ORIGINAL ARTICLE

CrossMark

The Residual Center of Mass: An Image Descriptor for the Diagnosis of Alzheimer Disease

Alexandre Yukio Yamashita[1] ⬤ · Alexandre Xavier Falcão[2] · Neucimar Jerônimo Leite[2] ·
Alzheimer's Disease Neuroimaging Initiative

Abstract

A crucial quest in neuroimaging is the discovery of image features (biomarkers) associated with neurodegenerative disorders. Recent works show that such biomarkers can be obtained by image analysis techniques. However, these techniques cannot be directly compared since they use different databases and validation protocols. In this paper, we present an extensive study of image descriptors for the diagnosis of Alzheimer Disease (AD) and introduce a new one, named *Residual Center of Mass* (RCM). The RCM descriptor explores image moments and other techniques to enhance brain regions and select discriminative features for the diagnosis of AD. For validation, a Support Vector Machine (SVM) is trained with the selected features to classify images from normal subjects and patients with AD. We show that RCM with SVM achieves the best accuracies on a considerable number of exams by 10-fold cross-validation — 95.1% on 507 FDG-PET scans and 90.3% on 1374 MRI scans.

Keywords Diagnosis · Neuroimaging · Alzheimer disease · Machine learning · Support vector machine · Image analysis · Biomarker

✉ Alexandre Yukio Yamashita
  alexandre.yamashita@eldorado.org.br

  Alexandre Xavier Falcão
  afalcao@ic.unicamp.br

  Neucimar Jerônimo Leite
  neucimar@ic.unicamp.br

[1] Departamento de Software B, Instituto de Pesquisas Eldorado, Campinas, Brazil

[2] Institute of Computing, University of Campinas, Campinas, Brazil

Introduction

The Alzheimer Disease (AD) is a neurodegenerative disorder that critically affects memory, reasoning, and behavior. It is the most common type of dementia, accounting for 60%–80% of the cases (Alzheimer's Association 2017). Worldwide, it is estimated that 47.5 million people live with dementia and such prevalence is expected to double in 20 years (World Health Organization 2017). There is no cure for AD, but a suitable treatment can relieve the symptoms and reduce its aggravation. Its early and accurate diagnosis is crucial to improve the life quality of the patient, but this is a complex task that requires cognitive and objective tests, patient records, clinical and laboratory exams.

Recent works have shown that biomarkers obtained by image processing and machine learning techniques can aid the diagnosis of AD (Chincarini et al. 2011; Garali et al. 2016; Liu et al. 2014). Machines might also provide a more accurate diagnosis than clinicians, because they are free from fatigue and can deal with neurodegenerative patterns of difficult visualization (Casanova et al. 2011; Klöppel et al. 2008). Effects of aging also cause brain changes, making more difficult the effective pattern identification (Ambastha 2015).

⌕ Springer

Many works have reported high accuracies when using different image databases, pattern classifiers, and validation protocols. This makes impossible a comparative analysis of their image descriptors. In this work, we present an extensive study of image descriptors for the diagnosis of AD and introduce a new one, named *Residual Center of Mass* (RCM). RCM explores image moments and other operations to enhance brain regions and select the most relevant features for the diagnosis of AD. For validation, a Support Vector Machine (SVM) is trained with the selected features to classify images from Normal Control (NC) subjects and patients with AD. We show that RCM with SVM achieves the best accuracies on a considerable number of exams — 507 Fluorodeoxyglucose-Positron Emission Tomography (FDG-PET) scans and 1,374 Magnetic Resonance Imaging (MRI) scans, as provided by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (Jack et al. 2008).

The related works are described in Section "Related Works", making clear the motivation and impact of this paper. Section "Materials" presents the databases and the image preprocessing techniques used for the experiments. The RCM descriptor is introduced in Section "Methods" and its comparative study is presented in Section "Results and Discussion". Finally, Section "Conclusion" states conclusion and provides direction to future research.

Related Works

Neuroimaging methods for Computer-Aided Diagnosis (CAD) are composed by different techniques of image processing and classification, as shown in Fig. 1. Each method has each its own pipeline of processes described in the following. The images are preprocessed to remove the skull and muscle tissues, both irrelevant for classification. Then, the images are spatially normalized (e.g., registered into a reference image space) to account for the natural differences in size and shape of the brain. The voxel intensities are also normalized to correct the large variations caused by the use of different scanners and parameters. Finally, the images may be resized to a lower resolution, for the sake of efficiency, and smoothed by a low-pass filter to reduce the effects of misregistration.

In the next step, some works extract image features and/or apply feature-space reduction techniques. Subsequently, the image descriptor results from feature selection techniques and/or regions/volumes of interest (ROIs/VOIs). ROIs are image regions that represent the spatial extension of brain structures, which are interactively delineated or automatically segmented based on object models, such as a probabilistic atlas (Carmichael et al. 2005). VOIs are volumetric blocks extracted from the image at given locations, as defined by prior knowledge in some reference image space. Finally, a classifier is trained from the image descriptor and evaluated by some validation method .

The performances of recent neuroimaging methods using MRI and FDG-PET scans are presented in Tables 1 and 2. The following metrics are reported:

$$
\begin{aligned}
Accuracy\ (ACC) &= \tfrac{TP+TN}{TP+TN+FP+FN}, \\
Precision\ (PRE) &= \tfrac{TP}{TP+FP}, \\
Recall\ (REC) &= \tfrac{TP}{TP+FN}, \\
Specificity\ (SPE) &= \tfrac{TN}{FP+TN},
\end{aligned}
\tag{1}
$$

where TP (true positives) are the AD patients correctly classified as AD, TN (true negatives) are the normal control (NC) individuals correctly classified as NC, FP (false positives) are the NC incorrectly identified as AD and FN (false negatives) are the AD patients incorrectly identified as NC. The positive class represents the data from AD patients and the negative, the data from healthy individuals.

The simplest approach for the detection of AD consists in classifying the images directly based on their voxel values. Some works classify voxels from segmented tissues (ROIs) (Casanova et al. 2011; Klöppel et al. 2008; Rao et al. 2011) by logistic regression (LR) or SVM. Others train deep learning architectures with the whole brain image volume (Ambastha 2015; Gupta et al. 2013; Liu et al. 2015; Payan and Montana 2015). Klöppel et al. (2008) achieved an accuracy rate of 96.4% when classifying images of the gray matter tissue by SVM, but they used images from only 68
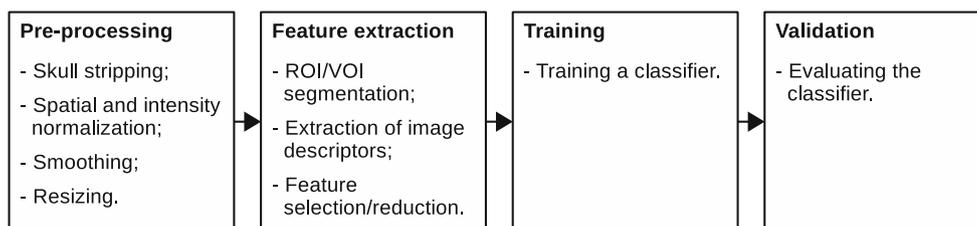

Fig. 1 Processes of neuroimaging methods for computer-aided diagnosis

Table 1 Classification performances of methods using MRI scans

Method	Subjects	ACC (%)	PRE (%)	REC (%)	SPE (%)
3D ConvNet + AdaBoost (Ambastha 2015)	50 AD, 50 NC	81.8	–	–	–
3D ConvNet with SAE (Ambastha 2015)	50 AD, 50 NC	81.8	–	–	–
White and gray matter + PLR (Casanova et al. 2011)	49 AD, 49 NC	85.7 ± 1.0	–	82.9 ± 1.9	90.0 ± 1.5
VOI + SVM (Chincarini et al. 2011)	144 AD, 189 NC	–	–	89.0	94.0
ConvNet 2D with SAE (Gupta et al. 2013)	200 AD, 232 NC	94.7	–	95.2	94.3
PCA + SVM (Khedher et al. 2015)	188 AD, 229 NC	87.8	–	90.0	85.1
Gray matter + SVM (Klöppel et al. 2008)	34 AD, 34 NC	96.4	–	100.0	92.9
Biomarkers + SVM (Liu et al. 2014)	198 AD, 229 NC	90.2 ± 0.5	–	85.3 ± 0.6	94.3 ± 0.4
Stacked AE (Liu et al. 2015)	85 AD, 77 NC	82.6 ± 5.3	84.3 ± 7.4	86.8 ± 6.8	77.8 ± 10.8
3D ConvNet with SAE (Payan and Montana 2015)	200 AD, 232 NC	95.4	–	–	–
Gray matter + SLR (Rao et al. 2011)	69 AD, 60 NC	85.3 ± 1.4	–	90.8 ± 3.7	80.3 ± 3.9
ICA + SVM (Yang et al. 2011)	202 AD, 236 NC	88.9	–	80.9	95.8

subjects. Casanova et al. (2011) tested the performance of a penalized logistic regression (PLR) to classify 98 subjects, resulting in an accuracy rate of 85.7%. In Rao et al. (2011), a sparse logistic regression (SLR) was trained, incorporating a sparsity penalty into the log-likelihood function, such that 85.3% of the images from 129 subjects were correctly classified.

Recent works (Ambastha 2015; Gupta et al. 2013; Liu et al. 2015; Payan and Montana 2015) adopt deep learning techniques for feature extraction and classification using images from the ADNI databases (Jack et al. 2008). Ambastha (2015) proposed a 3D convolutional neural network (ConvNet), reporting an accuracy rate of 81.8% in the classification of 100 individuals. In Payan and Montana (2015), the authors trained a 3D ConvNet with Sparse Autoencoders (SAE) and reported a high accuracy rate of 95.4% on images of 432 individuals. However, according to Ambastha (2015), the experiments in Payan and Montana (2015) indicate bias in the data. Liu et al. (2015) evaluated a multimodal approach, training Stacked Autoencoders (Stacked AE) with FDG-PET and MRI scans from 162 subjects. They reported a high accuracy rate of 91.4%. Gupta et al. (2013) trained a 2D ConvNet with SAE to extract features from MRI slices, achieving an accuracy rate of 94.7% on images of 432 individuals.

We may also say that the most promising approaches apply feature selection methods to discover biomarkers and remove non-informative features. Liu et al. (2014) selected the most relevant features using a tree construction method based on hierarchical clustering by taking into account spatial adjacency and feature similarity and discriminability. They achieved an accuracy rate of 90.2% for classifying 198 AD patients and 229 NC. In Garali et al. (2016), the brain was segmented into 116 regions to create a ranking of the most discriminative regions. Features were selected from 29 ROIs using the SelectKBest method (Kramer 2016), achieving an accuracy rate of 94.4% in the classification of 142 subjects. In Chincarini et al. (2011), different features were extracted from 9 VOIs to classify 144 AD and 189 NC. The features with the highest relative importance values, given by a Random Forest classifier (Breiman 2001), were selected to train an SVM classifier. This approach was able to discriminate NC from AD individuals with 89% of sensitivity and 94% of specificity.

Other approaches use Principal Component Analysis (PCA) and Independent Component Analysis (ICA) to reduce the feature space. Khedher et al. (2015) applied PCA to extract features from segmented MRI scans, achieving an accuracy rate of 87.7% on images from 417 individuals. Approaches with ICA achieved accuracies of

Table 2 Classification performances of methods using FDG-PET scans

Method	Subjects	ACC (%)	REC (%)	SPE (%)
ROI + SVM (Garali et al. 2016)	61 AD, 81 NC	94.4	–	–
ICA + SVM (Illán et al. 2011)	95 AD, 97 NC	87.0	87.8	86.4
PCA + SVM (Illán et al. 2011)	95 AD, 97 NC	88.2	87.7	88.6
ICA + SVM (Wenlu et al. 2011)	80 AD, 80 NC	86.8	73.9	99.7

Table 3 Demographic information of each dataset

Dataset	Group	Subjects	Female/Male	Age	No. of images
MRI	Normal	222	109/113	77.2 ± 5.2	865
	AD	181	85/96	76.4 ± 7.5	509
FDG-PET	Normal	93	35/58	77.5 ± 5.0	304
	AD	79	33/46	76.9 ± 6.8	203

88.9% (Yang et al. 2011) and 86.8% (Wenlu et al. 2011) in the classification of 438 and 160 subjects, respectively. In Illán et al. (2011), the experiments were performed by applying ICA and PCA to images from 192 subjects. The best accuracy was 88.24%, as obtained by PCA with SVM.

Therefore, given the differences in databases, classifiers, and validation protocols, it is impossible to indirectly compare the image descriptors or methods (descriptor and classifier) proposed in the aforementioned works. We have then selected some of them for a fair comparative analysis based on 10-fold cross-validation and a considerable number of FDG-PET and MRI scans.

Materials

ADNI Database

The data used in the preparation of this article were obtained from the ADNI databases (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 has been to test
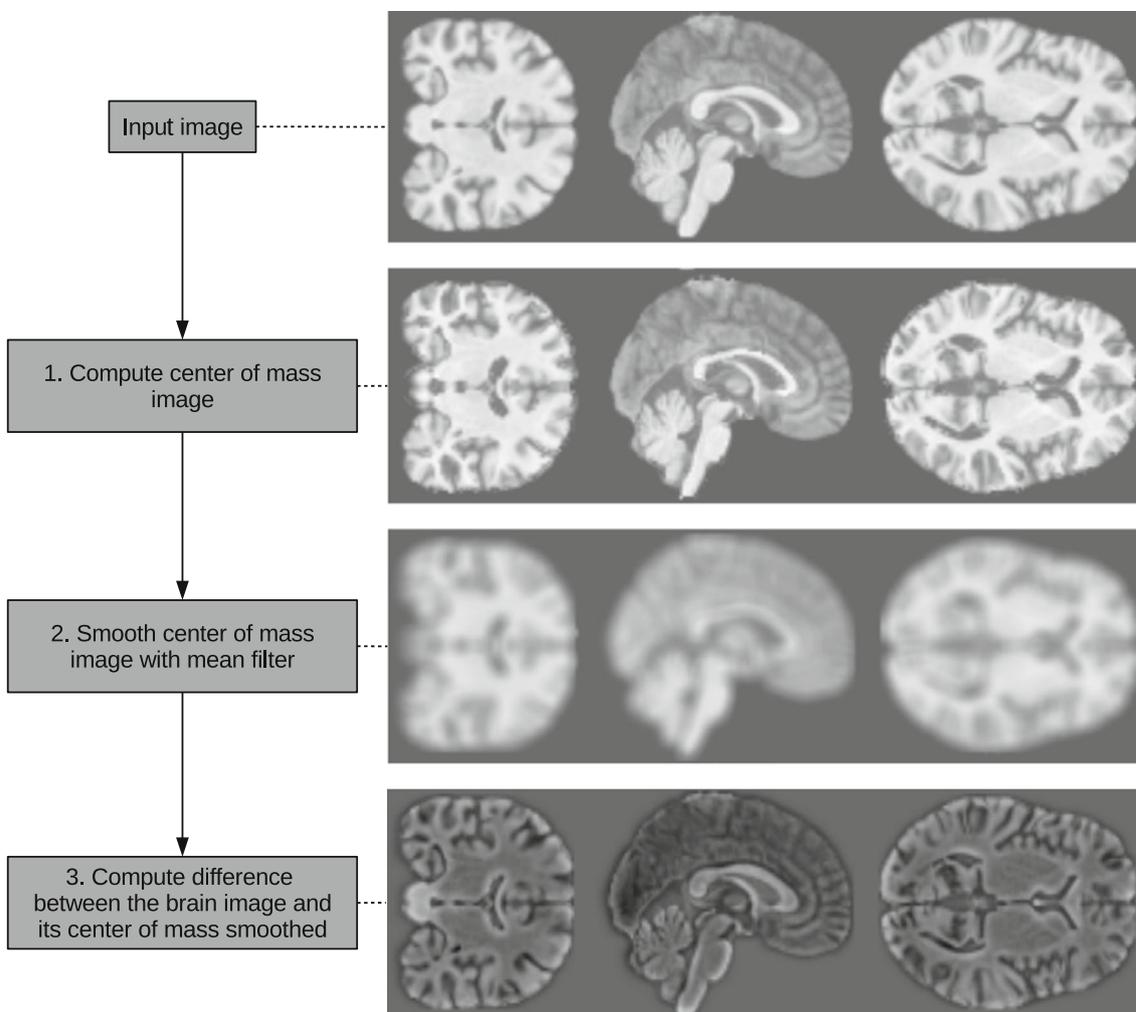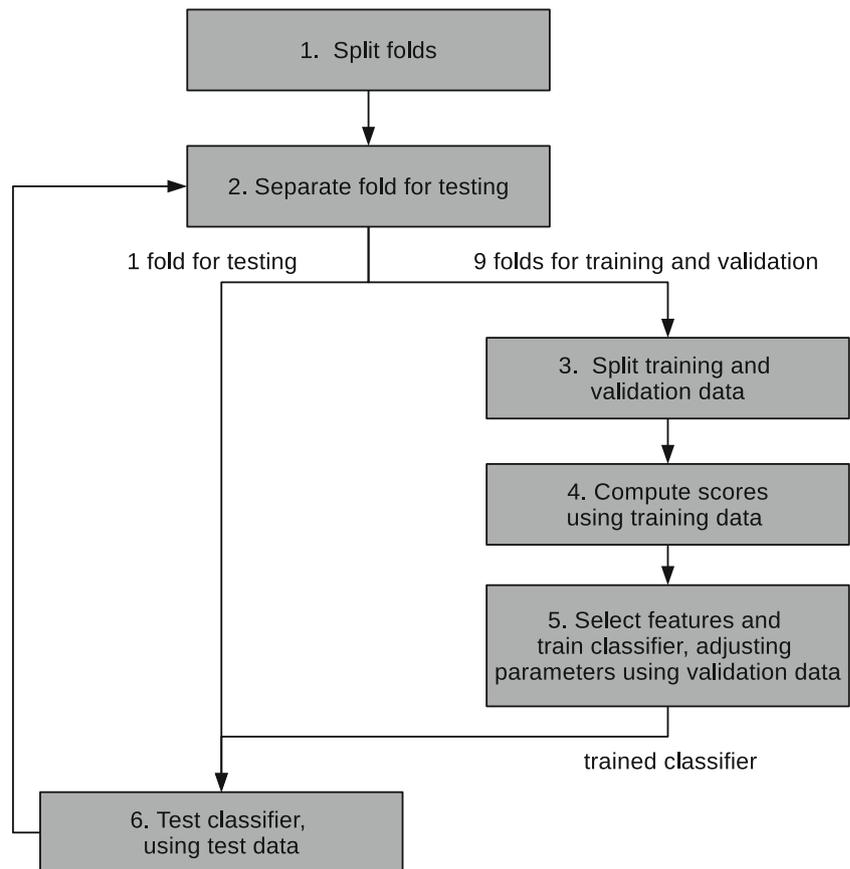

Fig. 2 Flowchart to compute the RCM features. The images on the right are the inputs and outputs of each feature extraction step

Fig. 3 Schematic representation
of the validation method

whether serial MRI, FDG-PET, other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (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 reduce the time and cost of clinical trials.

The Principal Investigator of this initiative is Michael W. Weiner, MD, VA Medical Center, 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, but ADNI has been followed by ADNI-GO and ADNI-2. To date, these three protocols have recruited over 1,500 adults of age from 55 to 90, consisting of cognitively normal older individuals, people with early or late MCI, and people 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.
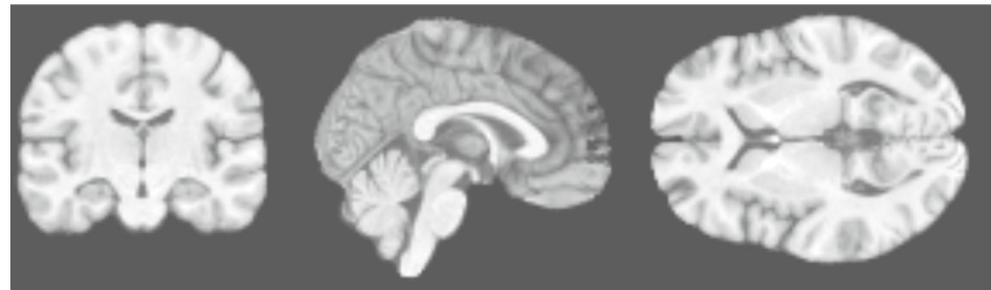
Our experiments were performed using scans acquired over a 2 or 3-year period in two imaging modalities: FDG-PET and T1-weighted MRI. The group and demographic information about the data are summarized in Table 3. The data include 507 FDG-PET scans from 172 individuals and 1374 MRI scans from 403 individuals.
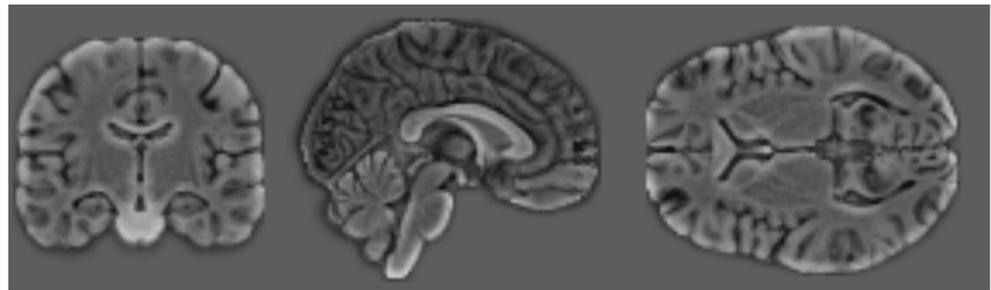
Preprocessing

Before extracting the features, the images are preprocessed through 4 steps: skull-stripping, spatial normalization, min-max normalization, and image downsampling.

Skull-stripping: firstly, the skull is stripped off using the Brain Extraction Tool version 2 (BET2) from the FMRIB Software Library 4.1 (FSL 4.1) (Jenkinson et al. 2005). This step removes the skull and the muscle tissue of the head. The skull-stripping algorithm creates a three-dimensional mesh with a spherical shape positioned at the center of gravity of the head. The mesh is iteratively expanded, adjusting its vertices to the borders detected between the brain tissues and the skull. Then, the brain is separated from the skull by applying the mask defined by this mesh. In the FDG-PET images, the skull region is not well-defined. Thus the skull-stripping was not performed
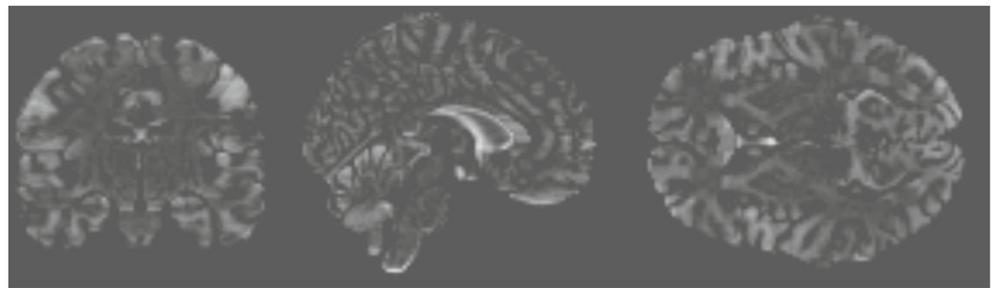
Fig. 4 Preprocessed image and features extracted with filter size 5
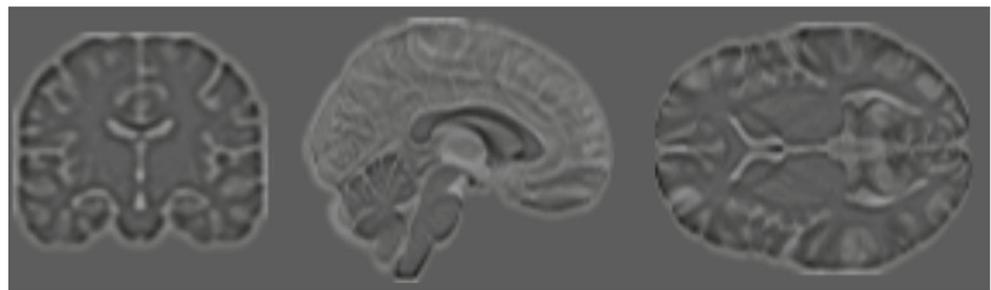

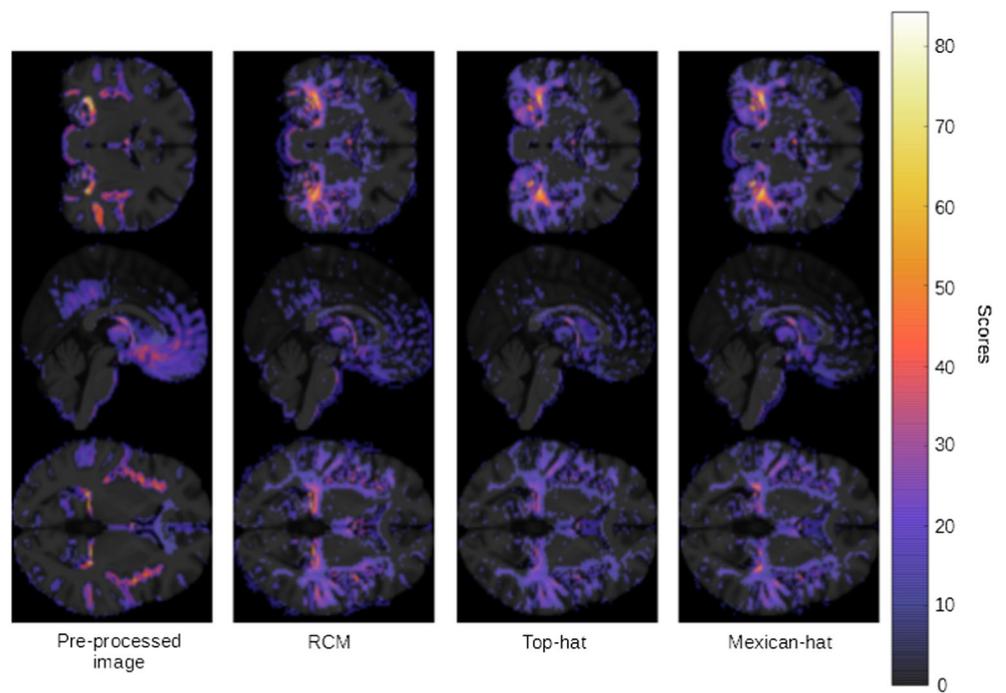(a) Preprocessed image

(b) RCM

(c) Top-hat

(d) Mexican-hat

and the images were only smoothed with a 12mm FWHM (Full width at half maximum) Gaussian filter (Landini et al. 2005).
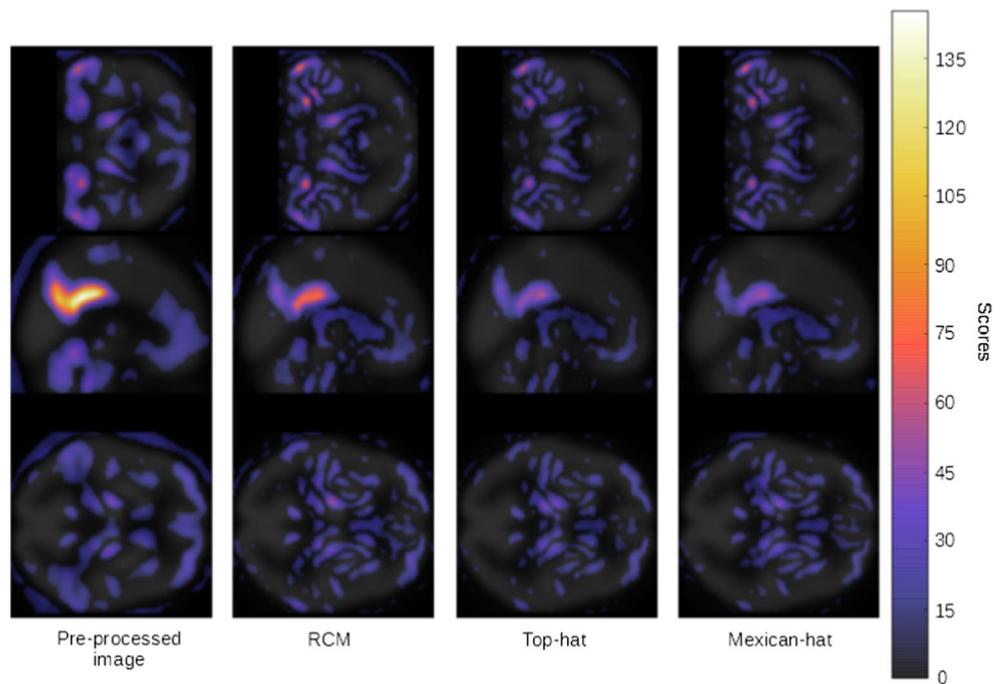
Spatial normalization: the images are spatially normalized onto the MNI (Montreal Neurological Institute) reference space (Fonov et al. 2011). This step reduces the brain anatomy variability among individuals, warping the images into a standard coordinate space. The MRI images are normalized using the symmetric diffeomorphic registration (SyN) (Avants et al. 2008). In an evaluation of 14 brain registration methods, the SyN was the algorithm that presented the best results according to overlap and distance measures (Klein et al. 2009). The FDG-PET scans are registered using the Statistical Parametric Mapping 5 (SPM 5) toolbox (Eickhoff et al. 2005) configured with its default parameters.

Fig. 5 Scores computed using preprocessed images and their features

(a) MRI scores

(b) PET scores

Min-max normalization: the voxel values are mapped in the range [0, 1], calculated by:

$$I_{norm}(\mathbf{p}) = \frac{I(\mathbf{p}) - I_{min}}{I_{max} - I_{min}}. \tag{2}$$

Each voxel $I(\mathbf{p})$ at its position $\mathbf{p} = (p_1, p_2, p_3)$ is subtracted to the minimum value I_{min} of all voxels. Then,

they are divided by the difference between the maximum I_{max} and the minimum I_{min} of their original values.

Image downsampling: After registration, each image is downsampled to the dimension $74 \times 92 \times 78$ with a voxel size of $2 \times 2 \times 2$ mm^3, which results in 531 024 voxels. This process helps to eliminate noise and compensate imprecisions in the registration. Also, the computational

time and memory requirements are reduced without losing discriminant information (Segovia et al. 2010, 2012).

Methods

Image Description

When analyzing biomarkers from other studies (Chincarini et al. 2011; Liu et al. 2014), we can observe that the brain boundaries concentrate the most discriminative voxels for the detection of AD. Thus, we explore in this work image operations to highlight theses areas. We extract three types of features for evaluation: top-hat (Heijmans and Roerdink 1998), Mexican-hat (Russ 2016) and RCM. The first two were already used in neuroimaging applications (Sensi et al. 2014; Somasundaram and Genish 2014). The last one, RCM, is proposed in this work.

In computer vision, image moments have been extensively used for pattern recognition (Chaumette 2004). In this work, we use image moments to extract a feature designed to highlight the brain boundaries. The process to extract the RCM features has three steps as presented in Fig. 2.

In the first step, we compute the center of mass locally in regions of the image using three-dimensional moments. For a continuous function $f(p_1, p_2, p_3)$, we can define the moment of order $(q + r + s)$ as:

$$M_{qrs} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} p_1^q p_2^r p_3^s f(p_1, p_2, p_3) dp_1 dp_2 dp_3 \quad (3)$$

where $q, r, s = 0, 1, 2, ..., \infty$. Adapting (3) to volumetric images with voxel intensities $I(p_1, p_2, p_3)$, image moments M_{ijk} are calculated by:

$$M_{ijk} = \sum_{p_1, p_2, p_3} p_1^i p_2^j p_3^k I(p_1, p_2, p_3) \quad (4)$$

where $i, j, k \in \mathbb{N}$ and $(i + j + k)$ is the order of the moment.

A set of moments up to order n consists of all moments $M_{i,j,k}$ such that $0 \leq i + j + k \leq n$. Moments of low orders provide geometric properties of the image. The zero-order moment M_{000} defines the total of mass of the image I. The ratios between the first-order moments and the zero-order moments $(M_{100}/M_{000}, M_{010}/M_{000}, M_{001}/M_{000})$ define the center of mass or centroid of the image.

To extract the RCM features, an image I_C is created with all voxel values of the centroids calculated locally in regions of the input image. We define a region R at each voxel $\mathbf{p}$ of the input image and compute the centroid $C_\mathbf{p}$ of R as follows:

$$C_\mathbf{p} = \frac{\sum_{\mathbf{p} \in R} \mathbf{p} I(\mathbf{p})}{\sum_{\mathbf{p} \in R} I(\mathbf{p})}, \quad (5)$$

where $I(\mathbf{p})$ is the voxel value at the position $\mathbf{p}$. The region R at a position $\mathbf{p} = (p_1, p_2, p_3)$ is defined by all valid voxels between $(p_1 - s/2, p_2 - s/2, p_3 - s/2)$ and $(p_1 + s/2, p_2 + s/2, p_3 + s/2)$, where $s \times s \times s$ is the region's size.

In the second step, the center of mass image I_C is smoothed by a mean filter, creating a smoothed image I_C'. Lastly, we subtract I_C' from the input image of the brain to output the RCM features. Figure 2 presents the resulting images from the RCM procedure. The first image is the input image used to extract the features. The second is the center of mass image I_C. The third image is I_C' and the voxel intensities of the last image represent the RCM features.

Feature Selection

One of the main challenges in working with neuroimages is the curse of dimensionality. In neuroimaging studies, there are generally hundreds of images to be analyzed for thousands of features. All that can easily lead to classifiers that overfit the data. Some studies avoid this problem by introducing regularization parameters to train sparse models (Rao et al. 2011). Others use feature selection techniques to identify biomarkers (Breiman 2001; Chincarini et al. 2011; Garali et al. 2016; Kramer 2016; Liu et al. 2014).

A popular technique used in machine learning applications is the feature selection by ANOVA (Analysis Of Variance) (Chen and Lin 2006; Costafreda et al. 2009; Costafreda et al. 2011; Elssied et al. 2014; Garali et al. 2016; Golugula et al. 2011; Grünauer and Vincze 2015). It consists of selecting the most relevant features for classification by calculating scores with the ANOVA F-test. The features with the lowest F-values are filtered out and the features with the highest F-values are maintained as the final image descriptor to train and test the classifier. We can define the F-value as:

$$F = \frac{S_B^2}{S_W^2}. \quad (6)$$

S_B^2 is the between-group variability (sets of samples per class), given by:

$$S_B^2 = \frac{\sum_i n_i (\bar{x}_i - \bar{x})^2}{K - 1}, \quad (7)$$

where n_i and $\bar{x}_i$ denote the number of observations and the sample mean in the ith group, respectively. $\bar{x}$ is the overall mean of the data and K denotes the number of groups (or classes).

S_W^2 is the within-group variability, defined as:

$$S_W^2 = \frac{\sum_{ij} (x_{ij} - \bar{x}_i)^2}{N - K}, \quad (8)$$

where x_{ij} is the jth observation in the ith group and N is the overall sample size.

Studies have shown that ANOVA is an efficient method to discover biomarkers (Costafreda et al. 2009, 2011; Garali et al. 2016). However, it is very sensitive to outliers (French et al. 2017; Halldestam 2016). Noisy data and errors in the spatial normalization can hamper the effectiveness of the scores. To avoid that, the F-values are computed in an iterative process by random sampling the data. The pseudo-code of the method is summarized as follows:

Algorithm 1 Stable feature selection by ANOVA

$scores \leftarrow initializeScoresWithZeros()$
while stopping criteria is false **do**
  $X' \leftarrow SelectRandomSubset(X)$
  $fValues \leftarrow ComputeFValues(X')$
  $scores \leftarrow max(scores, fValues)$
end while
$relevantPositions \leftarrow IdentifyRelevantPositions(X, scores, N)$
$X_{selected} \leftarrow SelectFeatures(X, relevantPositions)$

First, a score vector ($scores$) is initialized with zeros. Each voxel value (or feature) is associated with a score value. And X represents the set of vectors composed by the image voxels of each sample. We select a random subset of samples X' from X and compute the F-values ($fValues$) for each feature in X'. Then the scores of each feature are updated with the maximum values of the score values ($scores$) and the F-values ($fValues$). The random selection and the score update are executed until a stopping criteria is reached. We may stop the execution when a maximum iteration number is achieved or the scores converges to some value. The positions of the relevant features for classification are identified by selecting the regions associated with the N highest scores. Lastly, we select the N features in the corresponding relevant positions.

Results and Discussion

Validation Method

The experiments adopted the 10-fold cross-validation method, presented in the flowchart of Fig. 3. First, the data are split in ten folds, assigning the images of each subject to different folds to avoid biased results. Then, ten classification tests are performed. In each test, one fold is selected to be the test set and the others are split in training and validation sets. After splitting the data, we have about 10% of the data for testing, 85% for training and 5% for validation.

From the training set, we compute the scores for the feature selection. Then, the features are selected and the classifier is trained, adjusting its parameters and the number

Table 4 Rank of the regions with highest scores in MRI scans

sub_cortical_gray_nuclei_-_amygdala_l
limbic_lobe_-_hippocampus_l
limbic_lobe_-_hippocampus_r
sub_cortical_gray_nuclei_-_amygdala_r
limbic_lobe_-_parahippocampal_gyrus_l
limbic_lobe_-_parahippocampal_gyrus_r
temporal_lobe_-_lateral_surface_-_heschl_gyrus_l
insula_l
limbic_lobe_-_temporal_pole_superior_temporal_gyrus_l
temporal_lobe_-_lateral_surface_-_heschl_gyrus_r

of features selected using the validation data. At last, the classifier is tested using the test data and the performance metrics are reported.
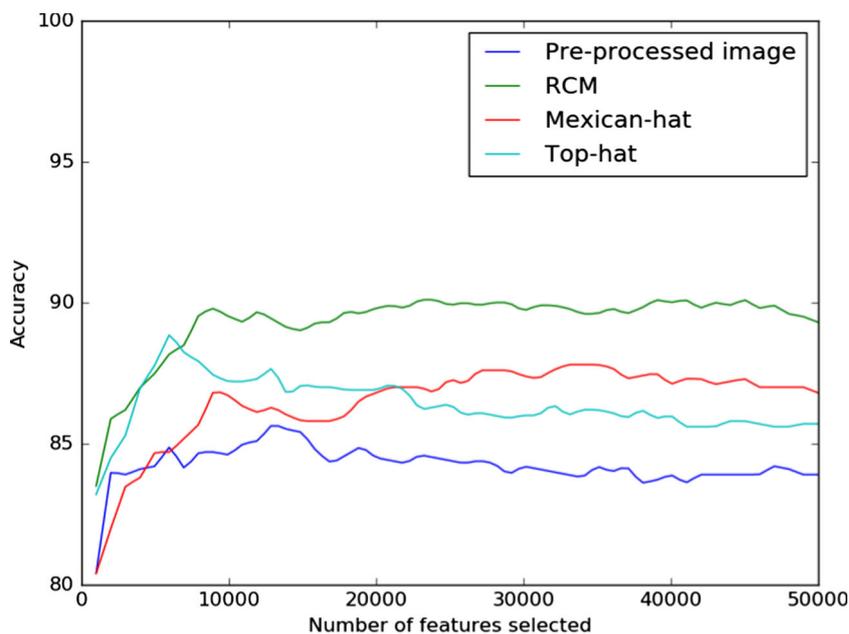
Experiments and Results

In the experiments of this work, we first analyze the classification performances of the preprocessed images and three descriptors: RCM and the preprocessed input image filtered by mexican-hat and by top-hat. Different patterns are generated with different filter scales. Thus, four filter sizes are used to extract the features: 3, 5, 7 and 9. The descriptors for classification are represented by vectors concatenating the features of each filter size, as selected with the scores computed by ANOVA. Examples of features extracted with filter size equal to 5 are shown in Fig. 4. All images are normalized to the range 0 to 255 of grayscale values for visualization. The images Fig. 4b, c and d are the RCM, top-hat and mexican-hat features extracted from the preprocessed image Fig. 4a, respectively.

In feature selection, the scores are computed separately for each type of feature and filter size using 100 iterations. Figure 5 presents the scores computed with the preprocessed images and their features. The filters applied in the images
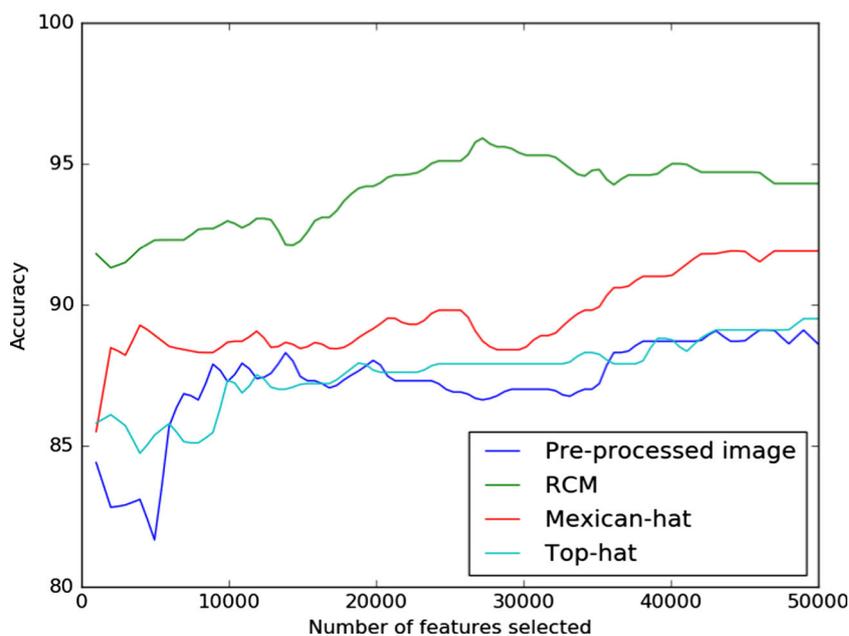
Table 5 Rank of the regions with highest scores in FDG-PET scans

limbic_lobe_-_posterior_cingulate_gyrus_l
limbic_lobe_-_posterior_cingulate_gyrus_r
parietal_lobe_-_lateral_surface_-_angular_gyrus_l
limbic_lobe_-_hippocampus_l
limbic_lobe_-_hippocampus_r
parietal_lobe_-_lateral_surface_-_angular_gyrus_r
sub_cortical_gray_nuclei_-_amygdala_l
limbic_lobe_-_parahippocampal_gyrus_l
sub_cortical_gray_nuclei_-_thalamus_l
parietal_lobe_-_medial_surface_-_precuneus_r

Fig. 6 Classification accuracy when the number of selected features varies

(a) Classification accuracy obtained from MRI scans

(b) Classification accuracy obtained from PET scans

help to select discriminant features, highlighting the scores of relevant voxels near the boundary. We ranked the brain regions by their scores and the results are consistent with the literature (Ambastha 2015; Garali et al. 2016). The brain is segmented in 116 anatomical regions using masks extracted from the MARINA software (Walter et al. 2003). The 10 regions with highest mean scores in MRI and FDG-PET scans are presented in the Tables 4 and 5. The MRI features

with the highest scores are located in the hippocampi, parahippocampi, and amygdalae. In the FDG-PET scans, the top ranked regions are the posterior cingulate gyrus, angular gyrus and hippocampi.

To calculate the scores, random subsets are selected choosing randomly images from 95% of the subjects from the training data. At each iteration, only one image is chosen by subject to keep the number of samples fixed for

Table 6 Classification performances of SVM with each descriptor

Dataset	Type of feature	ACC (%)	PRE (%)	REC (%)	SPE (%)	κ
MRI	Preprocessed image	83.8 ± 6.2	79.2 ± 10.8	78.1 ± 10.0	87.1 ± 7.9	0.65 ± 0.12
	Top-hat	86.2 ± 5.7	83.0 ± 9.8	80.5 ± 7.2	89.9 ± 6.5	0.73 ± 0.1
	Mexican-hat	87.5 ± 4.7	84.0 ± 9.1	81.7 ± 7.2	90.9 ± 4.9	0.7 ± 0.11
	RCM	**90.3 ± 4.7**	**88.5 ± 9.0**	**85.5 ± 6.2**	**93.3 ± 5.1**	**0.79 ± 0.1**
FDG-PET	Preprocessed image	86.7 ± 10.9	83.1 ± 15.3	91.2 ± 16.2	84.4 ± 17.5	0.72 ± 0.23
	Top-hat	87.9 ± 11.5	86.1 ± 15.5	86.1 ± 12.7	89.0 ± 14.2	0.77 ± 0.25
	Mexican-hat	89.3 ± 12.3	87.1 ± 15.5	86.5 ± 16.1	91.2 ± 11.9	0.75 ± 0.24
	RCM	**95.1 ± 6.0**	**93.8 ± 7.4**	**95.9 ± 8.7**	**94.6 ± 7.8**	**0.9 ± 0.12**

the calculation of the F-values. The number of features is selected by the value that achieved the highest accuracy on the validation set. Experiments were performed by training linear SVM classifiers with penalty parameter C fixed at value 1. Other values of C did not cause variations in classification performances.

Figure 6 shows the average accuracy on the test set with respect to different number of selected features. For each neuroimaging modality, we evaluated the classification performance of SVM with each descriptor: the preprocessed input image and the filtered images. The highest classification performance of SVM was obtained with the RCM descriptor. Correct classification rates of 90.3% e 95.1% were achieved for the MRI and FDG-PET modalities, respectively. Table 6 presents the results for each image descriptor. The highest means of the performance rates are emphasized in bold.

Figure 7 shows the Receiver Operating Characteristic (ROC) (Hanley and McNeil 1982) plots of the RCM descriptor. Analyzing the ROC curves, we can observe variations in the classification performances across the folds due to the diversity of individuals. This shows that there are some subjects more difficult to classify than others. Thus, our choice for 10-fold cross-validation is justified to obtain more reliable results.

The results reported in the literature are difficult to be compared given the differences in the validation methods, databases and amount of data used. In addition, problems such as the presence of images of the same individual in both training and test sets, or the lack of consideration to the variability among individuals in the validation method, may compromise the accuracy of the results. Therefore, for a more reliable comparison with other methods, the performances of different techniques were evaluated using k-fold cross-validation with $k = 10$ without overlapping images of the same individual in the training and test sets. The same images and folds were used to evaluate these methods. The results are reported in Table 7. The highest means of the performance rates are emphasized in bold.
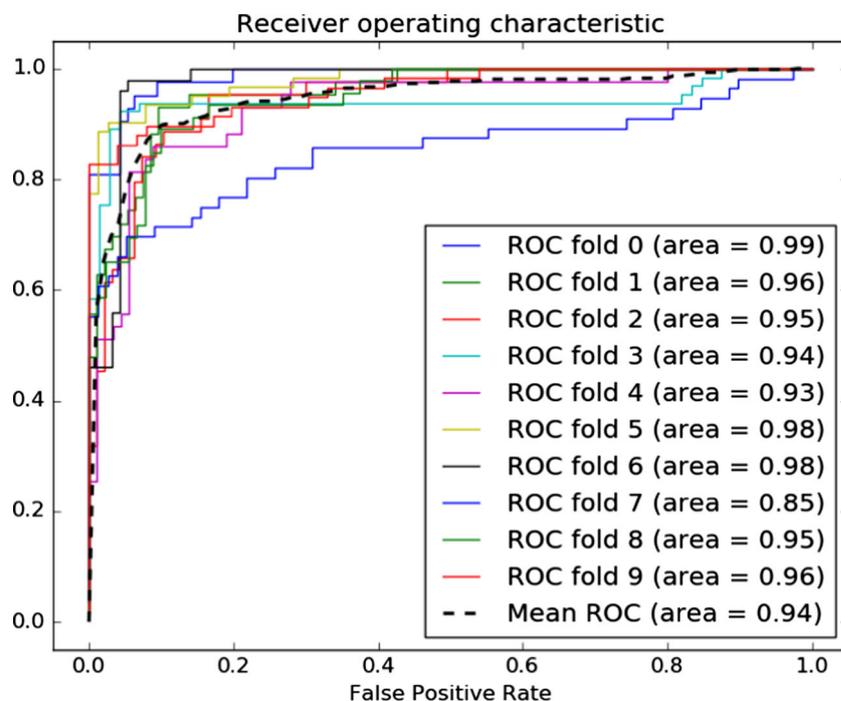
The classification performances of LR with the preprocessed images, feature-space reduction, and Deep Learning techniques were also evaluated. PLR models were trained with regularization parameter $\lambda = 0.5$, achieving accuracies of 82.2% and 91.6% on MRI and FDG-PET images. SLR with regularization and sparsity parameters, $\lambda = 0.5$ and $\beta = 0.5$, achieved correction classification rates of 81.0% and 91.6% in the classification of MRI and FDG-PET images, respectively.

Feature-space reduction techniques also resulted in high accuracy of classification. The space dimension of the images was reduced by PCA and ICA, followed by SVM classification. The best results were obtained with ICA, presenting correct classification rates of 82.3% and 92.1% on MRI and FDG-PET images, respectively. The PCA method resulted in a accuracy of 81.9% and 89.8% on MRI and FDG-PET images. The number of components chosen for the feature-space reduction was determined by the value that reached the highest accuracy in the classification of the validation data.
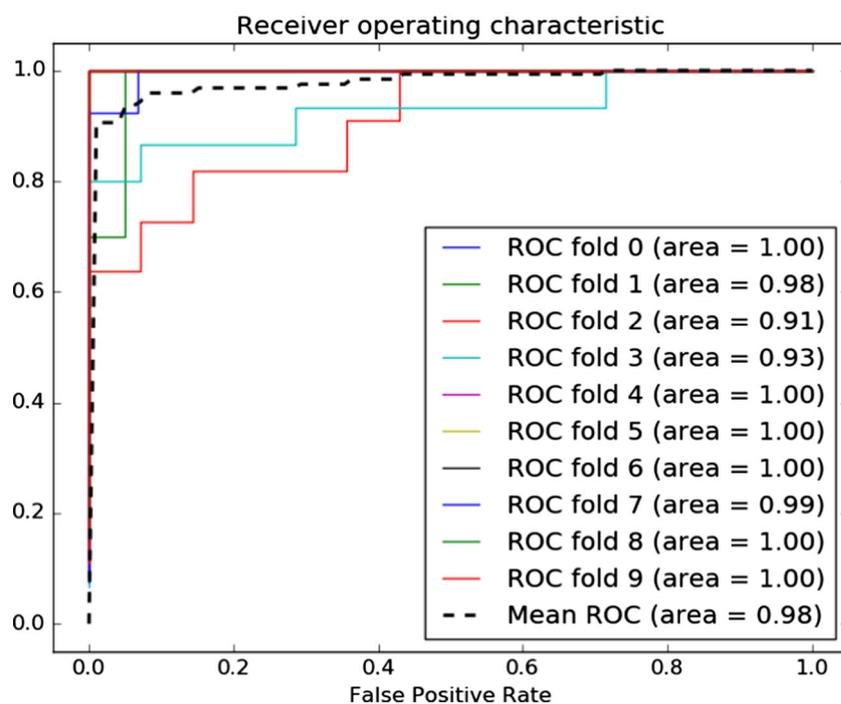
Deep learning techniques have proved to be effective for image representation and classification. For neuroimaging applications, different architectures have been exploited (Ambastha 2015; Gupta et al. 2013; Liu et al. 2015; Payan and Montana 2015). Most of the works train their models with Autoencoders (Gupta et al. 2013; Liu et al. 2015; Payan and Montana 2015) due to the high dimensionality of the data. Some of these works, however, do not split the training and the test sets by subjects. This implies in a high risk of overfitting, because some databases can have multiple images of a same subject. Thus, we evaluated in this work the performances of two ConvNet architectures (Gupta et al. 2013; Payan and Montana 2015) trained with SAE.

The 2D architecture of Gupta et al. (2013) resulted in low accuracies below 60%. The large pooling causes the loss of discriminant information, affecting the classification performances. The 43 084 800 features extracted by the convolution layer are reduced to 61 200. Generally, ConvNets use pooling sizes of 3×3 or 5×5, preserving

Fig. 7 ROC curves of experiments performed with RCM features

(a) ROC of experiments using MRI scans

(b) ROC of experiments using PET scans

discriminant information. But in this architecture, the size used was 24x30.

Payan and Montana (2015) trained a 3D ConvNet with the same data used in Gupta et al. (2013), reporting a greater accuracy. Our experiments with this architecture

have reached accuracies of 82.3% and 87.1% on MRI and FDG-PET images.

Considering the randomness on the data that can affect the classification performances, we also reported the Cohen's kappa coefficient (Association et al. 1999)

Table 7 Classification performances obtained from different techniques using 10-fold cross-validation

Dataset	Method	ACC (%)	PRE (%)	REC (%)	SPE (%)	κ
MRI	PLR	82.2 ± 6.9	72.3 ± 10.8	**87.3 ± 5.6**	79.4 ± 10.4	0.64 ± 0.13
	SLR	81.0 ± 7.0	70.5 ± 10.9	86.7 ± 5.5	77.6 ± 10.9	0.61 ± 0.13
	PCA + SVM	81.9 ± 6.4	78.5 ± 11.7	73.2 ± 7.2	87.4 ± 8.0	0.61 ± 0.13
	ICA + SVM	82.3 ± 4.2	79.5 ± 11.2	71.9 ± 8.9	88.8 ± 7.0	0.61 ± 0.09
	2D ConvNet with SAE	59.5 ± 6.2	52.8 ± 25.3	25.8 ± 24.1	81.6 ± 19.2	0.03 ± 0.05
	3D ConvNet with SAE	82.3 ± 5.0	78.7 ± 13.0	75.1 ± 8.0	86.9 ± 8.9	0.62 ± 0.1
	RCM + SVM	**90.3 ± 4.7**	**88.5 ± 9.0**	85.5 ± 6.2	**93.3 ± 5.1**	**0.79 ± 0.1**
FDG-PET	PLR	91.6 ± 9.0	88.5 ± 11.8	92.1 ± 13.0	91.2 ± 9.4	0.82 ± 0.19
	SLR	91.6 ± 8.9	89.0 ± 10.7	91.5 ± 13.1	91.7 ± 9.0	0.82 ± 0.19
	PCA + SVM	89.8 ± 7.6	90.0 ± 11.2	86.0 ± 14.4	92.5 ± 9.1	0.78 ± 0.16
	ICA + SVM	92.1 ± 7.7	92.4 ± 8.8	88.2 ± 14.8	**95.9 ± 4.3**	0.83 ± 0.17
	2D ConvNet with SAE	53.1 ± 9.9	46.0 ± 34.6	12.2 ± 7.8	87.2 ± 10.5	−0.07 ± 0.13
	3D ConvNet with SAE	87.1 ± 7.3	89.8 ± 9.5	78.6 ± 14.5	93.1 ± 7.3	0.72 ± 0.15
	RCM + SVM	**95.1 ± 6.0**	**93.8 ± 7.4**	**95.9 ± 8.7**	94.6 ± 7.8	**0.9 ± 0.12**

obtained in each experiment. Different from the accuracy, Cohen's Kappa calculates the correct classification rates independently for each class and then aggregates them into a single value. This metric is less sensitive to randomness caused by variations in the numbers of observations of each class. The coefficient κ can be defined by:

$$\kappa = 1 - \frac{1 - p_o}{1 - p_e}, \qquad (9)$$

where p_o is the relative agreement rate observed between the real value and the estimated value by the classifier, being equal to the value of the accuracy:

$$p_o = \frac{TP + TN}{TP + TN + FP + FN}. \qquad (10)$$

The term p_e is the hypothetical probability of chance agreement, calculated by:

$$p_e = \frac{(TP+FN)(TP+FP)+(FP+TN)(FN+TN)}{(TP+TN+FP+FN)^2}. \qquad (11)$$

According to Landis and Koch (1977), the κ value can be interpreted based on the intervals shown in Table 8.

Table 8 Interpretation of κ based on intervals of values (Landis and Koch 1977)

κ values	Interpretation
< 0	No agreement
0 − 0.19	Slight agreement
0.20 − 0.39	Fair agreement
0.40 − 0.59	Moderate agreement
0.60 − 0.79	Substantial agreement
0.80 − 0.99	Almost perfect agreement

The κ values obtained for each classification method are presented in Table 7. Analyzing κ, the performances with RCM indicate substantial agreement and almost perfect agreement with the clinical information of the MRI and FDG-PET scans, respectively.

Conclusion

In this paper, we analyzed the performances of different methods for the discovery of image biomarkers associated with AD. Experiments were performed with large and public datasets of MRI and FDG-PET scans. The results showed that image filtering techniques could be useful to improve the classification performances. We also proposed the RCM descriptor that extracts features from the brain boundaries. Our method was able to find relevant biomarkers not requiring prior knowledge of ROIs/VOIs. Also, the whole brain with gray matter and white matter tissues was used for feature extraction. Thus, tissue segmentation was not necessary as in other approaches. In comparison with other methods, the classification with the RCM descriptor obtained high performances with low variances.

AD severely damages the hippocampal region that begins to be affected from the earliest stages of the disease, even before it impairs the patient's cognitive ability. In this work, it is noted that this region is very important for the diagnosis as also reported in medical studies. The results indicate that relevant regions can be automatically found by computers and are useful for supporting the clinical diagnosis.

PET images indicated a high potential to aid the diagnosis of AD. It is expected that this type of exam will be an important tool for diagnosis and prognosis.

Furthermore, with advances in treatment methods, imaging exams will be of great importance for the discovery and determination of the stages of AD. In future work, we intend to investigate multimodal approaches with FDG-PET and MRI, since the combination of different modalities improved the classification performances in other studies. RCM can also be investigated to classify images of patients with other stages of dementia, such as early and late MCI.

Information Sharing Statement

The implementation of the methods to extract features, pre-process and classify the images is available at https://github.com/alexandreyy/alzheimer_cad. These methods requires open source softwares to be executed: the FSL library (RRID:SCR_002823) https://fsl.fmrib.ox.ac.uk/fsl/fslwiki and the Advanced Normalization Tools (RRID:SCR_004757) http://stnava.github.io/ANTs/. The image dataset is provided by the ADNI (RRID:SCR_003007) at http://adni.loni.usc.edu/.

References

Alzheimer's Association (2017). Alzheimer's disease and dementia. http://www.alz.org/. [Online; accessed 20 Dec 2017].

Ambastha, A.K. (2015). Neuroanatomical characterisation of Alzheimer's disease using deep learning. National University of Singapore.

Association, A.E.R., Association, A.P., on Measurement in Education, N.C., on Standards for Educational, J.C., (US), P.T. (1999). Standards for educational and psychological testing. American Educational Research Association.

Avants, B.B., Epstein, C.L., Grossman, M., Gee, J.C. (2008). Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. *Medical Image Analysis*, *12*(1), 26–41.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Carmichael, O.T., Aizenstein, H.A., Davis, S.W., Becker, J.T., Thompson, P.M., Meltzer, C.C., Liu, Y. (2005). Atlas-based hippocampus segmentation in Alzheimer's disease and mild cognitive impairment. *NeuroImage*, *27*(4), 979–990.

Casanova, R., Whitlow, C.T., Wagner, B., Williamson, J., Shumaker, S.A., Maldjian, J.A., Espeland, M.A. (2011). High dimensional classification of structural MRI Alzheimer's disease data based on large scale regularization. *Frontiers in Neuroinformatics*, *5*, 22.

Chaumette, F. (2004). Image moments: a general and useful set of features for visual servoing. *IEEE Transactions on Robotics*, *20*(4), 713–723.

Chen, Y.W., & Lin, C.J. (2006). Combining SVMs with various feature selection strategies. In *Feature extraction* (pp. 315–324). Springer.

Chincarini, A., Bosco, P., Calvini, P., Gemme, G., Esposito, M., Olivieri, C., Rei, L., Squarcia, S., Rodriguez, G., Bellotti, R., et al (2011). Local MRI analysis approach in the diagnosis of early and prodromal Alzheimer's disease. *NeuroImage*, *58*(2), 469–480.

Costafreda, S.G., Chu, C., Ashburner, J., Fu, C.H. (2009). Prognostic and diagnostic potential of the structural neuroanatomy of depression. *PLoS one*, *4*(7), e6353.

Costafreda, S.G., Fu, C.H., Picchioni, M., Toulopoulou, T., McDonald, C., Kravariti, E., Walshe, M., Prata, D., Murray, R.M., McGuire, P.K. (2011). Pattern of neural responses to verbal fluency shows diagnostic specificity for schizophrenia and bipolar disorder. *BMC Psychiatry*, *11*(1), 1.

Eickhoff, S.B., Stephan, K.E., Mohlberg, H., Grefkes, C., Fink, G.R., Amunts, K., Zilles, K. (2005). A new SPM toolbox for combining probabilistic cytoarchitectonic maps and functional imaging data. *NeuroImage*, *25*(4), 1325–1335.

Elssied, N.O.F., Ibrahim, O., Osman, A.H. (2014). A novel feature selection based on one-way ANOVA f-test for e-mail spam classification. *Research Journal of Applied Sciences Engineering and Technology*, *7*(3), 625–638.

Fonov, V., Evans, A.C., Botteron, K., Almli, C.R., McKinstry, R.C., Collins, D.L. (2011). Brain development cooperative group, others: unbiased average age-appropriate atlases for pediatric studies. *NeuroImage*, *54*(1), 313–327.

French, A., Macedo, M., Poulsen, J., Waterson, T., Yu, A. (2017). Multivariate analysis of variance (MANOVA). http://userwww.sfsu.edu/efc/classes/biol710/manova/MANOVAnewest.pdf. [Online; accessed 20 Dec 2017].

Garali, I., Adel, M., Bourennane, S., Guedj, E. (2016). Brain region ranking for 18FDG-PET computer-aided diagnosis of Alzheimer's disease. *Biomedical Signal Processing and Control*, *27*, 15–23.

Golugula, A., Lee, G., Madabhushi, A. (2011). Evaluating feature selection strategies for high dimensional, small sample size datasets. In *2011 Annual International conference of the IEEE engineering in medicine and biology society* (pp. 949–952). IEEE.

Grünauer, A., & Vincze, M. (2015). Using dimension reduction to improve the classification of high-dimensional data. arXiv:1505.06907.

Gupta, A., Ayhan, M., Maida, A. (2013). Natural image bases to represent neuroimaging data. In *ICML* (Vol. 3, pp. 987–994).

Halldestam, M. (2016). ANOVA-the effect of outliers.

Hanley, J.A., & McNeil, B.J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, *143*(1), 29–36.

Heijmans, H.J., & Roerdink, J. (1998). *Mathematical morphology and its applications to image and signal processing* (Vol. 12). Springer Science & Business Media.

Illán, I., Górriz, J., Ramírez, J., Salas-Gonzalez, D., López, M., Segovia, F., Chaves, R., Gómez-Rio, M., Puntonet, C.G., ADNI, et al (2011). 18 F-FDG PET imaging analysis for computer aided Alzheimer's diagnosis. *Information Sciences*, *181*(4), 903–916.

Jack, C.R., Bernstein, M.A., Fox, N.C., Thompson, P., Alexander, G., Harvey, D., Borowski, B., Britson, P.J., L Whitwell, J., Ward, C., et al (2008). The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods. *Journal of Magnetic Resonance Imaging*, *27*(4), 685–691.

Jenkinson, M., Pechaud, M., Smith, S. (2005). BET2: MR-based estimation of brain, skull and scalp surfaces. In: *Eleventh annual meeting of the organization for human brain mapping* (Vol. 17, p. 167).

Khedher, L., Ramírez, J., Górriz, J.M., Brahim, A., Segovia, F., ADNI, et al. (2015). Early diagnosis of Alzheimer's disease based on partial least squares, principal component analysis and support vector machine using segmented MRI images. *Neurocomputing*, *151*, 139–150.

Klein, A., Andersson, J., Ardekani, B.A., Ashburner, J., Avants, B., Chiang, M.C., Christensen, G.E., Collins, D.L., Gee, J., Hellier, P., et al (2009). Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration. *NeuroImage*, *46*(3), 786–802.

Klöppel, S., Stonnington, C.M., Barnes, J., Chen, F., Chu, C., Good, C.D., Mader, I., Mitchell, L.A., Patel, A.C., Roberts, C.C., et al (2008). Accuracy of dementia diagnosis - a direct comparison between radiologists and a computerized method. *Brain: A Journal of Neurology*, *131*(11), 2969–2974.

Kramer, O. (2016). Scikit-learn. In *Machine learning for evolution strategies* (pp. 45–53). Springer.

Landini, L., Positano, V., Santarelli, M. (2005). *Advanced image processing in magnetic resonance imaging*. CRC Press.

Landis, J.R., & Koch, G.G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 159–174.

Liu, M., Zhang, D., Shen, D., ADNI, et al. (2014). Identifying informative imaging biomarkers via tree structured sparse learning for AD diagnosis. *Neuroinformatics*, *12*(3), 381–394.

Liu, S., Liu, S., Cai, W., Che, H., Pujol, S., Kikinis, R., Feng, D., Fulham, M.J., et al. (2015). Multimodal neuroimaging feature learning for multiclass diagnosis of Alzheimer's disease. *IEEE Transactions on Biomedical Engineering*, *62*(4), 1132–1140.

Payan, A., & Montana, G. (2015). Predicting Alzheimer's disease: a neuroimaging study with 3d convolutional neural networks. arXiv:1502.02506.

Rao, A., Lee, Y., Gass, A., Monsch, A. (2011). Classification of Alzheimer's disease from structural MRI using sparse logistic regression with optional spatial regularization. In *2011 Annual International conference of the IEEE engineering in medicine and biology society, EMBC* (pp. 4499–4502). IEEE.

Russ, J.C. (2016). *The image processing handbook*. CRC Press.

Segovia, F., Górriz, J., Ramírez, J., Salas-Gonzalez, D., Álvarez, I., López, M., Chaves, R., ADNI, et al. (2012). A comparative study of feature extraction methods for the diagnosis of Alzheimer's disease using the ADNI database. *Neurocomputing*, *75*(1), 64–71.

Segovia, F., Ramírez, J., Górriz, J.M., Chaves, R., Salas-Gonzalez, D., López, M., Álvarez, I., Padilla, P., Puntonet, C.G. (2010). Partial least squares for feature extraction of SPECT images. In *International Conference on hybrid artificial intelligence systems* (pp. 476–483). Springer.

Sensi, F., Rei, L., Gemme, G., Bosco, P., Chincarini, A. (2014). Global disease index, a novel tool for MTL atrophy assessment. In *MICCAI workshop challenge on computer-aided diagnosis of dementia based on structural MRI data* (pp. 92–100).

Somasundaram, K., & Genish, T. (2014). The extraction of hippocampus from MRI of human brain using morphological and image binarization techniques. In *2014 International Conference on electronics and communication systems (ICECS)* (pp. 1–5). IEEE.

Walter, B., Blecker, C., Kirsch, P., Sammer, G., Schienle, A., Stark, R., Vaitl, D. (2003). MARINA: an easy to use tool for the creation of MAsks for Region of INterest analyses. In *9th International conference on functional mapping of the human brain* (Vol. 19).

Wenlu, Y., Fangyu, H., Xinyun, C., Xudong, H. (2011). ICA-based automatic classification of PET images from ADNI database. In *International Conference on neural information processing* (pp. 265–272). Springer.

World Health Organization (2017). Dementia fact sheet. http://www.who.int/mediacentre/factsheets/fs362/en/. [Online; accessed 20 Dec 2017].

Yang, W., Lui, R.L., Gao, J.H., Chan, T.F., Yau, S.T., Sperling, R.A., Huang, X. (2011). Independent component analysis-based classification of Alzheimer's disease MRI data. *Journal of Alzheimer's Disease*, *24*(4), 775–783.