International Journal of Neural Systems, Vol. 12, No. 6 (2015) 1–23

ENSEMBLES OF DEEP LEARNING ARCHITECTURES FOR THE EARLY DIAGNOSIS OF THE ALZHEIMER'S DISEASE

ANDRÉS ORTIZ
Communications Engineering Department. University of Málaga
Málaga, 29071/Spain
E-mail: aortiz@ic.uma.es
www.uma.es

JORGE MUNILLA
Communications Engineering Department. University of Málaga
Málaga, 29071/Spain
E-mail: munilla@ic.uma.es
www.uma.es

JUAN M. GÓRRIZ*
Department of Signal Theory, Communications and Networking. University of Granada.
Granada, 18060/Spain
E-mail: gorriz@ugr.es
www.ugr.es

JAVIER RAMíREZ
Department of Signal Theory, Communications and Networking. University of Granada.
Granada, 18060/Spain
E-mail: javierrp@ugr.es
www.ugr.es

FOR THE ALZHEIMER'S DISEASE NEUROIMAGING INITIATIVE†

Computer aided diagnosis (CAD) constitutes an important tool for the early diagnosis of Alzheimer's Disease (AD), which, in turn, allows the application of treatments that can be simpler and more likely to be effective. This paper explores the construction of classification methods based on deep learning architectures applied on brain regions defined by the Automated Anatomical Labelling (AAL). Gray Matter (GM) images from each brain area have been split into 3D patches according to the regions defined by the AAL atlas and these patches are used to train different deep belief networks. An ensemble of deep belief networks is then composed where the final prediction is determined by a voting scheme. Two deep learning based structures and four different voting schemes are implemented and compared, giving as a result a potent classification architecture where discriminative features are computed in an unsupervised fashion. The resulting method has been evaluated using a large dataset from the

*Department of Signal Theory, Communications and Networking. E.T.S. de Ingenierías Informática y de Telecomunicación. Periodista Daniel Saucedo Aranda s/n. 18014 - Granada (Spain)

†Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

Alzheimer's disease Neuroimaging Initiative (ADNI). Classification results assessed by cross-validation prove that the proposed method is not only valid for differentiate between controls (NC) and AD images, but it also provides good performances when tested for the more challenging case of classifying Mild Cognitive Impairment Subjects (MCI). In particular, the classification architecture provides accuracy values up to 0.90 and AUC of 0.95 for NC/AD classification, 0.84 and AUC of 0.91 for stable MCI /AD classification and 0.83 and AUC of 0.95 for NC/MCI converters classification.

Keywords: Deep Learning; Ensemble; Alzheimer's Disease classification

1. Introduction

Alzheimer's Disease (AD) is the most common cause of dementia among older people and a third of young people with dementia have AD, affecting 30 million people worldwide. Due to the increasing life expectancy and the ageing of the population in developed nations, it is expected that AD will affect 60 million people worldwide over the next 50 years. It is a slow neurodegenerative disease associated to the production of β-amyloid peptide (Aβ) and its extracellular deposition as well as the flame -shaped neurofibrillary tangles of the microtubule binding protein tau.[1] This causes the loss of nerve cells, whose symptoms usually start with mild memory problems, turning into severe brain damage in several years. There is no cure for AD, and currently developed drugs can only help to temporarily slow down the progression of the disease.[2] Thus, early diagnosis becomes the best way to have effective treatments.
Since the AD neurodegeneration process progressively affects different brain functions,functional images such as Single Emission Computerized Tomography (SPECT)[3–5] or Positron Emission Tomography (PET)[6,7] have been extensively used in Computer Aided Diagnosis systems.[8] Other works present different techniques that allow to discover alterations in electroencephalography (EEG) patterns associated to AD[9–11] that have been used for automated diagnosis.[9,12–15] For instance, in Ref. 16 and Ref. 17 a probabilistic neural network is used for classification between NC and AD by means of conventional and wavelet coherence-based features extracted from EEG data.

AD also causes structural changes in the brain and thus structural differences between controls and AD patients can be revealed by analysis of Magnetic Resonance Images (MRI). In fact, MRI has been used in many previous works for automatic diagnosis.[18–21] These works use White Matter (WM) or Grey Matter (GM) images on whole brain volume to classify controls and AD images[20,21] or to compute Regions of Interest (ROI). Other approaches define weak classifiers on small enough regions.[22,23] Specifically, Ref. 22 uses an ensemble of sparse representation classifiers (SRC) defined on equally-sized patches extracted from the GM image. By contrast, Ref. 23 uses an ensemble of Support Vector Machines (SVM) to classify separately each area defined by the Automated Anatomical Labelling Atlas (AAL). Despite showing good classification results, both proposals present different drawbacks. The former splits the brain into equally-sized patches, and instead of computing discriminative features, performs the classification directly using voxel values, in a similar way to using Voxel-as-Features (VAF) method over small regions, and therefore shares the *curse of dimensionality problem*. The latter extracts some first order statistics from each brain area to be used as features, not considering the spatial relationship among voxels. Moreover, these methods use supervised learning for both, computing the statistical relevance of each brain region and training the classifier, which could be a problem whenever not all the training samples are labelled, or the labels are not reliable enough to use them as ground truth. This is a relatively common problem in AD labels, as they are assigned from the Mini Mental State Examination (MMSE) score. Additionally, there are works that show clear advantages of using a reduced number of discriminative features, such as eigenbrains-based methods,[5] multivariate Gaussian methods,[7] codebook based methods[18] or SVM-based methods.[24] Nevertheless, (e.g. Ref.24) use a downsized cohort of subjects in the study, complicating the estimation of the generalization error. Specifically, we propose the use of deep learning architectures to extract representative features from each brain area defined by the AAL atlas in an unsupervised manner, avoiding the need for a ground truth at this stage. We implement and

compare here different architectures to define ensembles of Deep Belief Networks (DBN). Each brain area has been split into small three-dimensional patches which act as input samples of these DBNs. Different voting schemes to combine the DBNs are analyzed, and an alternative architecture where a SVM (Support Vector Machine) is used to fuse the DBN outcomes is presented. An important aspect of the latter is that each unit in the ensemble is responsible not only of classifying the corresponding patch but also of extracting representative features for the different brain regions.

The organization of the rest of this paper is as follows. Section 2 describes the database and the methods used in this work. In particular, image preprocessing and brain parcellation are explained in subsection 2.2, while backgrounds in Deep Belief Networks and Support Vector Classifiers (SVC) are given in Section 2.3 and Section 2.4, respectively. Section 3 shows details on the experiments performed and the results obtained using patient data from the ADNI database. In this Section, classification of NC/AD subjects is performed, but also experiments involving stable MCI /AD subjects and NC / MCI converters are addressed to deal with early diagnosis. Finally, the main conclusions are drawn in Section 4.

2. Materials and methods

2.1. *Database*

Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). For up-to-date information, see www.adni-info.org.

ADNI database collects a vast amount of MRI and Positron Emission Tomography (PET) images, as well as blood biomarkers and cerebrospinal fluid analyses, for three groups of subjects: healthy individuals (Controls, NC), Alzheimer disease patients (AD) and patients suffering from mild cognitive impairment symptoms (MCI). The database that has been used in this work, contains 1075 T1-weighted MRI images, comprising 229 NC, 401 MCI (312 stable MCI and 86 progressive MCI) and 188 AD images. Specifically, we have used the database *ADNI1:Screening 1.5T (subjects who have a screening data)*. This database contains MRI data from 818 subjects and repeated scans in some cases. When multiple scans of the same subject were available, the first one was selected. As a result, 818 MR images were first selected for assessing our approach. However, as our study includes multimodal data (i.e. MRI and PET images) and PET data are not available for all patients, we have only selected those patients having MRI and PET images simultaneously and taken on the same date. This way, 68 NC, 70 AD, 111 MCI and 26 Late MCI (LMCI patients) were selected. Demographic data of patients in the multimodal database is summarized in Tab. 1.

Table 1. Demographic data of patients in the database

Diagnosis	Number	Age	Gender M/F	MMSE
Control	68	75.81 ± 4.93	43/25	29.06 ± 1.08
MCI	111	76.39 ± 6.96	76/35	26.68 ± 2.16
AD	70	75.33 ± 7.17	46/24	22.84 ± 2.91
LMCI	26	73.06 ± 7.08	21/5	27.27 ± 1.89

In the ADNI2 database, MCI patients are split into two subclasses: Late MCI (LMCI in Tab. 1) and Early MCI (MCI in Tab. 1). Details regarding these groups can be found at http://adni.loni.usc.edu/wp-content/uploads/2008/07/adni2-procedures-manual.pdf. Hereafter, we consider MCI patients and these were taken into account when searching for converters (MCI patients who converted to AD within 2 years) in the ADNI database. It is also important to highlight that clinical labels in ADNI database are assigned according to MMSE values and are not 100% accurate. In other words, the presence/absence of AD pathology in controls and MCI patients is not verified either with cerebrospinal fluid or amyloid-PET biomarkers. This could cause that some of the controls subjects are in the asymptomatic stage of AD, or that neurode-

generation/amyloid load is below the standardized cut-off values in some of the MCI subjects. As a result, the use of ADNI clinical labels could underestimate the classification performance.

It is worth noting that women are exposed to a higher risk of AD and consequently, AD prevalence is higher in females.[25] Gender correction in AD prediction could improve the prediction of different AD-related biomarkers.[25,26] However, our work is focused on classification and previous works using MRI have demonstrated that genders were not significant predictors for the group separation.[27–29]

2.2. *Image preprocessing and brain parcellation*

MRI and PET images from the ADNI database have been spatially normalized according to the PET and VBM-T1 templates, respectively, ensuring each image voxel correspond to the same anatomical position. After image registration, all the MRI images from ADNI database were resized to 121x145x121 voxels with voxel-sizes of 1.5 mm (Sagittal) x 1.5 mm (coronal) x 1.5 mm (axial), and PET images were resized to 79x95x68 voxels with voxel-size of 3 mm (Sagittal) x 3 mm (Coronal) x 3 mm (Axial). Subsequently, MRI and PET images are treated differently. MRI images are segmented into White Matter (WM) and Grey Matter (GM) tissues using the VBM toolbox for SPM[30,31] . This process, which provides information about GM and WM tissue distributions, is guided by means of tissue probability maps of GM, WM or cerebro-spinal fluid (CSF). A nonlinear deformation field is estimated that best overlays the tissue probability maps on the individual subjects' images. The tissue probability maps provided by the International Consortium for Brain Mapping (ICBM) are derived from 452 T1-weighted scans, which were aligned with an atlas space, corrected for scan inhomogeneities, and classified into GM, WM and CSF. The segmentation process produces values in the range $[0, 1]$, which denotes the membership probability to a specific tissue.

PET images, for their part, are also normalized in intensity in order to compute comparable levels among the images. Since the cerebellum is considered as a constant activation region,[32] intensity normalization is performed by means of the mean cerebellum activation level, which is used as a normalization value. More specifically, the normalization value applied to each image is calculated as the mean of the 1% of the voxels with a higher activation level in the cerebellum. This normalization method, which is commonly used in radiology,[32,33] helps to homogenize the activation levels making them comparable by using the same scale. It is important to mention that studies combining FDG-PET and MRI data tend to under- and over- estimate the tracer concentration if partial volume correction (PVC) is not applied.[34,35] However, PVC could also introduce an additional error depending on the anatomical position.[36] On the other hand, a recent work by Teipel et al.[37] analyzes the use of PVC in the classification methods and concludes that PVC only provides a slight improvement in the predictive performance of FDG-PET data. As a consequence, since this work is focused on the classification methods, PVC has not been applied.

2.2.1. *Voxel preselection*

Voxel preselection has been applied to each image modality separately to remove low significance voxels and reduce the computational burden due to the high dimensionality of the input space. This feature preselection was performed by means of Welch's t-test hypothesis testing separately for each image type.

Welch's t-test allows to test the difference between the means of two populations (e.g. NC and AD) when the variances are unequal, and can be calculated using the following expression

$$I^t = \frac{I^\mu_{NC} - I^\mu_{AD}}{\sqrt{\frac{I^\sigma_{NC}}{N_{NC}} + \frac{I^\sigma_{AD}}{N_{NC}}}} \tag{1}$$

where I^μ_{NC} and I^μ_{AD} are the mean images for NC and AD respectively, I^σ_{NC} and I^σ_{AD} are the variance images, and N_{NC}, N_{AD} are the number of NC and AD images respectively. Mean images I^μ_{NC} and I^μ_{AD} are computed as

$$I^\mu_{NC} = \frac{1}{N_{NC}} \sum_{j=1}^{N_{NC}} I_j \qquad I^\mu_{AD} = \frac{1}{N_{AD}} \sum_{j=1}^{N_{AD}} I_j \tag{2}$$

and variance images I^σ_{NC} and I^σ_{AD} are computed as

$$I^\sigma_{NC} = \frac{1}{N_{NC}} \sum_{j=1}^{N_{NC}} \left(I_j - I^\mu_{NC}\right)^2 \quad I^\sigma_{AD} = \frac{1}{N_{AD}} \sum_{j=1}^{N_{AD}} \left(I_j - I^\mu_{AD}\right)^2 \tag{3}$$

I^t represent the image composed by the t-value provided by Welch's t-test for each image voxel, which is

a significance measurement on the means difference. Greater t-values correspond to lower p-values, where p is the probability of observing the given value t, or one more extreme, by chance if the null hypothesis, which argues for equal means, is true. Hence, small values of p lead to reject the null hypothesis. Thus, depending on the threshold chosen for the p-values, a different number of voxels will be selected. More specifically, the lower the threshold for the p-values the fewer voxels will be selected. In our case, only those voxels of the training set with p-value ≤ 0.05 (5% significance level) have been selected to build the ensembles.

2.2.2. *Brain parcellation*

A key aspect of this work consists in splitting the brain into patches to be classified separately, and therefore we use an atlas defining the different brain regions. In particular, we have used the AAL atlas[38] which defines 116 brain regions corresponding to different neuroanatomical areas.
PET and MRI atlases have been co-registered along with the PET and MRI images respectively, so that both atlas and images voxels correspond to the same neuroanatomical position. This allows us to extract the brain regions indicated in the atlas from the images, making it possible to process them separately. Fig. 1 shows MRI and PET example images and the corresponding atlases.

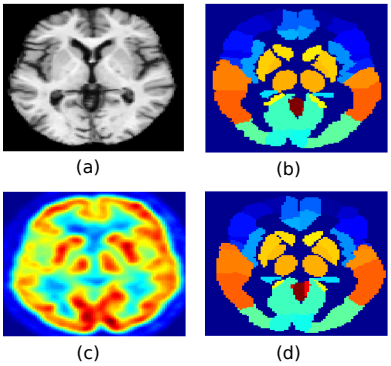

Fig. 1. MRI image (a), MRI atlas (b), (c) PET image and (d) PET atlas (same slice is shown in MRI and PET images).

Although all regions defined by the atlas, comprised of brain regions and cerebellum regions, can be used for classification, we have discarded cerebellum regions as these, according to medical literature,[32,39,40] do not contain discriminant information for the detection of the AD. Some works discard even more brain areas (e.g. Ref. 41) neglecting their influence in the Alzheimer's disease, but we have preferred not to assume this and work with all the rest. This way, as the cerebellum is split into 18 subregions in the AAL atlas, our samples are composed of voxels belonging to the 98 remaining areas.

2.3. *Deep Belief Networks*

A Deep Belief Network (DBN) can be seen as a neural network composed by multiple hidden layers with connections between the layers but not between units within each layer.[42] The core idea is not new and it was already used in multilayer perceptrons or multilayer back-propagation networks. The multilayer architecture tries to mimic the bioinspired model, as it is believed that human brain organizes the information in a hierarchical fashion, from simpler concepts to more abstract representations along with the relationships between these layers. As a typical example, visual cortex model is split into four areas: retina (stores the raw pixels), V1 area (which combines raw pixels and stores edges), V2 area (combining edges to form primitive shape detectors) and V4 area (storing higher level visual abstractions). Nevertheless, the main drawback when using deep architectures in the past stemmed from the training process. In fact, until 2006, many researchers tried to train deep architectures unsuccessfully, and as a result, many of them abandoned the use of multilayer neural architectures in favour of Support Vector Machines (SVM).[43] SVMs can be seen as a smart type of perceptron which uses an optimization technique to compute the weight associated to each feature. This way, SVMs clearly outperformed the multilayer neural networks. However, since 2006, when some specific training algorithms were devised,[44–46] multilayer neural architectures have become popular again. These algorithms facilitate the construction of deep architectures while trained in an unsupervised context, by using unsupervised networks such as Restricted Boltzman Machines (RBM)[47] or autoencoders[48] as single-layer building blocks. More specifically, RBMs are the basic building block of DBNs, as efficient algorithms have been devised to train them unsupervisedly and efficiently.
A RBM is a specific type of Markov random field

with a two-layer architecture that represents the density of the input data $x \in \{0,1\}^d$ (also called visible units) using binary latent variables $h \in \{0,1\}^r$ (also called hidden units). Its basic architecture is depicted in Fig. 2, where ω_{ij} is the weight between the units i and j. In RBM, units at one layer are connected to all the units in the another layer without lateral connections.

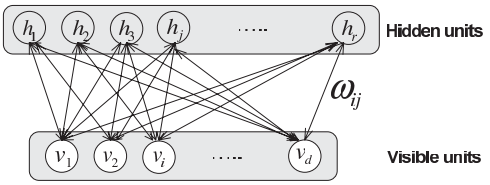

Fig. 2. Restricted Boltzman Machine Architecture

However, binary units provide a very poor representation in the case of natural images, as the data are real-valued. Fortunately, Ref. 49 generalized the RBMs to exponential family distributions allowing the use of real-valued data in the DBM learning process. These are the so-called *Gaussian RBM*. Binary units are replaced by linear units with independent Gaussian noise. This way, the probability of the state of a visible unit can be reconstructed given the state of the hidden units.
Finally, it is worth noting that the *sigmoid* or *logistic* function is used to estimate the activation of each unit.

2.3.1. *Unsupervised learning algorithm*

A fast unsupervised learning method that starts setting the visible units to a training vector was proposed by Hinton,[50] and called Contrastive Divergence (CD). This method updates the weights ω_{ij} by computing the error between the train data and its reconstruction using the current state of the hidden units, without using data labels. Thus, the core equation of the weight updating is

$$\Delta\omega_{ij} = \eta\left(\langle v_i h_j\rangle_{data} - \langle v_i h_j\rangle_{rec}\right) \tag{4}$$

where η is the learning rate and $\langle v_i h_j\rangle_{rec}$ is the reconstruction error.[50] This way, the network trained on a set of examples learns to probabilistically reconstruct the inputs by a learning process that is addressed as an iterative minimization problem.

2.3.2. *Deep Belief Networks*

A Deep Belief Network (DBN) consists of an stack of RBM layers, which are trained using the greedy layer-wise algorithm proposed by Hinton et al.,[42] which allows to train one RBM layer at a time. The core idea of the greedy layer-wise algorithm is to start to train the first RBM using the training data, and continue training higher level RBMs using the current state of the hidden layer at the previous level. This process, sketched in Fig. 3, learns different levels of features; low-level features are located at the bottom (i.e. visible layer), corresponding to raw data, while features encoding higher abstraction levels are hierarchically computed at higher levels of the network.

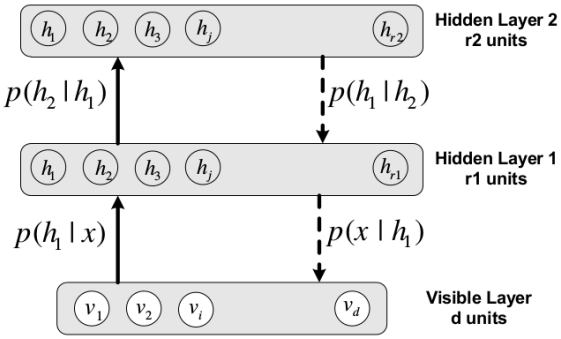

Fig. 3. Deep Belief Network

Following this scheme, RMBs are trained unsupervisedly, as explained in the previous section (using the CD method), to minimize the reconstruction error of the samples. Once this learning step has finished, a DBN can be further trained in a supervised way to perform classification by means of backpropagation algorithms.[51,52] This allows fine-tuning the weights of the network in order to improve its discriminative capabilities.
The network devised for this work includes a layer on the top for the labels when used as a feedforward network. Thus, the constructed network is trained using the backpropagation algorithm and the gradient descent method, which basically consists in computing the error for each sample according to the training labels and then backpropagating it to the first hidden layer. The backpropagation algorithm applied to the proposed network can be summarized as follows. Let t_k the target (desired) output of unit k, and y_k the actual output. The input layer is first feedforwarded

layer wise to the output. Thus, the activation of the neuron k at the output layer is computed using a sigmoid activation function of the weighted sum:

$$y_k = sigm\left(\sum_{j=1}^{r_l} \omega_{jk} h_j + \theta_k\right) \tag{5}$$

where ω_{jk} is the weight of the connection between units j and k and θ_k is a bias term.

The total sum of squared error is computed from the target activation t_k as

$$\epsilon = \sum_k (t_k - y_k)^2. \tag{6}$$

Subsequently, the update rule for the weights between the output layer and the top most hidden layer can be written as

$$\omega_{j,k}(n+1) = \omega_{j,k}(n) + \eta \cdot t_k \cdot \delta_j \tag{7}$$

where n is the epoch number, η the learning rate and δ_j denotes the backpropagated error, computed as

$$\delta_j = (t_k - y_k) \cdot y_k(1 - t_k), \tag{8}$$

or as follows in the case that unit j belongs to a hidden layer:

$$\delta_j = \left(\sum_j \delta_j \cdot \omega_{jk}\right) \cdot y_k \cdot (1 - y_k) \tag{9}$$

This process is repeated in subsequent epochs until ϵ is below a predefined threshold (error tolerance).

2.3.3. *Extracting Features using DBN*

Although DBNs are usually used for classification, in this work we have also focused on their abilities as feature extractors. This is addressed by using the activation of the RBM units at different levels as features that represent different abstraction levels generated during the training process.

DBNs can be thus used to extract features in an unsupervised way, due to the unsupervised training algorithms for RBMs. This approach has been addressed in different works such as Ref. 53, where the Sparse Encoding Symmetric Machine (SESM) is proposed to produce sparse overcomplete representations of the data. Moreover, unsupervised feature learning is also addressed in Ref. 54 using convolutional DBNs to learn feature representations from unlabeled audio data, showing very good performance for different audio classification tasks. In addition, Ref. 55 uses deep autoencoders to extract features from image and CSF biomarkers as well as from MMSE data. Alternatively, supervised training can be used to fine-tune the features computed at each layer by means of backpropagation which aims to minimize the classification error, improving the representation capabilities of the features. This approach, consisting in a classification DBN with unsupervised pre-training is used in Ref. 56 to classify audio data, showing that DBN computed features from raw data performs similarly to MFCC-based features. Nevertheless, Ref. 56 uses the DBN as a classifier but not to extract features from a specific layer. It is worth noting that features produced at different layers represent the data at different abstraction levels and eventually have different discriminative capabilities. Similarly, Ref. 57 uses discriminative DBNs for visual data classification. In this work we not only consider the use of a discriminative DBN but also the features generated at each DBN layer during the training stage, using a SVM as classifier. The entire learning architecture is shown in Fig. 4.

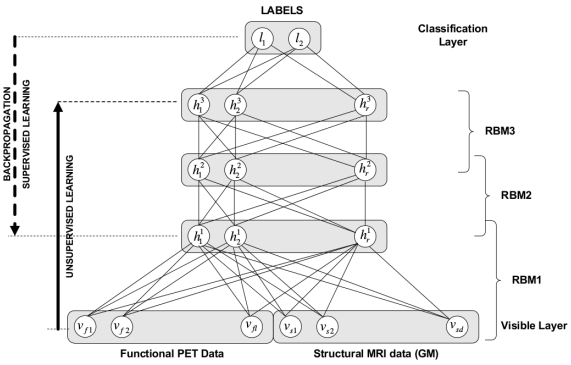

Fig. 4. Proposed architecture for a discriminative DBN

2.4. *Support Vector Machines*

Support Vector Machines (SVM) are a set of supervised learning methods widely used for classification and regression,[43,58–60] designed to separate a

set of binary-labelled data by means of a hyperplane. Specifically, they use a smart optimization method to compute the maximal margin hyperplane to achieve maximum separation between classes, using a decision function in the form $h : \mathbb{R}^n \longrightarrow \{\pm 1\}$, corresponding to n-dimensional training vectors and class labels y_i:

$$(f_1, y_1), (f_2, y_2), ..., (f_s, y_s) \in \mathbb{R}^n \times \{\pm 1\} \quad (10)$$

in such a way that g is able to correctly classify new samples (f, y). Linear discriminant functions define decision hyperplanes in a multidimensional feature space:

$$g(f) = \omega^T f + \upsilon_0 \quad (11)$$

where ω is the weight vector and υ_0 is a bias (threshold). This way, $\omega^T f + \upsilon_0 \geq 1$ if class $y_i = +1$ and $\omega^T f + \upsilon_0 \leq 1$ if class $y_i = -1$, being the weight vector ω orthogonal to the decision hyperplane. The optimization task finds the unknown parameters ω and υ_0 which define the decision hyperplane that separates the two classes optimally.

Additionally, a measure of the relative importance of each feature can be computed. In fact, let *Ns* be the number of *support vectors* within the margin chosen during the training phase, the following vector can be computed:

$$W = \sum_{j=1}^{Ns} y_j \lambda_j f_j \quad (12)$$

where y_j are the labels, λ_j are the corresponding Lagrangian parameters, which are also optimized during the training phase, and f_j the training samples. The coordinate i of the vector W, W_i with $1 \leq i \leq n$, informs us about the relevance of the i-th dimension of the feature vectors.[61] More precisely, the higher the $|W_i|$, the more the relevance of the i-th dimension in the feature vectors. By contrast, $|W_i| = 0$ indicates that the i-th feature does not have any influence in the classification process.

2.5. *Ensemble of Deep Learning Architectures*

A combination of weak classifiers is generally considered to be more accurate than individual classifiers.[22,62] When the dimension of the feature space is high, the use of weak classifiers fed with a reduced number of features each can also help to avoid the *curse of dimensionality problem*[63,64] . Thus, the use of weak classifiers have been previously used to leverage the performance in MRI classification problems. For instance, in Ref. 22, weak Sparse Representation Classifiers (SRC) are defined using randomly extracted 3D patches from GM MRI images. Then, these classifiers are combined following a classical rule based on the SRC residuals. In Ref. 65 an ensemble of SVM classifiers is used over all the brain ROI defined by an atlas. However, it is very important that weak classifiers are properly combined to take full advantage of the ensemble: different methods may be possible that can be more or less accurate depending on the specific individual classifiers and their decision boundaries.[66] In this work we define an individual DBN for each brain region. Thus, although DBNs cannot be considered as weak classifiers, the ensemble of DBNs aims to combine the expertise of individual good classifiers, but specialized in separate domains (i.e. different brain areas). One of the most popular techniques to combine the classifiers in order to compose the ensemble is majority voting. Nevertheless, this method does not weight the individual decision of each classifier, considering all of them equally relevant in the final prediction. A more elaborated combination technique consists in computing a relevance measure associated to each classifier in order to weight the individual decisions. A similar method is used in Ref.22, where the residuals are averaged so that SRCs providing higher residuals have a lower weight in the final decision.

A different technique to combine the classifiers consists in using a new classifier which is fed with the outputs of individual classifiers[67] . For instance, when using a SVM to fuse all the classifiers, the supervised optimization process executed on the training samples will compute the weights that determine the relative importance of each classifier in the final decision[68,69] .

In this work two DBN-based classification methods have been implemented and compared. The first of them is analyzed for four different voting schemes. These schemes are described next.

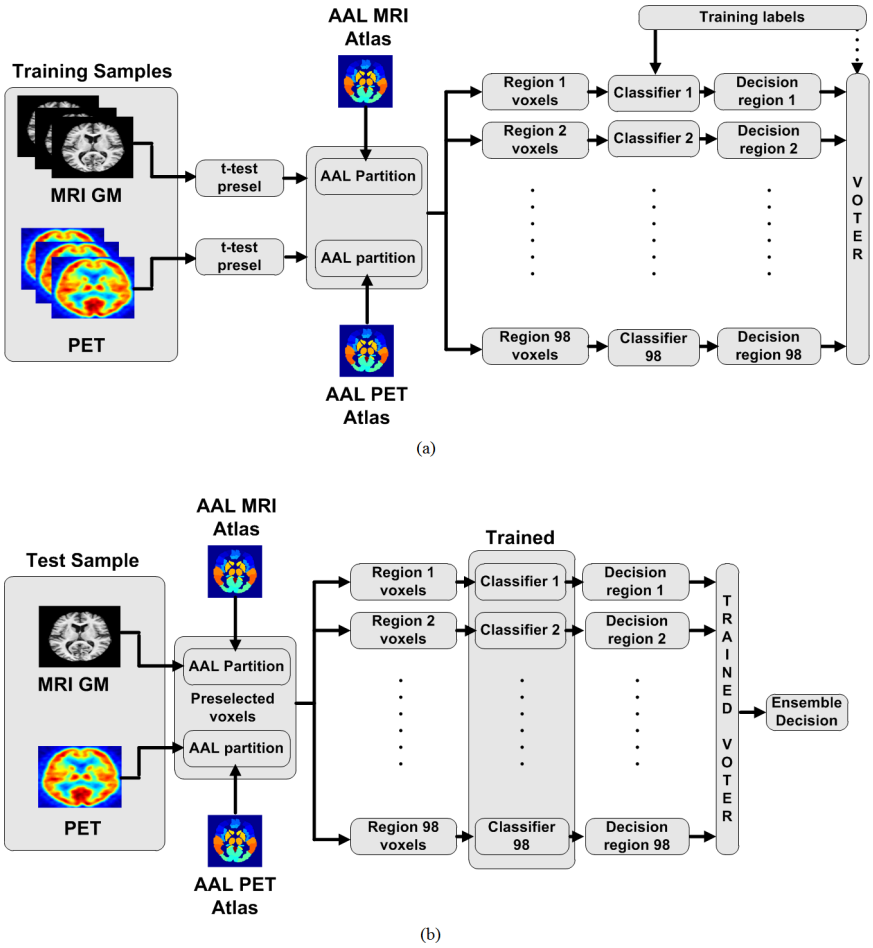

Fig. 5. Ensemble of DBN classifiers and voting stage. In this case, individual DBN are used as weak classifiers defined for each brain region: a) training Phase, b) classification phase. The voter block implements one of the voting mechanisms described in the text.

2.5.1. *Ensemble of DBN classifiers plus voting scheme (DBN-voting)*

The first of the DBN-based classification method consists of an ensemble of discriminative DBNs, in which the top layer is composed of two neurons (for the binary classification problem treated here) in combination with a voting scheme. Four different voting schemes have been compared:

- Majority Voting (MV). The classification outcomes from each region are summed up, so that each region contributes with one vote for a specific subject. The final prediction is determined as the class with the higher number of votes.
- Weighted Voting (WV). We devised a method trying to circumvent the problem that arises in majority voting, due to the fact that not all regions have the same relevance in terms of their discriminative power. This way, a two-sample Welch's test is used to rank the voxels in each region by means of the p-value derived from the hypothesis test. Consequently, the number of voxels in each region with p-value < 0.01 (corresponding to 1% of significance level) is computed, ranking the i-region using the score:

$$SC_i = \frac{\#voxels_i^p(PET) + \#voxels_i^p(GM)}{\#voxels_i} \quad (13)$$

where $\#voxels_i^p(PET)$ and $\#voxels_i^p(GM)$ correspond to the number of voxels in the region i with p-value < 0.01 for PET and MRI-GM images, respectively. The scores computed by this method are used to weight the votes from each region. Latter, the weighted votes of each region are summed up and the final prediction is determined as the

class with the highest overall score. If the p-value used increases, more voxels are included for the computation of the weights and the scheme approaches to the majority voting.

- Classifiers fusion using SVM. This method fuses the elements of the ensemble through a SVM. This way, a SVM is trained with the predictions of each DBN. The support vector weights generated during the training of the SVM will determine the most relevant DBNs and indirectly weight the decision of each individual classifier.
- Classifiers fusion using a DBN. In this case, a discriminative DBN is trained with the classification outcomes from the individual classifiers of the ensemble.

Fig. 5 depicts the block diagram of the ensemble of DBNs. The voter block represents one of the voting mechanisms described above. It is worth noting that training samples are only required in the voter block for weighted voting, as this weights each vote by means of the discriminative power of the corresponding region, computed by the Welch's test on the training samples. Specific details on the implementation of the weighted voting scheme are provided in Section 3.2.

2.5.2. *Extracting features using DBN (FEDBN-SVM)*

In our second approach we take a further step in order to leverage the classification results provided by the ensemble of DBN classifiers, and we use a different implementation in which DBNs are not used as weak classifies but as weak feature extractors. Henceforth, we will denote this architecture as FEDBN, while the term DBN will be used to refer to the former one.

In FEDBN, voxels extracted from each region are used as training samples for a DBN composing a DBN-per-region structure. The activations of the neurons in a hidden layer are then computed and used as features. Finally, the features extracted from each region are concatenated into a unique feature vector to train a SVM. This way, the SVM will compute the relative relevance of each feature during the training stage and avoid the need of the voting phase. The use of a DBN as a feature extractor is based on the idea that it generates a different model in each hidden layer, encoding the sample features in a different number of new features corresponding to the activations of the neurons in each hidden layer. As previously explained, hidden layers in DBNs represent features at different abstraction levels in such a way that higher levels in the network represent higher levels of abstraction (see Fig. 6). However, there is no a priory way to determine the discriminative capability of the features generated at each layer, and the one providing the best performances has to be determined by testing.

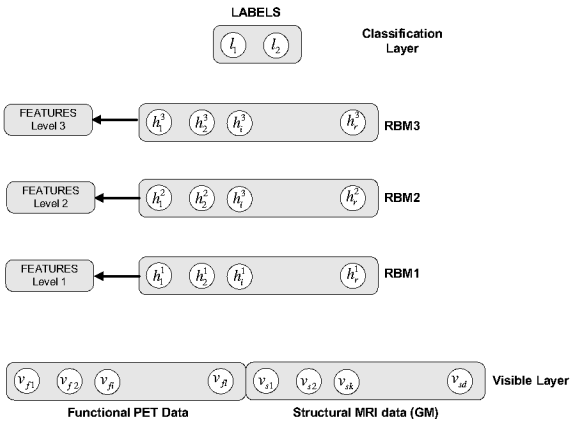

Fig. 6. Feature extraction from different levels of each DBN

Fig. 7 shows the block diagram of the proposed ensemble of DBNs, each of them extracting features from a different brain region, according to the regions defined by the AAL atlas, and fused by a SVM.

3. Results and discussion

In this section, classification outcomes using the proposed classification methods are shown. These results have been also compared with those obtained using other methods. Classification results are assessed by *k-fold* (k=10) cross-validation, namely stratified cross validation, ensuring that each fold has roughly equal size and roughly the same class proportions as in the data manifold. To avoid double dipping, training and testing subsets are disjoint sets and thus they do not share any sample. This process is repeated for the 10 folds and the results provided here are computed as the average of 10 evaluations throughout 10 folds. The main purpose of cross-validation is to estimate the generalization error, ensuring that similar results will be obtained on new data (i.e. low generalization error). In practice, this

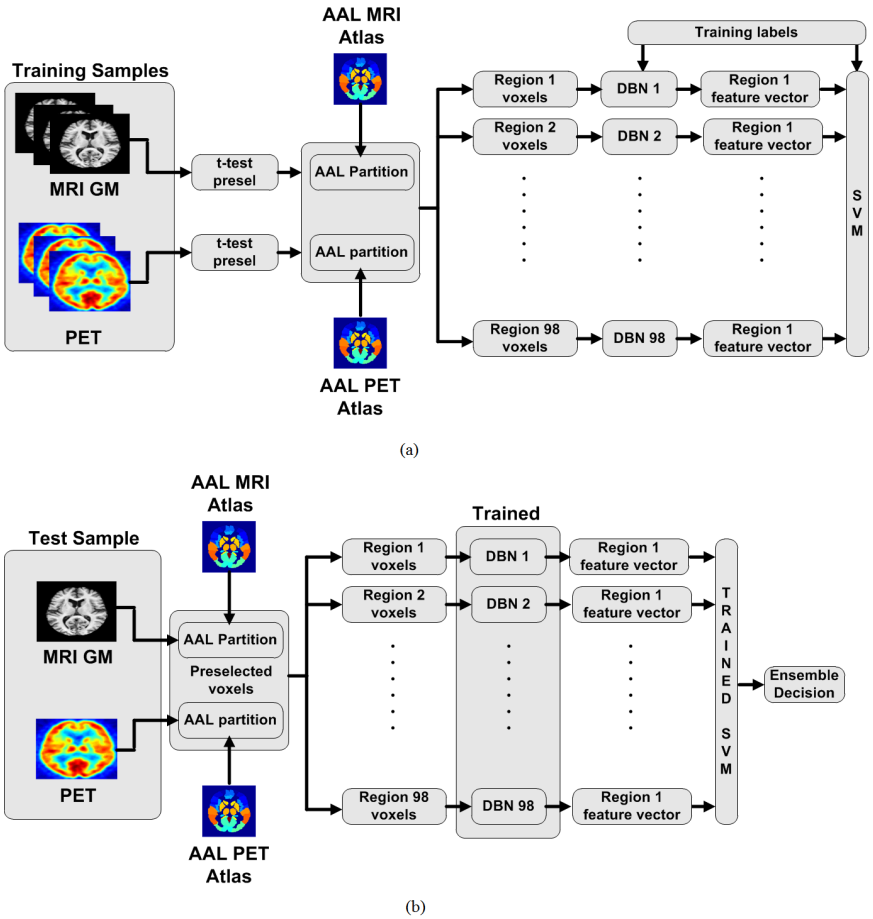

Fig. 7. Block diagram of the proposed ensemble of DBNs for extracting features and classification. In this case, each individual DBN extract features from a specific brain region: a) training Phase, b) classification phase.

error will always result in an overestimate of the true prediction error, since the models obtained during the training phase are not computed using all the training set but $k-1$ folds. This overestimation will depend on the slope of the learning curve of the classifier and reduces when k increases. Thus, the leave-one-out cross-validation ($k = N$, with N the number of available samples) has the lowest bias but can have high variance because the training sets are so similar to one another. Overall, five- or tenfold cross-validation are recommended as a good compromise.

The first experiments, used also to validate and compare different configurations sets, have consisted in classifying between Controls and AD patients. Then, we have addressed the much more challenging case that involves the classification of mild cognitive impairment patients (MCI).[70] In particular, we have performed classification experiments between stable MCI (MCIs) and AD patients, and between controls and MCI converters (MCIc). MCI converters are patients who were diagnosed as MCI but finally converted to AD in the term of 2 years, while Stable MCI are those who remain MCI after this period. The latter case, involving MCIc, deals directly with early AD diagnosis, which constitutes the most relevant issue in AD diagnosis due to its importance in the treatment success.

This section compares the different classification methods that have been implemented to determine that which provides the best results.

3.1. *Parameter set analysis for DBN*

The performances of the DBN-based classification methods will obviously depend on the performances

of the DBNs. These, in turn, will depend on the used parameter configuration. It is well-known that the number of hidden layers and the number of neurons have a direct influence on the representation capabilities and the convergence of the network. Thus, to increase the confidence in our comparisons, we have carried out a previous analysis to determine some optimal values for the number of layers and hidden neurons of the deep learning networks.

Increasing the number of hidden layers implies more epochs to converge and consequently a higher training time[48,51,71] . It has been also shown that, in practice, structures composed of more than 3 hidden layers slightly increases the performance of the network. Therefore, for this work, we have assumed networks with 3 hidden layers.

Although a rough range of possible values could be inferred from the characteristics of the input, the specific number of neurons in the hidden layer has to be determined by testing. To limit the number of possible solutions and since our objective is to compare the representation capabilities at each layer, corresponding to different abstraction levels for the same number of features, we assume here that the three layers are equally sized. Fig. 8 graphs the classification accuracies for the DBN-SVM (SVM as voting mechanisms) and FEDBN-SVM architectures where different numbers of neurons are used. It seems that 400 hidden neurons per layer provide the best results. Consequently, networks with 400 units at each hidden layers are considered hereafter in the experiments.

The contrastive divergence method used to train each RBM requires a certain number of iterations to converge. This number of iterations has to be chosen for a trade-off between representation error and computing time while avoiding over-fitting the data, which would decrease the generalization capabilities of the network. Moreover, the supervised training used to fine-tune the weights by back-propagation has to be also stopped before over-fitting occurs.[72] In the experiments performed, we trained the network across 20 and 200 epochs in the unsupervised (RBM training) and supervised (backpropagation) stages. Experiments conducted using 10 and 100 epochs (for the unsupervised and supervised part, respectively) provided lower accuracy values, while 30 and 300 iterations respectively, do not improve the accuracy. As a conclusion, using more than 20 and 200 iterations could tend to over-fit the data. Additionally, the learning rates in both unsupervised and supervised stages control the portion of weight updating during training. These were also tuned by experimentation and we eventually used the values 0.1 and 0.01 for the unsupervised and supervised phases, respectively. In general, lower learning rates tend to increase the learning rate and the network could be stuck in a local minima. On the contrary, higher values speed up the training process but the network may not converge. Experiments performed using 0.01 and 0.001 for the unsupervised and supervised stages respectively, slowed down the training stage and slightly decreased the classification accuracy. Higher values tested (0.5 and 0.1) decreased considerably the accuracy.

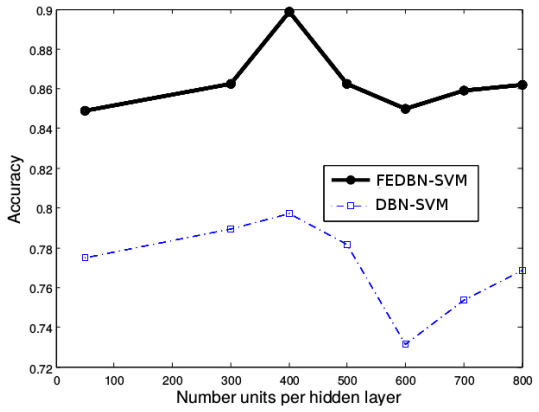

Fig. 8. Accuracy obtained for different number of units in the hidden layer for both implementations, DBN-SVM and FEDBN-SVM. The architecture used consists of three equally sized hidden layers

In Table 2 we summarize the parameters used to train the i-th component (DBN) of the ensemble ($1 \leq i \leq 98$), corresponding to the i-th region, including the learning rates used during the unsupervised (RBM training) and supervised (backpropagation) phases. Note that the only difference between the different DBNs is the number of neurons of the input layer since this corresponds to the number of preselected voxels of each region.

Additionally, when using the DBN as a feature extractor (i.e. FEDBN-SVM), it is also important to determine the layer providing the most discriminative features. As there is no way to know the discrim-

inative power of the features at different levels a priori, different experiments were conducted to evaluate these features. With this aim, different tests have been carried out to compare the classification accuracies when features are extracted from the different hidden layers: L1, L2 or L3. The results are collected in Tab. 3.

Table 2. DBN Parameters. $voxels_i$ refers to voxels of region i

Parameter	Value
# hidden layers	3
# neurons per hidden layer	400
DBN_i Structure	$voxels_i$-400-400-400-2
Unsupervised training epochs	20
Supervised training epochs	200
Unsupervised learning rate	0.1
Backpropagation learning rate	0.01

Although L1, L2 and L3 seem to provide representative features, we consider the use of L2 features as it provides the best performance in terms of AUC metric, measured for a 5% of significance level, which measures the robustness of the classifier taking into account not only the accuracy but also sensitivity and specificity. Consequently, features extracted from hidden layer 2 are used hereafter.

Table 3. Accuracy, Sensitivity and Specificity obtained with features extracted from different layers, corresponding to NC/AD classification with FEDBN-SVM. AUC values are computed for a 5% of significance level ($p < 0.05$).

Layer	Accuracy	Sensitivity	Specificity	AUC
L1	0.88 ± 0.08	0.84 ± 0.10	0.94 ± 0.17	0.94
L2	**0.90 ± 0.09**	**0.86 ± 0.12**	**0.94 ± 0.10**	0.95
L3	0.87 ± 0.09	0.79 ± 0.17	0.96 ± 0.10	0.94

3.2. *Voting Methods comparison*

This section compares the voting methods described in Section 2.5. Experiments combining the architecture DBN with the four different voting schemes are carried out. Table 4 shows the results of these experiments. As previously explained, the DBN voter case consists in fusing the decisions given by the individual classifiers[67] composing the ensemble by means of another DBN. In this case, by experimentation, a 98-100-100-100-2 network structure was selected.

Table 4. Comparison of Voting methods. Results in this Table refers to NC/AD classification using an ensemble of DBN classifiers. AUC values are computed for 5% of significance level.

Voting Method	Accuracy	AUC
Majority Voting (MV)	0.85±0.05	0.83
Weighted voting (WV)	0.86±0.06	0.85
DBN voter (DBN-DBN)	0.78±0.06	0.78
SVM voter (DBN-SVM)	**0.90±0.08**	**0.90**

Statistical significance test of the results aiming to state the best-performing method is addressed by ANOVA[73] analysis using the accuracy values. This revealed that null hypothesis (H_0) can be rejected, which means that at least one group mean differs from the rest. A multiple comparison test was eventually performed to identify these differences by means of the 95% confidence intervals. Consequently, DBN-SVM method outperforms the MV and DBN methods (confidence intervals of [-0.17, -0.06] and [-0.11,-0.02], respectively). At the same time, the superiority of the DBN-SVM method over the WV method cannot be statistically assessed at 5% of significance (confidence interval of [-0.10,-0.01]). However, as SVM voting method provides a higher AUC computed for 5% of significance level, it can be regarded as superior to the other methods.

3.3. *Classification Experiments*

Once the best-performing architectures have been determined, we have carried out the previously described classification experiments between the different subject groups. Tab. 5 shows the classification outcomes obtained for different classification approaches for NC/AD classification. In the first three rows, the *Voxel as Features (VAF)* method,[74] which considers individual voxels as different features, is shown. For VAF experiments we used a linear SVM trained with GM, PET and GM+PET data. Although SVMs are less prone to suffer from this issue than other classifiers,[43] the main drawback of this method is the *curse of dimensionality problem.* Thus, dimensionality of raw data used in VAF experiments was reduced by means of Principal Com-

ponent Analysis (PCA)[75,76] to 10 principal components (under PCA PET+GM row in Tab. 5). Furthermore, an ensemble of linear SVMs, labelled as SVM-e, is included for comparison. In SVM-e each SVM acts as a weak classifier being trained with voxels from one brain region, and then, a new linear SVM is trained using the outputs of each individual SVM composing the ensemble to deliver the final decision. Finally, the last two rows correspond to the proposed DBN-based classification architectures defined in the previous sections. That is, DBN with the four different voting mechanisms, and FEDBN-SVM, where DBNs are not used as classifiers but as feature extractors from each brain region, which allows defining a new space composed of the concatenated features extracted by each DBN. These are then used to train a linear SVM.

Classification performance is assessed by measuring the accuracy, sensitivity and specificity for each method as this is a widely accepted method to evaluate the classification performance[66,77,78] and to estimate the generalizaton error.[66] Moreover, the ROC curve which graphically shows the ability to discriminate between different classes[79] is also provided along with the AUC (Area Under Roc Curve) metric measured for a 5% of significance level, which can be defined as the probability of the classifier to rank a randomly chosen positive sample higher than a randomly chosen negative sample.[79,80] Thus, AUC values fall in the [0,1] range, where 1 indicates perfect discrimination between classes, 0.5 indicates no ability to discriminate (random classifier) and 0 indicates that negative data are always ranked higher than positive data.[66,78,79]

In addition, the Receiver Operating Curve (ROC)[79] computed for the best-performing classification alternatives are also shown in Figure 9, providing AUC values of 0.95, 0.94, 0.94 and 0.79 for FEDBN-SVM, SVM-e, DBN-SVM and VAF PET+GM methods, respectively. AUC values in the Tab. 5 are computed for a significance level of 5% ($p < 0.05$) in all cases. Thus, it can be observed that FEDBN-SVM shows best results that the rest of classification approaches.

3.4. *Ranking ROIs*

The proposed methods where a SVM is used to compose the ensemble, also allow us to rank atlas ROIs according to their relative discriminant capabilities. In fact, as explained in Section 2.4, a weighted sum vector W of the Ns support vectors can be computed. This vector gives us information about the relative importance of each feature.

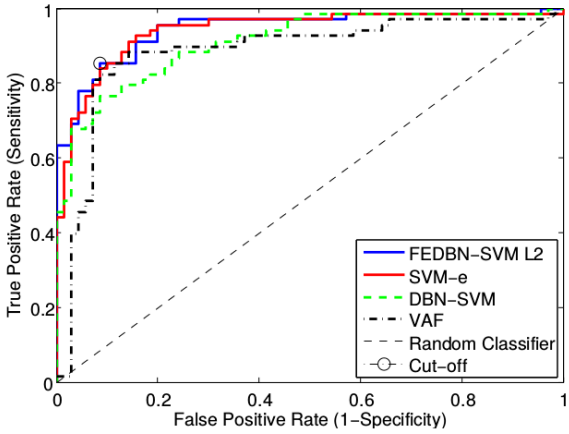

Fig. 9. ROC curves for different NC/AD classification methods.

For the FEDBN-SVM, the selected architecture consists of 400 units at each RBM layer, and therefore 400 features are extracted for each region (including MRI and PET information). The SVM which fuses the features generated in each DBN to compose the ensemble is thus fed with 400*98 features. The SVM then computes 400*98 weights, and the weight of the k-th region can be defined as the sum of all the weights of the features corresponding to that region:

$$W^k = \frac{1}{400} \sum_{i=1}^{400} |w_{k,i}| \tag{14}$$

where $w_{k,i}$ is the SVM weight corresponding to the activation of the neuron i in the region k. These values can be normalized by dividing them by the maximum W^k_{max} for $1 \leq k \leq 98$, so that all the values are in the range [0,1].

For NC/AD classification, Figure 10 and 11 illustrate the computed relative importance of the ROIs in the axial and coronal planes, and the most discriminative brain regions, respectively. These selected regions are according to the medical bibliography.

Table 5. Accuracy, Sensitivity, Specificity and Area Under ROC Curve (AUC) for different classification methods. These results correspond to NC/AD classification. AUC values are computed for a significance level of 5% ($p < 0.05$).

Method	Accuracy	Sensitivity	Specificity	AUC
VAF PET	0.85 ± 0.09	0.89 ± 0.13	0.81 ± 0.12	0.91
VAF GM	0.82 ± 0.12	0.82 ± 0.18	0.81 ± 0.14	0.91
VAF PET+GM	0.86 ± 0.11	0.85 ± 0.13	0.87 ± 0.16	0.88
PCA PET+GM	0.87 ± 0.10	0.85 ± 0.15	0.90 ± 0.10	0.79
SVM-e	0.88 ± 0.08	0.84 ± 0.13	0.92 ± 0.11	0.94
DBN-MV	0.84 ± 0.07	0.80 ± 0.14	0.88 ± 0.09	0.84
DBN-WV	0.87 ± 0.09	0.84 ± 0.16	0.90 ± 0.12	0.87
DBN-SVM	0.88 ± 0.08	0.87 ± 0.14	0.90 ± 0.12	0.93
DBN-DBN	0.78 ± 0.05	0.99 ± 0.05	0.57 ± 0.16	0.77
FEDBN-SVM L2	**0.90 ± 0.09**	**0.86 ± 0.12**	**0.94 ± 0.10**	**0.95**

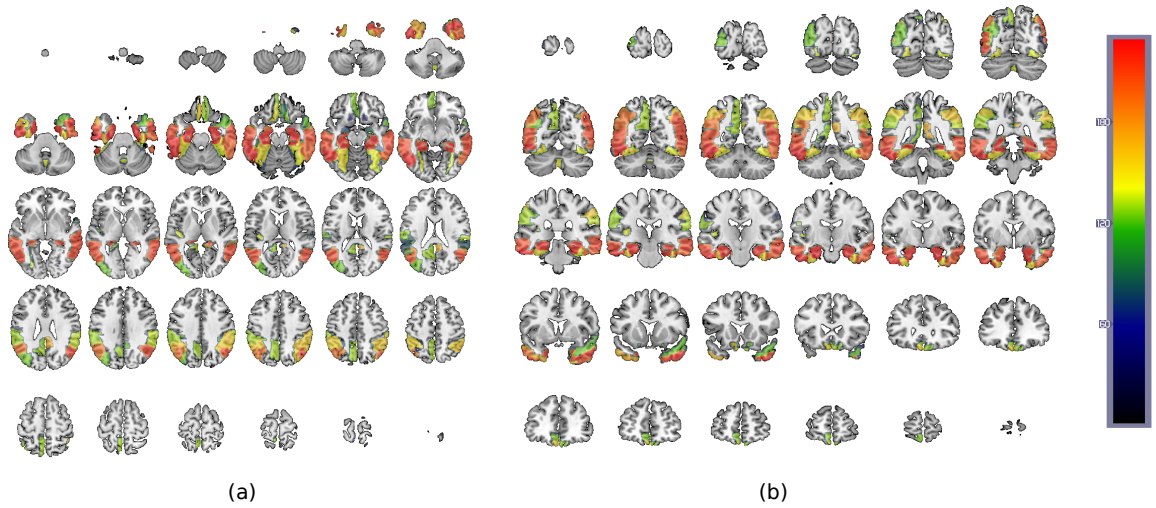

Fig. 10. ROIs computed for NC/AD in the axial (a) and coronal (b) planes. Relative importance is shown in the colorbar (red colour indicates the most discriminative regions).

3.5. *Classification of MCI subjects*

In this section, we address the more complex problem of MCI/AD classification. As previously explained, MCI can be considered as an intermediate state between controls and AD patients, and not all MCI subjects have to develop AD necessarily. Those MCI whose diagnostic changed to AD, according to the MMSE value, within the next two years after being diagnosed as MCI are labelled as MCI converters. Otherwise, they are considered stable MCI. The differences, functional and structural, between the groups are here much more subtle.

3.5.1. *MCIs/AD Classification*

We first tackle the MCIs/AD classification issue. This is possible since ADNI database provides information to identify these patients in the MCI cohort. Indeed, Tab. 6 shows the demographic data of the MCI patients in the database when MCIs and MCIc subjects are differentiated.

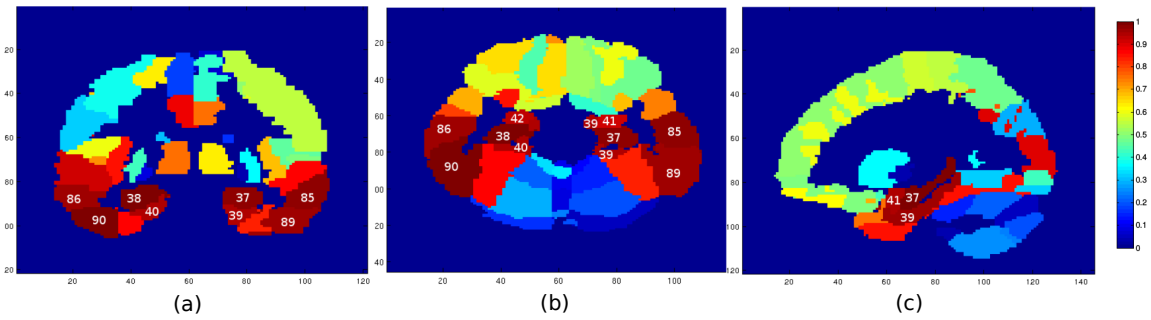

Fig. 11. Most discriminative brain regions for NC/AD classification, identified in axial (a), coronal (b) and sagittal (c) planes. Sorted by discriminative capability, (38) Right Hippocampus, (90) Right Inferior temporal Gyrus, (37) Left Hippocampus, (85) Left Middle Temporal gyrus, (86) Right Middle Temporal gyrus, (89) Inferior Temporal gyrus, (40) Right ParaHippocampal gyrus, (39) Left ParaHippocampal gyrus, (42) Right Amygdala, (41) Left Amygdala.

Table 6. Demographic data of MCIs and MCIc in the ADNI database

Diag.	Num.	Age	Gender M/F	MMSE
MCIs	64	76.46 ± 6.56	45/19	14.73 ± 12.58
MCIc	39	77.02 ± 7.06	25/14	17.05 ± 12.41

We apply the the same classification and analysis methods that those described for NC/AD subjects. Tab. 7 shows the classification performances obtained for the different classification approaches. The results obtained shows that using the raw voxels as features (VAF approach) and a unique classifier is not enough to differentiate between MCIs and AD subjects due to the subtle differences between them. Nevertheless, alternatives using ensembles of classifiers clearly outperform the VAF approach. Specifically, the ensemble of SVMs provides similar performances to those provided by the FEDBN-SVM proposal.

In the same way as for NC/AD, Fig. 12 shows the ROC curves corresponding to the best performing approaches, i.e. FEDBN-SVM, SVM-e, DBN-SVM and VAF PET+GM, according to the AUC value computed from the ROC curve for a 5% of significance level ($p < 0.05$).

ROIs computed for the case of MCIs/AD classification are shown in Fig. 13. Different layers in the axial and coronal planes are depicted to indicate the most discriminative regions. More in particular, the top ten most discriminative regions correspond to (according to the AAL atlas notation): Left Angular Gyrus (65), Right Angular Gyrus (66), Posterior Cingulate Gyrus (35), Left Amygdala (41), Left Hippocampus (37), Right Hippocampus (38), Parahippocampal gyrus (39), Left Inferior Parietal, but supramarginal and Angular Gyri (61), Right Posterior Cingulate Gyrus (36), Left Precuneus (67). Structure or functionality associated to these regions appear in medical literature to be affected in different stages of the AD development.[32,39,40] Note that ROIs appear at different layers, making it difficult to provide a figure similar to Fig. 11.

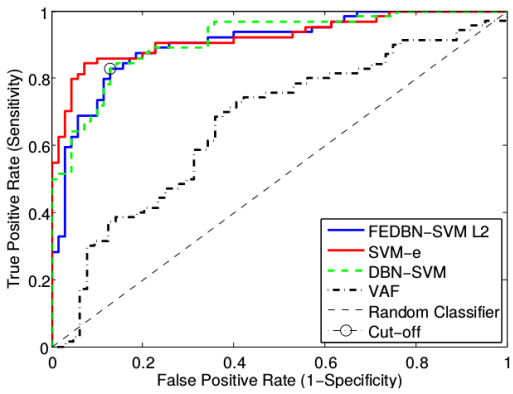

Fig. 12. ROC curves for different MCIs/AD classification methods.

3.5.2. *NC/MCIc Classification: early AD diagnosis*

A further step towards early AD diagnosis involves differentiating between controls and patients who converted to AD (MCIc) in subsequent evaluations.

For NC/MCIc classification, Tab. 8 collects the classification results obtained for the different ap-

Table 7. Accuracy, Sensitivity, Specificity and Area Under ROC Curve (AUC) for different classification methods. These results correspond to stable MCIs/AD classification. AUC values are computed for a significance level of 5% ($p < 0.05$).

Method	Accuracy	Sensitivity	Specificity	AUC
VAF PET	0.65 ± 0.13	0.67 ± 0.17	0.63 ± 0.17	0.72
VAF GM	0.55 ± 0.08	0.53 ± 0.14	0.57 ± 0.13	0.58
VAF PET+GM	0.66 ± 0.11	0.64 ± 0.19	0.69 ± 0.13	0.66
PCA PET+GM	0.70 ± 0.09	0.72 ± 0.11	0.69 ± 0.15	0.77
SVM-e	0.84 ± 0.10	0.80± 0.15	0.88 ± 0.14	0.91
DBN-MV	0.84 ± 0.12	0.83 ± 0.18	0.86 ± 0.13	0.84
DBN-WV	0.84 ± 0.10	0.82 ± 0.15	0.85 ± 0.15	0.80
DBN-SVM	0.86 ± 0.08	0.90 ± 0.12	0.81 ± 0.14	0.85
DBN-DBN	0.69 ± 0.12	1.00 ± 0.05	0.35 ± 0.20	0.67
FEDBN-SVM L2	**0.84 ± 0.09**	**0.79 ± 0.12**	**0.89 ± 0.12**	**0.90**

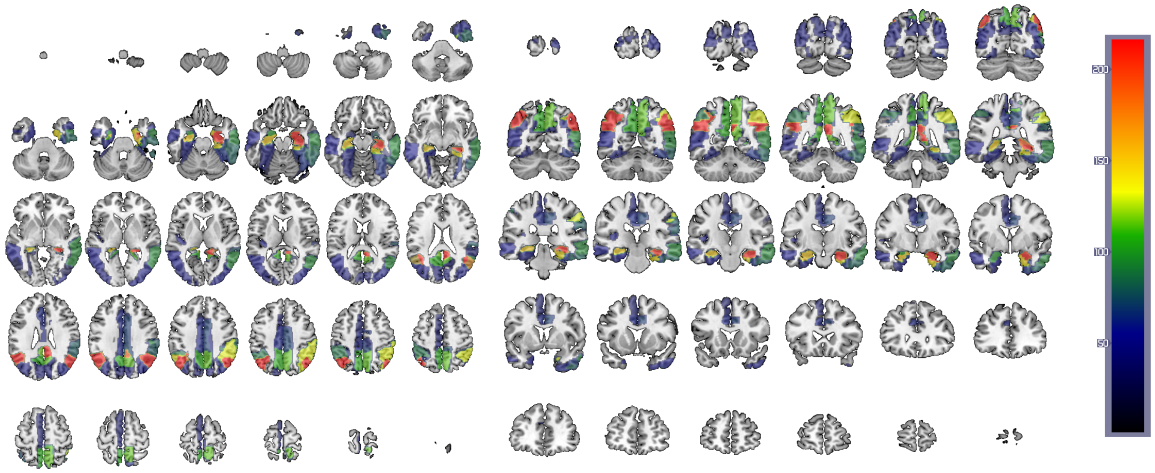

Fig. 13. ROIs computed for MCIs/AD in the axial (a) and coronal (b) planes. Relative importance is shown in the colorbar (red colour indicates the most discriminative regions). These regions include Left Angular Gyrus (65), Right Angular Gyrus (66), Posterior Cingulate Gyrus (35), Left Amygdala (41), Left Hippocampus (37), Right Hippocampus (38), Parahippocampal gyrus (39), Left Inferior Parietal, bu supramarginal and Angular Gyri (61), Right Posterior Cingulate Gyrus (36), Left Precuneus (67) as most discriminative

proaches, while Fig. 14 shows the ROC curve for the best performing ones. The top ten most discriminative ROIs in this case are: Left Amygdala (41), Right Parahippocampal gyrus (40), Right Amygdala (42), Right Hippocampus (38), Right Angular Gyrus (66), Left Hippocampus (37), Left Parahippocampal Gyrus (39), Right Gyrus Rectus (28), Left Temporal Pole: middle Temporal Gyrus (87), Middle Temporal Gyrus (86). In general, as previously commented, information for AD diagnosis is available in GM due to the atrophy produced by GM shrinkage at the time brain ventricles grow larger. By contrast, only mild brain changes are present in MCI patients and therefore most information is contained in PET data.

In this case, the superiority of the FEDBN-SVM method can be stated according to the AUC value computed for 5% of significance level, showing that FEDBN-SVM classifier is more robust than the other approaches, as it provides the highest AUC value.

3.6. *Comparison with other published alternatives*

Finally, this section compares the FEDBN-SVM with other classification methods published in the liter-

Table 8. Accuracy, Sensitivity, Specificity and Area Under ROC Curve (AUC) for different classification methods. These results correspond to NC/MCI converter classification. AUC values are computed for a significance level of 5% ($p < 0.05$).

Method	Accuracy	Sensitivity	Specificity	AUC
VAF PET	0.72 ± 0.14	0.79 ± 0.18	0.62 ± 0.21	0.84
VAF GM	0.63 ± 0.15	0.83 ± 0.18	0.35 ± 0.17	0.65
VAF PET+GM	0.71 ± 0.13	0.86 ± 0.18	0.48 ± 0.13	0.75
PCA PET+GM	0.71 ± 0.19	0.68 ± 0.19	0.75 ± 0.24	0.80
SVM-e	0.83 ± 0.07	0.81 ± 0.13	0.85 ± 0.12	0.94
DBN-MV	0.83 ± 0.11	0.66 ± 0.23	0.95 ± 0.09	0.80
DBN-WV	0.82 ± 0.10	0.60 ± 0.15	0.90 ± 0.15	0.77
DBN-SVM	0.85 ± 0.14	0.69 ± 0.28	0.96 ± 0.08	0.83
DBN-DBN	0.73 ± 0.12	0.95 ± 0.05	0.55 ± 0.21	0.77
FEDBN-SVM L2	**0.83 ± 0.14**	**0.67 ± 0.26**	**0.95 ± 0.09**	**0.95**

ature. Although these comparisons are always arguable as a means to identify the best option, since it is almost impossible to replicate the same initial conditions (i.e. input data), they can help to determine if our proposal is consistent with the state of the art. With this aim, Tab. 10 compares the performances of FEDBN-SVM with others reported in the bibliography. According to these data, FEDBN-SVM outperforms slightly previous results, which, at least, indicates that this proposal must be considered as a serious classification alternative. In addition, works such as Ref. 55 propose the use of deep autoencoders to learn features from image, CSF biomarkers and MMSE data. However, it is worth noting that using MMSE results along with the labels could boost the classification performance, as ADNI labels are based on these MMSE values. On the contrary, our proposal tries to exploit all the information contained in the image data by computing specific features from each brain region by individual DBN-based feature extractors that are eventually combined in an ensemble.

Comment 8.1

On the other hand, we show results obtained using the FEDBN-SVM method when classifying between MCIs and MCIc. At the same time, we also provide CN vs. MCIs classification results for completeness. The classification performances using all the groups exposed in this work are summarized in Tab.9.

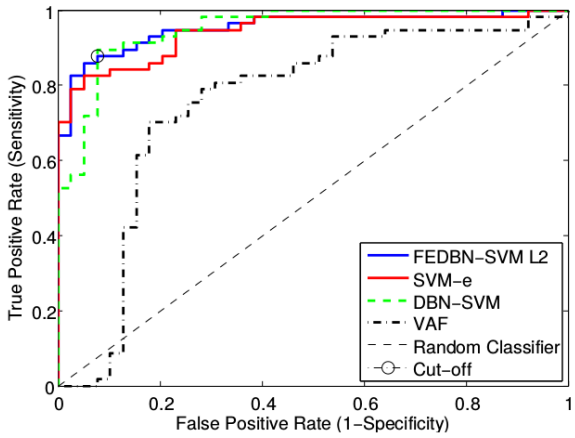

Fig. 14. ROC curves for different NC/MCIc classification methods.

4. Conclusions

This paper presents a method for AD and early AD diagnosis by fusing functional and structural imaging data based on the use of the Deep Learning paradigm, and more specifically, Deep Belief Networks (DBN). A set of DBNs is trained using data from each brain region, according to the AAL atlas, composing an ensemble of DBNs. This concept is used to implement and compare two different DBN-based alternatives: DBN-voter and FEDBN-SVM. The first consists in the use of an ensemble of DBNs classifiers, while the latter is based on the use of DBNs as feature extractors, making use of their capability of representing the information at different abstraction layers.

Four different methods to fuse the decisions of

Table 9. Classification performance of FEDBN-SVM approach for different groups

Method	Subjects (NC/AD)	Acc	Sens	Spec	AUC
CN / AD	68/70	0.90 ± 0.09	0.86±0.12	0.94±0.10	0.95
CN / MCIc	68/39	0.83 ± 0.14	0.67 ± 0.26	0.95 ± 0.09	0.95
CN / MCIs	68/64	0.80 ± 0.12	0.60 ± 0.20	0.90 ± 0.10	0.84
MCIs / AD	64/70	0.84 ± 0.10	0.79 ± 0.12	0.89 ± 0.12	0.90
MCIs / MCIc	64/39	0.78 ± 0.10	0.61 ± 0.15	0.88 ± 0.13	0.82

Table 10. Comparison of NC / AD classification results reported in the literature using MRI image data from the ADNI database

Method	Subjects (NC/AD)	Acc	Sens	Spec
VAF(GM)/(LP) Boosting[81]	183/172	0.82	0.85	0.80
VAF(GM)/SVM[21]	162/137	0.88	0.91	0.95
93 ROI (GM)[82]	52/51	0.86	0.86	0.86
VAF(GM)/SRC-ensemble[22]	228/198	0.90	0.86	0.94
FEDBN-SVM	**68/70**	**0.90**	**0.86**	**0.94**

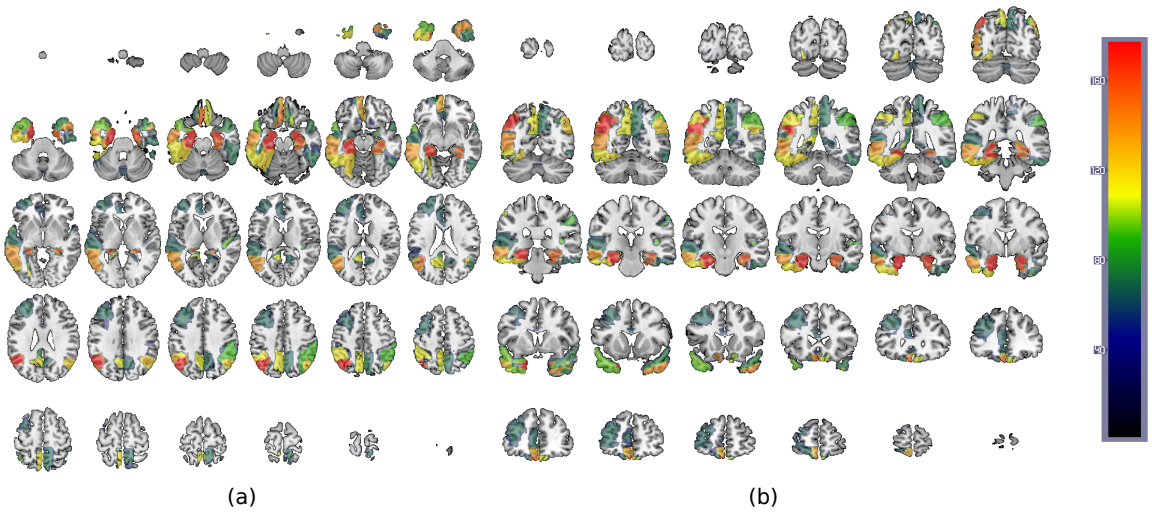

Fig. 15. ROIs computed for NC/MCIc in the axial (a) and coronal (b) planes. Relative importance is shown in the colorbar (red colour indicates the most discriminative regions).

the individual classifiers in the DBN method have been analyzed, and our experiments showed that the best results are obtained with the SVM voter; i.e. DBN-SVM. The best classification outcomes have been, however, obtained using DBNs as feature extractors; i.e. with FEDBN-SVM. It provides higher classification performances in terms of AUC than using discriminative DBNs as classifiers. In order to compare it with other ensemble alternatives, different options have been implemented. For the different classification experiments, FEDBN-SVM proposal outperforms the VAF technique and the results obtained using PCA to reduce the feature space, and offers similar performances to those provided by an ensemble of linear SVMs (SVM-e).

Classification experiments using different groups of subjects have been carried out. Firstly, experiments using the FEDBN-SVM between Controls and AD patients reported an accuracy of 0.90±0.09

and an AUC of 0.95. The proposed classification approach also allowed devising a method to determine the most discriminative ROIs, by using the SVM weights computed during the optimization process that defined the hyperplane; regions associated to AD such as the Hippocampus, the Temporal Gyrus and the Parahippocampal gyrus are pointed out as discriminative by our method, which is according to the medical literature. Next, taking advantage of the possibilities of the ADNI database to identify MCIs and MCIc, regions allowing to differentiate between stable MCIs and AD patients, such as Angular gyrus, Posterior cingulate gyrus, Parahippocampal gyrus and the Hippocampus, were also determined. Finally, classification experiments between NC and MCIc (early AD diagnosis) were performed. The classification outcome in this case reported accuracy of 0.84±0.14 and AUC of 0.95. Two facts were corroborated. First, the classification performance for NC/MCIc is higher than for MCIs/AD, and second, the regions involved in early AD diagnosis (NC/MCIc case) include regions computed as discriminative for NC/AD but with different relative importance.

Acknowledgments

This work was partly supported by the MICINN under the projects TEC2012-34306 and PSI2015-65848-R, and the Consejería de Innovación, Ciencia y Empresa (Junta de Andalucía, Spain) under the Excellence Projects P09-TIC-4530, P11-TIC-7103 and the Universidad de Málaga. Programa de fortalecimiento de las capacidades de I+D+I en las Universidades 2014-2015, de la Consejería de Economía, Innovación, Ciencia y Empleo, cofinanciado por el fondo europeo de desarrollo regional (FEDER) under the project FC14-SAF30.

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.;Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity ; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

References

1. M. P. Murphy and H. LeVine, Alzheimers disease and the β-amyloid peptide, *Journal of Alzheimer's disease* **19**(1) (2010) 311–318.
2. A. D. Society, Factsheet: Drug treatments for alzheimer's disease (2014).
3. J. Ramirez, R. Chaves, J. M. Gorriz, M. Lopez, I. A. Alvarez, D. Salas-Gonzalez, F. Segovia and P. Padilla, Computer aided diagnosis of the Alzheimer's disease combining spect-based feature selection and random forest classifiers, *Proc. IEEE Nuclear Science Symp. Conf. Record (NSS/MIC)*, 2009, pp. 2738–2742.
4. J. Górriz, F. Segovia, J. Ramírez, A. Lassl and D. Salas-González, Gmm based spect image classification for the diagnosis of Alzheimer's disease, *Applied Soft Computing* **11** (2011) 2313–2325.
5. M. López, J. Ramírez, J. Górriz, I. Álvarez, D. Salas-González, F. Segovia, R. Chaves, P. Padilla and M. Gómez-Río, Principal component analysis-based techniques and supervised classification schemes for the early detection of Alzheimer's disease, *Neurocomputing* **74(8)** (2011) 1260–1271.
6. I. Álvarez, J. Gorriz, J. Ramirez, D. Salas-Gonzalez, M. Lopez, F. Segovia, R. Chaves, M. Gomez-Rio and C. Garcia-Puntonet, 18f-fdg pet imaging analysis for computer aided Alzheimer's diagnosis, *Information Sciences* **184(4)** (2011) 903–196.
7. F. Segovia, J. Górriz, J. Ramírez, D. Salas-González, I. Álvarez, M. López, R. Chaves and The Alzheimer's

Disease Neuroimaging Initiative, A comparative study of the feature extraction methods for the diagnosis of Alzheimer's disease using the adni database, *Neurocomputing* **75** (2012) 64–71.
8. H. Adeli, S. Ghosh-Dastidar and N. Dadmehr, Alzheimer's Disease and Models of Computation: Imaging, Classification, and Neural Models, *Journal of Alzheimer's Disease* **7**(3) (2005) 187–199.
9. H. Adeli, S. Ghosh-Dastidar and N. Dadmehr, Alzheimer's Disease: Models of Computation and Analysis of EEGs, *Clinical EEG and Neuroscience* **36**(3) (2005) 131–140.
10. A. H. Ahmadlou, M. and A. Adeli, New Diagnostic EEG Markers of the Alzheimer's Disease Using Visibility Graph, *Journal of Neural Transmission* **117**(9) (2010) 1099–1109.
11. Z. Sankari and H. Adeli, Intrahemispheric, Interhemispheric and Distal EEG Coherence in Alzheimer's Disease, *Clinical Neurophysiology* **122**(5) (2011) 897–906.
12. F. C. Morabito, M. Campolo, D. Labate, G. Morabito, L. Bonanno, A. Bramanti, S. de Salvo, A. Marra and P. Bramanti, A Longitudinal EEG Study of Alzheimer's Disease Progression Based on a Complex Network Approach, *International Journal of Neural Systems* **25**(2) (2015) 1–18.
13. H. Adeli, S. Ghosh-Dastidar and N. Dadmehr, A Spatio-temporal Wavelet-Chaos Methodology for EEG-based Diagnosis of Alzheimer's Disease, *Neuroscience Letters* **444**(2) (2008) 190–194.
14. H. Adeli and S. Ghosh-Dastidar, *Automated EEG-based Diagnosis of Neurological Disorders - Inventing the Future of Neurology* (CRC Press, Taylor & Francis, Boca Raton, Florida, 2010).
15. A. H. Ahmadlou, M. and A. Adeli, Fractality and a Wavelet-Chao Methodology for EEG-based Diagnosis of Alzheimer's Disease, *Alzheimer Disease and Associated Disorders* **25**(1) (2011) 85–92.
16. Z. Sankari and H. Adeli, Probabilistic Neural Networks for EEG-based Diagnosis of Alzheimer's Disease Using Conventional and Wavelet Coherence, *Journal of Neuroscience Methods* **197**(1) (2011) 165–170.
17. Z. Sankari and H. Adeli, Wavelet Coherence Model for Diagnosis of Alzheimer's Disease, *Clinical EEG and Neuroscience* **43**(3) (2012) 268–278.
18. A. Ortiz, J. Górriz, J. Ramírez and F. Martínez-Murcia, Automatic roi selection in structural brain mri using som 3d projection, *PLOS One* **9**(4) (2014).
19. A. Ortiz, J. Górriz, J. Ramírez and F. Martínez-Murcia, LVQ-SVM based CAD tool applied to structural MRI for the diagnosis of the Alzheimers disease, *Pattern Recognition Letters* **34**(14) (2013) 1725–1733.
20. D. Chyzhyk, M. Graña, A. Savio and J. Maiora, Hybrid dendritic computing with kernel-lica applied to Alzheimer's disease detection in mri, *Neurocomputing* **75(1)** (2012) 72–77.
21. R. Cuingnet, E. Gerardin, J. Tessieras, G. Auzias, S. Lehéricy, M. Habert, M. Chupin, H. Benali, O. Colliot and Alzheimer's Disease Neuroimaging Initiative, Automatic classification of patients with Alzheimer's disease from structural MRI: a comparison of ten methods using the adni database, *Neuroimage* **56(2)** (2010) 766–781.
22. Liu, M. and Zhang, D. and Shen, D. and Alzheimer's Disease Neuroimaging Initiative, Ensemble sparse classification of alzheimer's disease, *Neuroimage* **60**(2) (2012) 1106–1116.
23. A. Savio and M. Graa, An ensemble of classifiers guided by the aal brain atlas for alzheimers disease detection, *Advances in Computational Intelligence*, eds. I. Rojas, G. Joya and J. Cabestany, *Lecture Notes in Computer Science* **7903** (Springer Berlin Heidelberg, 2013), pp. 107–114.
24. M. K. B. H. V. A. S. O. S. M. L. . A. D. N. I. Dukart, J., Meta-analysis based svm classification enables accurate detection of alzheimer's disease across different clinical centers using fdg-pet and mri, *Psychiatry Research: Neuroimaging* **212**(3) (2013) 230 – 236.
25. C. J. Jack, R. Petersen, X. Y.C., P. O'Brien, G. Smith, R. Ivnik, B. Boeve, S. Waring, E. Tangalos and K. E., Prediction of ad with mri-based hippocampal volume in mild cognitive impairment, *Neurology* **52**(7) (1999) 1397–1403.
26. K. A. Lin and P. M. Doraiswamy, When mars versus venus is not a clich: gender differences in the neurobiology of alzheimers disease, *Frontiers in Neurology* **5**(288) (2015).
27. C. Plant, S. J. Teipel, A. Oswald, C. Bhm, T. Meindl, J. Mourao-Miranda, A. W. Bokde, H. Hampel and M. Ewers, Automated detection of brain atrophy patterns based on {MRI} for the prediction of alzheimer's disease, *NeuroImage* **50**(1) (2010) 162 – 174.
28. S. J. Teipel, C. Born, M. Ewers, A. L. Bokde, M. F. Reiser, H.-J. Mller and H. Hampel, Multivariate deformation-based analysis of brain atrophy to predict alzheimer's disease in mild cognitive impairment, *NeuroImage* **38**(1) (2007) 13 – 24.
29. J. Barnes, G. R. Ridgway, J. Bartlett, S. M. Henley, M. Lehmann, N. Hobbs, M. J. Clarkson, D. G. MacManus, S. Ourselin and N. C. Fox, Head size, age and gender adjustment in {MRI} studies: a necessary nuisance?, *NeuroImage* **53**(4) (2010) 1244 – 1255.
30. J. Ashburner and T. Group, *SPM8*. Functional Imaging Laboratory, Institute of Neurology, 12, Queen Square, Lonon WC1N 3BG, UK (August, 2011).
31. Structural Brain Mapping Group. Department of Psychiatry, Available: http://dbm.neuro.uni-jena.de/vbm8/VBM8-Manual.pdf. Accessed 2014 March 10 (2014).
32. N. S., V. Villemagne, S. Berlangieri, S. Lee, M. Cherk, S. Gong, U. Ackermann, T. Saunder, H. Tochon-Danguy, G. Jones, C. Smith, G. O'Keefe,

C. Masters and C. Rowe, Visual assessment versus quantitative assessment of 11c-pib pet and 18f-fdg pet for detection of Alzheimer's disease., *Journal of Nuclear Medicine* **48**(4) (2004) 34–41.
33. D. Perani, V. D. Nero, G. Vallar, S. Cappa, C. Messa, G. Bottini, A. Berti, D. Passafiume, G. Scarlato, P. Gerundini, G. L. Lenzi and F. Fazi, Technetium-99m hm-pao-spect study of regional cerebral perfusion in early alzheimer's disease, *Journal of Nuclear Medicine* **29**(9) (1988) 1507–1514.
34. G. Giovacchini, A. Lerner, M. Maria, T. Toczek, M. C. Fraser, K. Ma, M. James, C. Demar, P. P. Herscovitch, M. William, C. Eckelman, S. I. Rapoport and R. E. Carson, Brain incorporation of 11c-arachidonic acid, blood volume, and blood flow in healthy aging: a study with partial-volume correction, *J Nucl Med* (2004) 1471–1479.
35. J. Aston, V. Cunningham, M. Asselin, A. Hammers, A. Evans and G. R.N., Positron emission tomography partial volume correction: estimation and algorithms, *J Cereb Blood Flow Metab.* **22** (August 2002) 1019–1034.
36. W. Jagust, D. Bandy, K. Chen, N. L. Foster, S. M. Landau, C. A. Mathis and the ADNI Investigators, The ADNI core PET, *Alzheimer's & Dementia: The Journal of the Alzheimer's Association* **6**(3) (2010) 221–229.
37. S. J. Teipel, J. Kurth, B. Krause, M. J. Grothe and Alzheimer's Disease Neuroimaging Initiative, The relative importance of imaging markers for the prediction of alzheimer's disease dementia in mild cognitive impairment - beyond classical regression, *NeuroImage. Clinical* **8** (2015) p. 583593.
38. G. Flandin, F. Kherif, X. Pennec, D. Riviere, N. Ayache, J.-B. Poline, G. Fl, F. K. , X. Pennec, D. R. , N. Ayache and J. b. Poline , Parcellation of brain images with anatomical and functional constraints for fmri data analysis (2002).
39. A. Minoshima, N. Foster and D. Kuhl, Metabolic reduction in the posterior cinculate cortex in very early Alzheimer's disease, *Lancet* **3440**(8926) (1994) p. 895.
40. P. J. Nestor, P. Scheltens and J. R. Hodges, Advances in the early detection of Alzheimer's disease, *Nature Reviews Neuroscience* **5**(1) (2004) 34–41.
41. S. Huang, J. Li, L. Sun, J. Ye, A. Fleisher, T. Wu, K. Chen, E. Reiman and A. D. N. Initiative, Learning brain connectivity of Alzheimer's disease by sparse inverse covariance estimation, *Neuroimage* **50**(3) (2010) 935–949.
42. G. E. Hinton and S. Osindero, A fast learning algorithm for deep belief nets, *Neural Computation* **18** (2006) p. 2006.
43. V. N. Vapnik, *Statistical Learning Theory* (Wiley-Interscience, 1998).
44. Y. Bengio, P. Lamblin, D. Popovici, H. Larochelle, U. D. Montral and M. Qubec, Greedy layer-wise training of deep networks, *In NIPS*, (MIT Press, 2007).
45. C. Poultney, S. Chopra and Y. Lecun, Efficient learning of sparse representations with an energy-based model, *Advances in Neural Information Processing Systems (NIPS 2006*, (MIT Press, 2006).
46. S. Dura-Bernal, G. Garreau, A. Andreou, S. Denham and T. wennekers, Multimodal integration of micro-doppler sonar and auditory signals for behabior classification with convolutional networks, *International Journal of Neural Systems* **13**(1) (2013) 1–15.
47. P. Smolensky, Parallel distributed processing: Explorations in the microstructure of cognition, vol. 1 (MIT Press, Cambridge, MA, USA, 1986), pp. 194–281.
48. Y. Bengio, Learning deep architectures for ai, *Found. Trends Mach. Learn.* **2** (January 2009) 1–127.
49. M. Welling, M. Rosen-Zvi and G. E. Hinton, Exponential family harmoniums with an application to information retrieval, *Neural Information Processing Systems (NIPS) 17*, 2004.
50. G. Hinton, Training products of experts by minimizing contrastive divergence, *Neural Computation* **14** (2000) p. 2002.
51. H. Adeli and S.-L. Hung, *Machine Learning: Neural Networks, Genetic Algorithms, and Fuzzy Systems* (John Wiley & Sons, Inc., New York, NY, USA, 1994).
52. R. Rojas, *Neural Networks: A Systematic Introduction* (Springer-Verlag New York, Inc., New York, NY, USA, 1996).
53. M. aurelio Ranzato, Y. lan Boureau and Y. L. Cun, Sparse feature learning for deep belief networks, *Advances in Neural Information Processing Systems 20*, eds. J. Platt, D. Koller, Y. Singer and S. Roweis (Curran Associates, Inc., 2008), pp. 1185–1192.
54. H. Lee, P. Pham, Y. Largman and A. Y. Ng, Unsupervised feature learning for audio classification using convolutional deep belief networks, *Advances in Neural Information Processing Systems 22*, eds. Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams and A. Culotta (Curran Associates, Inc., 2009), pp. 1096–1104.
55. H. Suk and D. Shen, *Deep learning-based feature representation for AD/MCI classification.*, *Medical image computing and computer-assisted intervention : MICCAI ... International Conference on Medical Image Computing and Computer-Assisted Intervention* 2013, pp. 583–590, pt 2 edn.
56. D. Eck and U. D. Montr/'eal, Learning features from music audio with deep belief networks, *11th International Society for Music Information Retrieval Conference (ISMIR 2010*, 2010.
57. Y. Liu, S. Zhou and Q. Chen, Discriminative deep belief networks for visual data classification, *Pattern Recogn.* **44** (October 2011) 2287–2296.
58. C. Sammut and G. I. Webb, *Statistical Learning Theory* (Springer, 2010).
59. Y. Zhang and W. Zhou, Multifractal Analysis and

Relevance Vector Machine-based Automatic Seizure Detection in Intracranial, *International Journal of Neural Systems* **25**(6) (2015) 1–14.
60. E. Castillo, D. Peteiro-Barral, B. Guijarro Berdinas and O. Fontenla-Romero, Distributed One-class Support Vector Machine, *International Journal of Neural Systems* **25**(7) (2015) 1–17.
61. A. Hidalgo-Munñoz, J. Górriz, J. Ramírez and P. Padilla, Regions of interest computed by svm wrapped method for alzheimers disease examination from segmented mri, *Frontiers in Aging Neuroscience* **6**(20) (2014).
62. J. Kittler, M. Hatef, R. Duin and J. Matas, On combining classifiers, *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE* **20** (1998) 226–239.
63. R. Duin, Classifiers in almost empty spaces, *Proceedings 15th International Conference on Pattern Recognition*, **2**2000, pp. 1–7.
64. G. Hughes, On the mean accuracy of statistical pattern recognizers, *IEEE Trans. Inf. Theor.* **14** (September 2006) 55–63.
65. A. Savio and M. Graña, An ensemble of classifiers guided by the AAL brain atlas for alzheimer's disease detection, *Advances in Computational Intelligence - 12th International Work-Conference on Artificial Neural Networks, IWANN 2013, Puerto de la Cruz, Tenerife, Spain, June 12-14, 2013, Proceedings, Part II*, 2013, pp. 107–114.
66. T. Hastie, R. Tibshirani and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, corrected edn. (Springer, August 2003).
67. M. Galar, A. Fernández, E. Barrenechea, H. Bustince and F. Herrera, An overview of ensemble methods for binary classifiers in multi-class problems: Experimental study on one-vs-one and one-vs-all schemes, *Pattern Recogn.* **44** (August 2011) 1761–1776.
68. M. A. H. Akhand, M. Islam and M. KAZUYUKI, A comparative study of data sampling techniques for constructing neural network ensembles, *International Journal of Neural Systems* **19**(02) (2009) 67–89, PMID: 19496204.
69. B. Baruque, E. Corchado and Y. Hujun, The s-2 ensemble fusion algorithm, *International Journal of Neural Systems* **21**(06) (2011) 505–525, PMID: 22131302.
70. M. Ahmadlou, A. Adeli, R. Bajo and H. Adeli, Complexity of Functional Connectivity Networks in Mild Cognitive Impairment Patients during a Working Memory Task, *Clinical Neurophysiology* **125**(4) (2014) 694–702.
71. S. Haykin, *Neural Networks*, 2nd edn. (Prentice-Hall, 1999).
72. H. Hebbo and J. W. Kim, Classification with deep belief networks, project report, TU Berlin (2013).
73. W. Navidi, *Statistics for Engineers and Scientists* (McGraw-Hill Science, 2010).
74. J. Stoeckel and G. Fung, Svm feature selection for classification of spect images of Alzheimer's disease using spatial information, *Proc. Fifth IEEE Int Data Mining Conf*, 2005.
75. S. Theodoridis and K. Koutroumbas, *Pattern Recognition* (Academic Press, 2009).
76. H. Abdi and L. Williams, Principal component analysis, *Wiley Interdisciplinary Reviews: Computational Statistics* **2** (2010) 433–459.
77. D. M. W. Powers, Evaluation: From precision, recall and f-measure to roc., informedness, markedness & correlation, *Journal of Machine Learning Technologies* **2**(1) (2011) 37–63.
78. K. P. Murphy, *Machine Learning: A Probabilistic Perspective* (The MIT Press, 2012).
79. T. Fawcett, An introduction to roc analysis, *Pattern Recogn. Lett.* **27** (June 2006) 861–874.
80. S. Smith and S. Cagnoni, *Genetic and Evolutionary Computation: Medical Applications* (Wiley Publishing, 2011).
81. C. Hinrichs, V. Singh, G. Xu, S. Johnson and the Alzheimers Disease NEuroimaging Initiative, Predictive markers for ad in multi-modality framework: An analysis of mci progression in the adni population, *Neuroimage* **55** (2011) 574–589.
82. D. Zhang, Y. Wang, L. Zhou, H. Yuan and D. Shen, Multimodal classification of Alzheimer's disease and mild cognitive impairment, *Neuroimage* **55**(1) (2011) p. 856867.