected number of patches. We display classification rates for a patch size of $3 \times 3 \times 3$ as this patch size resulted in best classification rates.

Figure 4 compares the obtained classification results. With all three examined maps we observe a significant raise of the OOB error rate if we choose a threshold selecting less than 750 patches. With a threshold that selects more than about one thousand patches, the OOB error rate is stable for all three maps. The best OOB error rate of 4.48% is achieved with the fusioned RF importance map using 2500 patches. The $2 \times 2 \times 2$ importance map achieved an OOB error rate of 4.52% using 2000 patches. The difference between these results (0.04%) is not significant (p-value < 0.5). For both importance maps we can assume that the best classification outcome with this setup is achieved selecting 2000 to 2500 patches, because both the classification error and the standard deviation of the classification error over the bootstrap samples attain their minimum in this range of patches. The class-separation distance map performed best with 5300 patches with an OOB error rate of 5.9%. The numerical results, presented in Table S.1 (supplementary material), show that the OOB error rate is slightly increasing for large numbers of patches.

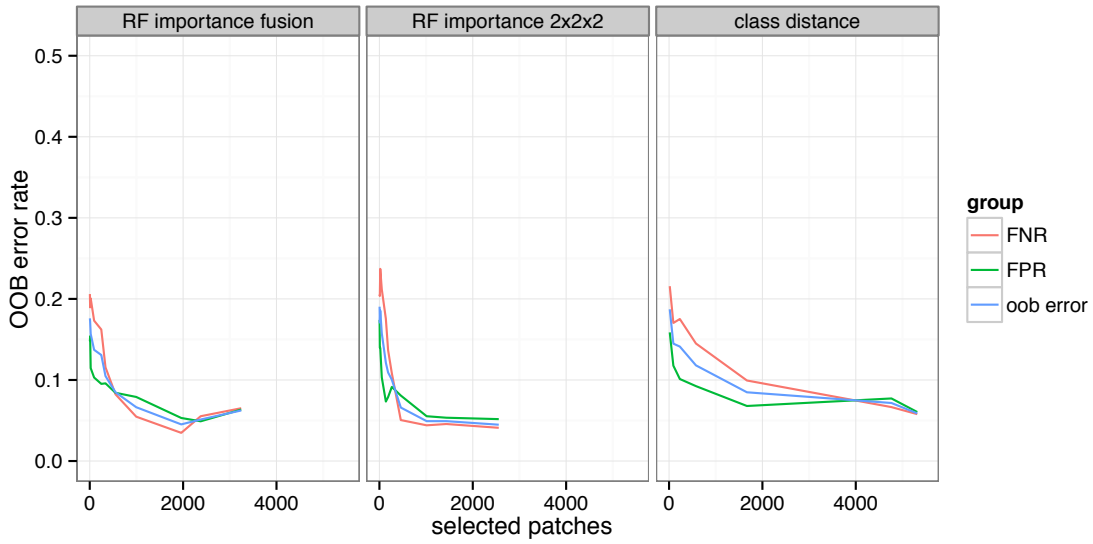


Figure 4: Feature, or equivalently ROI selection, by applying thresholds to RF importance map (left & middle) and class-separation distance map (right). The patch size is $3 \times 3 \times 3$ voxels.

6 Discussion

The classification rates of the RFs used to extract the feature importance reduced with growing patch size. According to the numeric results, we conclude that the changes in brain metabolism due to AD have a typical scale of 6mm or smaller as best classification results were obtained for a patch size of 2 and 3 voxels cube.

The maps for the importance of different brain areas that were obtained by our machine learning approach using the RF classifier indicated very similar areas of interest over all patch sizes. Those areas are much smaller than expected after literature study, but, in accordance with literature, localized mainly in parietal and temporal lobe. As regions of highest interest for the distinction of NC and AD using ^{18}F-FDG PET scans we found posterior cingulate and precuneus. A certain amount of voxels not included in any of the defined ROIs was found to be also of high interest. In fact, the voxels of maximum RF importance and class-separation distance were not contained in any defined ROI but in close vicinity to posterior cingulate and/or precuneus. This finding might be due to the imperfectness of registration to the MNI standard brain. We suggest therefore to consider slightly enlarging ROIs when using a model similar to ours.

Comparing the fusioned RF importance map with a class-separation distance map in a feature selection experiment we observed several characteristics that show an advantage of using the RF importance: overall classification rate is higher, difference between sensitivity (rate of correctly classified AD examples) and specificity (rate of correctly classified NC examples) is lower and the OOB error rate remains low for a smaller number of voxels.

The differences between fusioned importance map and the $2 \times 2 \times 2$ importance map were insignificant. The fusioned importance map yields a slightly superior classification rate.

In all our experiments the classification rate attains a maximum at $\sim 95.5\%$. We ascribe this limit to the use of the mean intensity as feature. By calculating the mean of each patch the structural information contained in each patch is lost. These high classification rates once more confirm the quality of the map we have extracted from the data.

7 Conclusion

The results of our experiments show that RF feature importance is a major indicator for clustering brain areas in the classification of early stage AD. All extracted maps are fully consistent with medical findings. As a map of zones of interest, it is localizing the classification relevant information more precisely than class-separation distance and therefore performs better than the same in feature selection. We reduced the number of features from 450000 to 2000 without suffering a loss in classification accuracy. This reduction of features will allow us to apply features that are computationally costly in our future research.

Our rather simple classification approach achieved a high reduction of the number of features and outstanding classification rates of up to 95.5%. We think our results contribute important information about the location of AD related changes in the brain and can be a base for further, very localized, analysis of PET images with computationally intensive image analysis and machine learning techniques.

8 Acknowledgements

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: Abbott; Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Amorfix Life Sciences Ltd. ; AstraZeneca; Bayer HealthCare; 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.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Medpace, Inc.; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; Novartis Pharmaceuticals Corporation; Pfizer Inc.; 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 California, Los Angeles. This research was also supported by NIH grants P30 AG010129, K01 AG030514, and the Dana Foundation.

References

[1] A. Wimo and M. Prince, "World alzheimer report 2010", `http://www.alz.co.uk/research/world-report`, September 2010.

[2] H. Braak and E. Braak, "Neuropathological stageing of alzheimer-related changes", *Acta Neuropathol*, 82(4):239-259, 1991.

[3] J. M. Hoffman, K. A. Welsh-Bohmer, M. Hanson, B. Crain, C. Hulette, N. Earl, and R. E. Coleman, "FDG PET imaging in patients with pathologically verified dementia", *J Nucl Med*, 41(11):1920-1928, 2000.

[4] L. Mosconi, M. Brys, L. Glodzik-Sobanska, S. D. Santi, H. Rusinek, and M. J. de Leon, "Early detection of alzheimer's disease using neuroimaging", *Experimental Gerontology*, 42:129-138, 2007.

[5] S. Ng, V. L. Villemagne, S. Berlangieri, S.-T. Lee, M. Cherk, S. J. Gong, U. Ackermann, T. Saunder, H. Tochon-Danguy, G. Jones, C. Smith, G. O'Keefe, C. L. Masters, and C. C. Rowe, "Visual Assessment Versus Quantitative Assessment of 11C-PIB PET and 18F-FDG PET for Detection of Alzheimer's Disease", *J Nucl Med*, 48(4):547-552, 2007.

[6] L. Sun, R. Patel, J. Liu, K. Chen, T. Wu, J. Li, E. Reiman, and J. Ye, "Mining brain region connectivity for alzheimer's disease study via sparse inverse covariance estimation", in *KDD '09 Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, Paris, 1:1335-1344, 2009.

[7] K. Ishii, F. Willoch, S. Minoshima, A. Drzezga, E. P. Ficaro, D. J. Cross, D. E. Kuhl, and M. Schwaiger, "Statistical brain mapping of 18F-FDG pet in alzheimer's disease: Validation of anatomic standardization for atrophied brains," *J Nucl Med*, 42(4):548-557, 2001.

[8] A. Drzezga, N. Lautenschlager, H. Siebner, M. Riemenschneider, F. Willoch, S. Minoshima, M. Schwaiger, and A. Kurz, "Cerebral metabolic changes accompanying conversion of mild cognitive impairment into alzheimer's disease: a pet follow-up study", *European Journal of Nuclear Medicine and Molecular Imaging*, 30:1104-1113, 2003.

[9] N. Scarmeas, C. G. Habeck, E. Zarahn, K. E. Anderson, A. Park, J. Hilton, G. H. Pelton, M. H. Tabert, L. S. Honig, J. R. Moeller, D. P. Devanand, and Y. Stern, "Covariance pet patterns in early alzheimer's disease and subjects with cognitive impairment but no dementia: utility in group discrimination and correlations with functional performance", *NeuroImage*, 23(1):35-45, 2004.

[10] P. Markiewicz, J. Matthews, J. Declerck, and K. Herholz, "Robustness of multivariate image analysis assessed by resampling techniques and applied to FDG-PET scans of patients with alzheimer's disease," *NeuroImage*, 46(2):472-485, 2009.

[11] F. Nobili, D. Salmaso, S. Morbelli, N. Girtler, A. Piccardo, A. Brugnolo, B. Dessi, S. Larsson, G. Rodriguez, and M. Pagani, "Principal component analysis of FDG PET in amnestic MCI", *European Journal of Nuclear Medicine and Molecular Imaging*, 35:2191-2202, 2008.

[12] E. Salmon, N. Kerrouche, D. Perani, F. Lekeu, V. Holthoff, B. Beuthien-Baumann, S. Sorbi, C. Lemaire, F. Collette, and K. Herholz, "On the multivariate nature of brain metabolic impairment in alzheimer's disease", *Neurobiology of Aging*, 30(2):186-197, 2009.

[13] 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*, 34(3):985-995, 2007.

[14] I. A. Illán, J. M. Górriz, J. Ramírez, D. Salas-Gonzalez, M. López, F. Segovia, C. G. Puntonet, and M. Gómez-Rio, "^{18}F-FDG PET imaging for computer aided Alzheimer's diagnosis", *Information Sciences*, 181(4):903-916, 2011.

[15] J. M. Górriz, A. Lassl, J. RamÌrez, D. Salas-Gonzalez, C. Puntonet, and E. Lang, "Automatic selection of rois in functional imaging using gaussian mixture models", *Neuroscience Letters*, 460(2):108-111, 2009.

[16] J. Ramírez, R. Chaves, J. M. Górriz, I. Álvarez, D. Salas-Gonzalez, M. López, and F. Segovia, "Effective detection of the alzheimer disease by means of coronal NMSE SVM feature classification", in *ISNN 2009 Proceedings of the 6th International Symposium on Neural Networks: Advances in Neural Networks - Part II*, Wuhan, 337-344, 2009.

[17] ADNI, "Alzheimer's disease neuroimaging initiative", `http://www.loni.ucla.edu/ADNI/`, 2011.

[18] K. Friston, J. Ashburner, J. Heather *et al.*, "Statistical parametric mapping (SPM)", `www.fil.ion.ucl.ac.uk/spm`, 2011.

[19] C. Montagne, A. Kodewitz, V. Vigneron, V. Giraud, and S. Lelandais, "3D Local Binary Pattern for PET image classification by SVM, Application to early Alzheimer disease diagnosis", *Proc. of the 6th International Conference on Bio-Inspired Systems and Signal Processing (BIOSIGNALS 2013)*, Barcelona, 2013, (to appear). [Online]. Available: `http://hal.archives-ouvertes.fr/hal-00762837`

[20] L. Breiman, *Setting Up, Using, And Understanding Random Forests V4.0*, February 2003. [Online]. Available: `ftp://ftp.stat.berkeley.edu/pub/users/breiman/Using_random_forests_v4.0.pdf`

[21] T. G. Dietterich, "An experimental comparison of three methods for constructing ensembles of decision trees : Bagging, boosting, and randomization", *Machine Learning*, 40(2):139-158, 2000.

[22] L. Breiman, J. Friedman, R. Olshen, and C. Stone, *Classification and Regression Trees*, Wadsworth and Brooks, 1984.

[23] C. Strobl, A.-L. Boulesteix, T. Kneib, T. Augustin, and A. Zeileis, "Conditional variable importance for random forests", *BMC Bioinformatics*, 9:307, 2008.

[24] Y. Fan, S. M. Resnick, X. Wu, and C. Davatzikos, "Structural and functional biomarkers of prodromal alzheimer's disease: A high-dimensional pattern classification study," *NeuroImage*, 41(2):277-285, 2008.

[25] J. Valla, J. D. Berndt, and F. Gonzalez-Lima, "Energy hypometabolism in posterior cingulate cortex of alzheimer's patients: superficial laminar cytochrome oxidase associated with disease duration," *J Neurosci*, 21(13):4923-4930, 2001.

[26] J. A. Maldjian, P. J. Laurienti, J. H. Burdette, and R. A. Kraft, "An automated method for neuroanatomic and cytoarchitectonic atlas-based interrogation of fMRI data sets", *NeuroImage*, 19:1233-1239, 2003.

[27] J. L. Lancaster, M. G. Woldorff, L. M. Parsons, M. Liotti, C. S. Freitas, L. Rainey, P. V. Kochunov, D. Nickerson, S. A. Mikiten, and P. T. Fox, "Automated talairach atlas labels for functional brain mapping", *Human Brain Mapping*, 10(3):120-131, 2000.

[28] U. Gromping, "Estimators of relative importance in linear regression based on variance decomposition", *The American Statistician*, 61:139-147, 2007.