Medical Image Analysis 67 (2021) 101850

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Original research

Multi-scale graph-based grading for Alzheimer's disease prediction

Kilian Hett[a,c,*], Vinh-Thong Ta[a], Ipek Oguz[c], José V. Manjón[b], Pierrick Coupé[a], Alzheimer's Disease Neuroimaging Initiative[1]

[a] CNRS, Univ. Bordeaux, Bordeaux INP, LABRI, UMR5800, PICTURA, Talence F-33400, France
[b] Universitat Politècnica de València, ITACA, Valencia 46022, Spain
[c] Vanderbilt University, Department of Electrical Engineering and Computer Science, Nashville, TN, USA

ARTICLE INFO

Article history:
Received 25 July 2019
Revised 18 August 2020
Accepted 31 August 2020
Available online 6 October 2020

Keywords:
Patch-based grading
Graph-based method
Whole brain analysis
Hippocampal subfields
Intra-subject variability
Inter-subject similarity
Alzheimer's disease classification
Mild cognitive impairment

ABSTRACT

The prediction of subjects with mild cognitive impairment (MCI) who will progress to Alzheimer's disease (AD) is clinically relevant, and may above all have a significant impact on accelerating the development of new treatments. In this paper, we present a new MRI-based biomarker that enables us to accurately predict conversion of MCI subjects to AD. In order to better capture the AD signature, we introduce two main contributions. First, we present a new graph-based grading framework to combine inter-subject similarity features and intra-subject variability features. This framework involves patch-based grading of anatomical structures and graph-based modeling of structure alteration relationships. Second, we propose an innovative multiscale brain analysis to capture alterations caused by AD at different anatomical levels. Based on a cascade of classifiers, this multiscale approach enables the analysis of alterations of whole brain structures and hippocampus subfields at the same time. During our experiments using the ADNI-1 dataset, the proposed multiscale graph-based grading method obtained an area under the curve (AUC) of 81% to predict conversion of MCI subjects to AD within three years. Moreover, when combined with cognitive scores, the proposed method obtained 85% of AUC. These results are competitive in comparison to state-of-the-art methods evaluated on the same dataset.

1. Introduction

Alzheimer's disease (AD) is the most prevalent dementia affecting elderly people (Petrella et al., 2003). According to the World Health Organization, the number of patients with AD will double in the next 20 years (Duthey, 2013). AD is a serious condition characterized by an irreversible neurodegenerative process that causes mental dysfunctions such as longterm memory loss, language impairment, disorientation, change in personality, and ultimately causes death (Alzheimer's Association, 2015). The disease is characterized by an accumulation of beta-amyloid plaques and neurofibrillary tangles composed of tau protein (Hardy, 2006) leading to synapse and neuronal losses. To date, no known therapy has been able to stop or hinder the progression of AD. Moreover, because neuroimaging studies have revealed that brain changes occur decades before the diagnosis is established (Coupé et al., 2015; 2019), the pathological load is already high when the diagnosis is made (DeCarli, 2003).

During this pre-diagnosis phase of neurodegeneration, the patient is already suffering from amnesic mild cognitive impairment (MCI). It is noteworthy that although current definition tends to describe AD evolution as a continuum of beta-amyloid accumulation (Jack et al., 2018), MCI is considered a prodromal phase of AD. The clinical symptoms of MCI are slight but still result in a measurable decrease in cognitive abilities. Previous studies have suggested that approximately 12% of subjects suffering from MCI progress to AD in the four years following the first symptoms (Petersen et al., 1999). Therefore, although MCI subjects present a high risk of AD development, subjects suffering from MCI can remain stable (*i.e.*, do not convert to AD). The early prediction of the subjects suffering from MCI symptoms who will convert to AD is thus crucial. This can improve the effectiveness of the future therapies by reducing the brain changes before the therapy starts. Further, the prediction of conversion to AD can accelerate the development of new therapies by making the subject selection more accurate. This would decrease the cost of clinical trials and enable more accurate clinical studies.

* Corresponding author.
E-mail address: kilian.hett@vanderbilt.edu (K. Hett).

[1] 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.

https://doi.org/10.1016/j.media.2020.101850
1361-8415/

With the improvement of medical imaging techniques such as magnetic resonance imaging (MRI), many methods were developed to increase the ability of computer-aided diagnosis systems to help early AD detection (Arbabshirani et al., 2017; Rathore et al., 2017). Computer-aided diagnosis systems describe anatomical information using different type of image-based features that operate at varying scales of analysis. Among these different scales, ranging from local to global, we can cite voxel, patch, shape, thickness, and volume. Considered features are then integrated into machine-learning methods to estimate the pathologic status of the subject under study. These methods can be grouped into two categories related to how they analyze AD alterations:

- Inter-subject similarity: This first group of methods focus on the detection of similarities between individuals from different groups that represent specific disease severities. Among these approaches, a popular method to estimate similarity at a voxel scale is the voxel-based morphometry (VBM) (Ashburner and Friston, 2000; Moradi et al., 2015). Methods based on region of interest (ROI) have also been proposed. A widely used approach is based on a volumetric measurement of gray matter within brain structures (Bron et al., 2015; Ledig et al., 2018). Other ROI-based methods such as thickness measurement were developed to capture the variations of gray matter along the cerebral cortex (Wolz et al., 2011; Wee et al., 2013). Among advanced methods, patch-based grading (PBG) framework was proposed to capture subtler alterations caused by the pathology. This exemplar-based framework aims to detect the similarity in terms of local anatomical patterns by comparing size-restrained area from the subject under study to a template library composed of two different population of subjects. PBG methods has demonstrated state-of-the-art performance to detect alterations of hippocampus (Coupé et al., 2012b; Hett et al., 2018b). This framework has also been extended to perform a whole brain analysis (Tong et al., 2017a). This extension has shown competitive performance for AD prediction especially compared to other approaches based on deep-learning architectures (Basaia et al., 2018; Lian et al., 2018).
- Intra-subject variability: Several methods were proposed to capture the intra-subject variability; such methods assume that AD does not occur at isolated areas but in several inter-related regions. Although similarity-based biomarkers provide helpful tools for detecting the first signs of AD, the structural alterations leading to cognitive decline are not homogeneous within a given subject. Therefore, intra-subject variability features could encode relevant information. Some methods proposed to capture the relationship of spread cortical atrophy with a network-based framework (Wee et al., 2013). Other approaches estimate inter-regional correlation of brain tissue volumes (Zhou et al., 2011). A study has also proposed a generic framework that embeds spatial and anatomical priors within a graph model (Cuingnet et al., 2013). This method extracts intra-subject variability from different features (for instance, voxel-based and cortical thickness) and various MRI modalities (*i.e.*, in their work, they evaluate their method using structural MRI) using an anatomical regularization scheme based on a graph model. More recently, convolutional neural networks (CNN) were used to capture relationships between anatomical structures volumes (Suk et al., 2017), and cortical thickness (Wee et al., 2019). These two last methods, model the structures abnormality relationships using a deep-learning approach. It is interesting to note that methods based on inter-subject similarities and intra-subject variability have performed similarly for AD prediction.

All these elements indicate that inter-subject similarity and intra-subject variability features provide important information for predicting the subject's conversion. Consequently, we proposed to develop a new method that efficiently combines inter-subject similarities estimated with a patch-based grading approach and intra-subjects' variability modeled by a graph-based approach. In our works, given the essential aspect of the interpretability nature of the produced results and the lack of prior knowledge related to the topology of alteration relationships, we opt for a sparse representation of a fully-connected graph defined by an adjacency matrix based on a gaussian kernel.

As a new contribution of the previously published work in conference proceedings (Hett et al., 2018c; 2018a), we applied our new method to two different anatomical scales: hippocampal subfields and whole brain structures. The experiments carried out show an increase in prediction performances for both anatomical scales. We also present a novel method based on a cascade of classifiers to efficiently and simultaneously combine information related to hippocampal subfields and whole brain structures alterations.

2. Materials

2.1. Dataset

Data used in this work were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset[2]. ADNI is a North American campaign launched in 2003 with aims to provide MRI, positron emission tomography scans, clinical neurological measures and other biomarkers. We use T1-weighted (T1w) MRI from the baseline of the standardized ADNI1 dataset (Wyman et al., 2013). This dataset includes AD patients, subjects with mild cognitive impairment (MCI) and cognitive normal (CN) subjects. MCI is a presymptomatic phase of AD composed of subjects who have abnormal memory dysfunctions. In our experiments we consider two groups of MCI. The first group is composed of patients who have stable MCI (sMCI) who did not convert to AD during the 36 months following their first visit, and the second one is composed of patients having MCI symptoms at the baseline and then converted to AD in the following 36 months. This group is named progressive MCI (pMCI). The information of the dataset used in our work is summarized in Table 1. The list of subjects and the code used to construct the graph-based grading features and to evaluate the proposed method are openly available online[3].

2.2. Preprocessing

The data are preprocessed using the following steps: (1) denoising using a spatially adaptive non-local means filter (Manjón et al., 2010), (2) inhomogeneity correction using N4 method (Tustison et al., 2010), (3) affine registration to MNI152 space using ANTS software (Avants et al., 2011), (4) intensity standardization using a piece-wise linear histogram normalization (Manjón et al., 2014). All experiments were conducted with images in the MNI space.

3. Method

3.1. Method overview

As illustrated in Fig. 1, our graph of structure grading method that combines inter-subjects' similarities and intra-subjects' variability is composed of several steps.

First, a segmentation of the structures of interest is computed on the input images. Then, a patch-based grading (PBG) approach

[2] http://adni.loni.ucla.edu.

[3] https://github.com/hettk/multi-scale_graph-based_grading.

Table 1
Description of the dataset used in this work. Data are provided by ADNI.

	CN	sMCI	pMCI	AD	P value
Number	213	90	126	130	
Ages (years)	75.7 ± 5.0	74.9 ± 7.5	73.7 ± 7.0	74.1 ± 7.7	$p = 0.63$[b]
Sex (M/F)	108/105	58/32	68/58	64/66	χ^2=5.29, $p = 0.15$[c]
MMSE	29.1 ± 1.0	27.6 ± 1.7	26.5 ± 1.6	23.5 ± 1.9	$p < 0.01$[a,b]
CDR-SB	3.5 ± 2.7	4.5 ± 2.3	4.8 ± 2.1	4.7 ± 1.9	$p < 0.01$[a,b]
RAVLT	45.4 ± 9.7	35.5 ± 10.2	27.7 ± 8.9	24.6 ± 7.0	$p < 0.01$[a,b]
FAQ	8.4 ± 4.4	13.3 ± 5.4	20.2 ± 6.7	30.0 ± 9.0	$p < 0.01$[a,b]
ADAS11	5.2 ± 3.0	8.1 ± 3.6	12.5 ± 4.9	20.2 ± 7.6	$p < 0.01$[a,b]
ADAS13	0.2 ± 0.9	2.3 ± 3.7	4.3 ± 4.8	14.6 ± 6.6	$p < 0.01$[a,b]

[a] Significant at $p < 0.05$.
[b] Kruskal–Wallis test (df = 3).
[c] Chi-square test (df = 3).

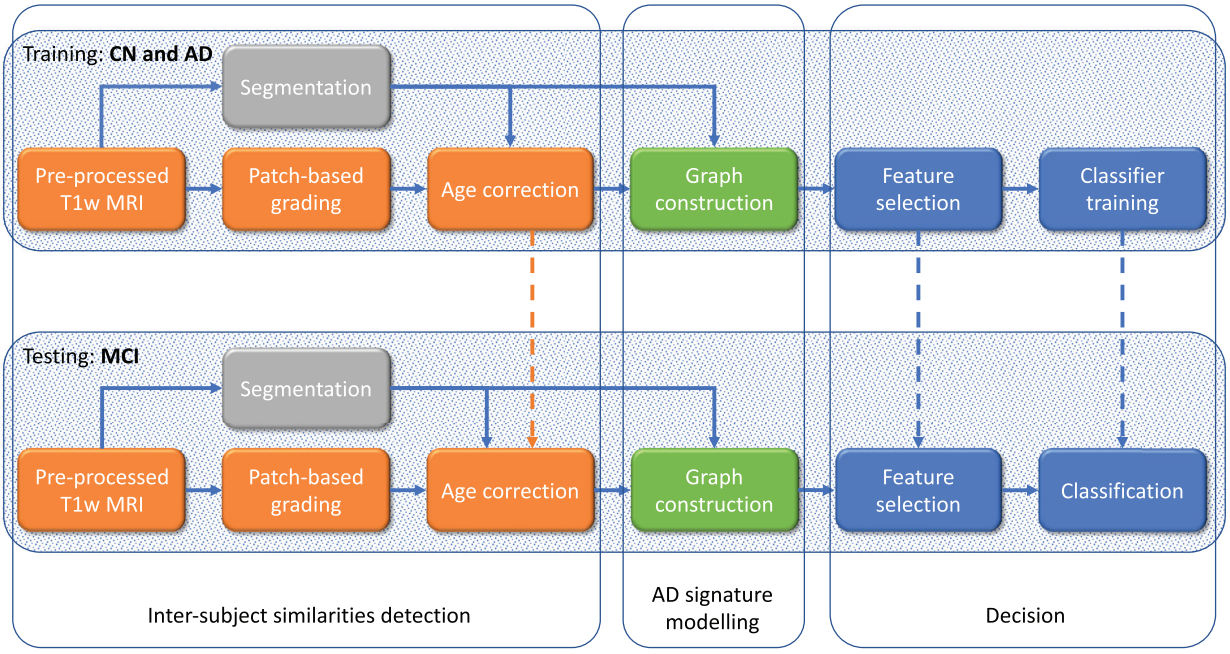

Fig. 1. Pipeline of the proposed graph-based grading method. PBG is computed using CN and AD training groups. CN group is also used to correct the bias related to age. Then, this estimation is applied to AD and MCI subjects. Afterwards, the graph is constructed, and the feature selection is trained on CN and AD and then is applied to CN, AD and MCI. Finally the classifier is trained with CN and AD.

is conducted over every segmented structures (*e.g.*, hippocampal subfields and brain structures). The two main alterations impacting the brain structures captured with PBG methods are the changes caused by normal aging (Koikkalainen et al., 2012) and the alterations caused by the progression of AD. Therefore, at each voxel, the grading values are age-corrected to avoid bias due to normal aging. After the patch-based grading maps are age-corrected, we construct an undirected graph to model the topology of alterations caused by Alzheimer's disease. This results in a high dimensional feature vector. Consequently, to reduce the dimensionality of the feature vector computed by our graph-based method, we use an elastic net that provides a sparse representation of the most discriminative elements of our graph (*i.e.*, edges and vertices). We use only the most discriminative features of our graph as the input to a random forest method which predicts the subject's conversion.

3.2. Segmentation

First, to enable analysis of the alterations that occur over different brain structures, segmentation using a non-local label fusion (Giraud et al., 2016) and a systematic error correction (Wang et al., 2011) at two different anatomical scales are performed, the hippocampal subfields and the whole brain structures.

Segmentation of hippocampal subfields was performed with HIPS, which is a method based on a combination of non-linear registration and patch-based label fusion (Romero et al., 2017). This technique uses a training library based on a dataset composed of high resolution T1w images manually labeled according to the protocol proposed in (Winterburn et al., 2013). To perform the segmentation, the ADNI images were up-sampled to 0.5 × 0.5 × 0.5 mm using a local adaptive super-resolution method to fit in the training image resolution (Coupé et al., 2013). The method provides automatic segmentation of hippocampal subfields grouped into 5 labels: Subiculum, CA1SP, CA1SR-L-M, CA2-3 and CA4/DG. Afterwards, the segmentation maps obtained on the up-sampled T1w images were down-sampled to fit in the previous MNI space resolution. All further hippocampal subfields analyzes were achieved in the MNI space resolution.

Whole brain structures were labeled with a patch-based multi-template segmentation (Manjón and Coupé, 2016). This

method was performed using 35 images manually labeled by Neuromorphometrics, Inc.[4] following the brain-COLOR labeling protocol composed of 133 structures.

Finally, visual quality control was conducted to remove all incorrect segmentations from the dataset. To prevent any bias in the dataset, the pathological status of each subject was hidden during the entire quality control process.

3.3. Patch-based grading

Following image segmentation, a patch-based grading of the entire brain was performed using the method described in (Hett et al., 2018b). This method was first proposed to detect hippocampus structural alterations with a new scale of analysis (Coupé et al., 2012a; 2012b). The patch-based grading approach provides the probability that the disease has impacted the underlying structure at each voxel. This probability is estimated via an inter-subject similarity measurement derived from a non-local approach.

The method begins by building a training library T from two datasets of images: one with images from CN subjects and the other one from AD patients. Then, for each voxel x_i of the region of interest in the considered subject x, the PBG method produces a weak classifier denoted, g_{x_i}, that provides a surrogate for the pathological grading at the considered position i. A PBG value is computed using a measurement of the similarity between the patch P_{x_i} surrounding the voxel x_i belonging to the image under study and a set $K_{x_i} = \{P_{t_j}\}$ of the closest patches P_{t_j}, surrounding the voxel t_j, extracted from the template $t \in T$. The grading value g_{x_i} at x_i is defined as:

$$g_{x_i} = \frac{\sum_{t_j \in K_{x_i}} w(P_{x_i}, P_{t_j}) p_t}{\sum_{t_j \in K_{x_i}} w(P_{x_i}, P_{t_j})} \tag{1}$$

where $w(x_i, t_j)$ is the weight assigned to the pathological status p_t of the training image t. We estimate w as:

$$w(P_{x_i}, P_{t_j}) = \exp\left(- \frac{||P_{x_i} - P_{t_j}||_2^2}{h^2} \right) \tag{2}$$

where $h = \min ||P_{x_i} - P_{t_j}||_2^2 + \epsilon$ and $\epsilon \to 0$. The pathological status p_t is set to -1 for patches extracted from AD patients and to 1 for patches extracted from CN subjects. Therefore, the PBG method provides a score representing an estimate of the alterations caused by AD at each voxel. Consequently, cerebral tissues strongly altered by AD have scores close to -1 while healthy tissues have scores close to 1.

3.4. Graph construction

Once structure alterations were estimated using patch-based grading, we modeled intra-subject variability for each subject using a graph to better capture the AD signature. Indeed, within the last decade, graph modeling has been widely used for its ability to capture the patterns of different diseases (Tong et al., 2017b; Parisot et al., 2018). This is achieved by encoding the relationships of abnormalities between different structures in the edges of the graph. Furthermore, graph modeling can also depict inter-subject similarity, by independently encoding the abnormality of each structure in the vertices measurement. Consequently, we proposed a graph-based grading approach that uses a graph model to combine inter-subject similarities computed with the PBG and intra-subjects' variability computed with the difference of the grading value distributions for each structure.

[4] http://Neuromorphometrics.com.

In our graph-based grading method, the segmentation maps were used to fuse grading values into each ROI, and to build our graph. We defined an undirected graph $G = (V, E, \gamma, \omega)$, where $V = \{v_1, \ldots, v_N\}$ is the set of vertices for the N considered brain structures, $E = V \times V$ is the set of edges, γ and ω are two functions of the vertices and the edges, respectively. In our work, γ is the mean of the grading values for a given structure while ω computes grading distribution distance between two structures. To this end, the probability distributions of PBG values were estimated with a histogram H_v for each structure v. The number of bins was computed with Sturge's rule (Sturges, 1926) using the average of number voxels that compose each brain structure or hippocampal subfields (*i.e.*, histogram of whole brain structures and hippocampal subfields were estimated using different bin number). For each vertex we assigned a function $\gamma : V \to \mathbb{R}$ defined as $\gamma(v) = \mu_{H_v}$, where μ_{H_v} is the mean of H_v. For each edge we assigned a weight given by the function $\omega : E \to \mathbb{R}$ defined as follows:

$$\omega(v_i, v_j) = \exp(-W(H_{v_i}, H_{v_j})^2 / \sigma^2) \tag{3}$$

where W is the Wasserstein distance with L_1 norm (Rubner et al., 2000) that showed best performance during our experiments. Indeed, this metric introduced by the optimal transport theory, aims to minimize the amount of work needed to rearrange the histogram H_{v_i} to H_{v_j}. The Wasserstein distance between two histograms is defined as the minimization of the following equation,

$$W(H_{v_i}, H_{v_j}, F) = \min_{F=\{f_{k;l}\}} \sum_{k,l} f_{k;l} d_{k;l} \tag{4}$$

subject to,

$$\begin{aligned} &\sum_{k \in I} f_{k;l} = p_k \quad \forall k \in I \\ &\sum_{l \in I} f_{k;l} = q_l \quad \forall l \in I \\ &f_{k;l} \geq 0 \quad \forall (k;l) \in J \end{aligned} \tag{5}$$

where $I = \{k | 1 \leq k \leq m\}$ is the index set for bins, $H_{v_i} = \{p_k | k \in I\}$ and $H_{v_j} = \{q_k | k \in I\}$ are the two normalized histograms. $J = \{(k,l) | k \in I, l \in I\}$ is the set for flows, and $d_{k;l} = ||k - l||_1$ is the group distance defined by a L_1 norm.

3.5. Selection of discriminant graph components

Completion of the previous step results in a high-dimensional feature vector. Because features computed from the graph-based grading method have varying significance levels, in this work, we used an elastic net regression method to provide a sparse representation of the most discriminating edges and vertices. This results in reducing the feature dimensionality by capturing the key structures and the key relationships between the different brain structures (see Fig. 1). Indeed, it has been demonstrated that combining the $L1$ and $L2$ norms takes into account possible inter-feature correlation while imposing sparsity (Zou and Hastie, 2005). Finally, after normalization, the resulting feature vector is given as the input of the feature selection, defined as the minimization of the following equation:

$$\hat{\beta} = \min_{\beta} \frac{1}{2} ||X\beta - y||_2^2 + \rho ||\beta||_2^2 + \lambda ||\beta||_1 \tag{6}$$

where $\hat{\beta}$ is a sparse vector that represents the regression coefficients and X is a matrix with rows corresponding to the subjects and columns corresponding to the features, including: the vertices, the edges or a concatenation of both for the full graph of grading feature vector. ρ and λ are the regularization hyper-parameters set to balance the sparsity and the correlation inter-feature, and y represents the pathological status of each patient.

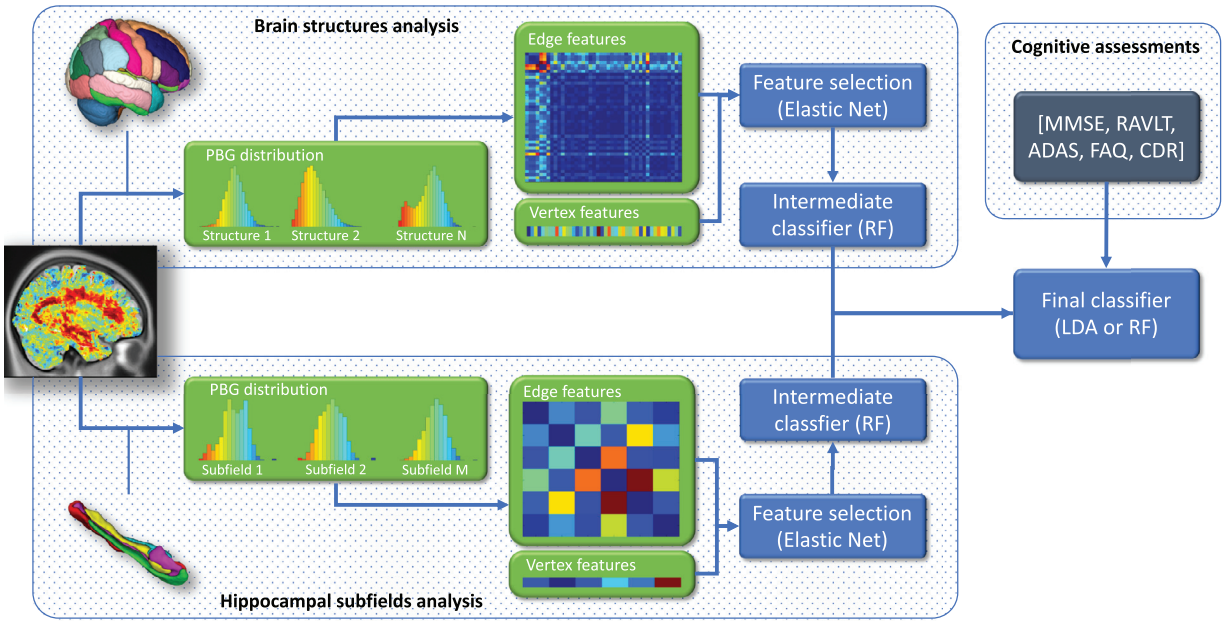

Fig. 2. Schema of the proposed multi-scale graph-based grading method. First, the segmentation maps are used to aggregate grading values. Our method computes a histogram for each structure/subfield. Once the graphs are built, an elastic net is computed to select the most discriminating graph features for each anatomical scale. A first layer of random forest classifiers are computed to estimate *a posteriori* probabilities. Finally, a linear classifier is trained with the *a posteriori* probabilities from each anatomical scale to compute the final decision. A random forest classifier replaces the linear classifier for the multimodal experiments to deal with the feature heterogeneity resulting from the concatenation of *a posteriori* probabilities and cognitive scores.

3.6. Application to different anatomical scales

In our experiments, we considered two different anatomical scales. First, as presented in (Hett et al., 2018a), we applied our graph of structure grading method within a definition of the hippocampal subfields. A histogram was computed to estimate the probability distribution of the grading values for each hippocampal subfields. Thus, $GG_{subfields} = (V, E, \Gamma, \Omega)$, represents the graph of the hippocampal subfields grading. The vertices V represent alteration of hippocampal subfields measured with patch-based grading, and the edges E represent the relationship between hippocampal subfield alterations embedded in graph modeling.

Second, we applied our graph-based approach to a whole brain parcellisation. Here, the histograms are computed to estimate the probability distribution of the grading values within each brain structure as proposed in (Hett et al., 2018c). Thus, for this second anatomical scale of analysis, $GG_{brain} = (V, E, \Gamma, \Omega)$ represents the graph of brain structure grading, where V represents the measures of alteration of brain structures, and E represents the alteration relationship between two brain structures.

3.7. Multi-scale graph-based grading

To combine multiple anatomical scales (for instance, brain structures and hippocampal subfields), we developed a multi-scale graph-based grading (*MGG*) approach based on a cascade of classifiers. In this approach, the graph of brain structures and the graph of hippocampal subfields were computed separately as it is described in the previous sections (see Fig. 2). The elastic net regression method was then used to select the most discriminating features of each graph. Afterward, a first layer of RF classifier was used to compute both *a posteriori* probabilities $P(p_t|X_{GG_{brain}})$ and $P(p_t|X_{GG_{subfields}})$ for whole brain and hippocampal subfields, respectively. As defined in Eq (1), p_t represents the pathological status of the subject under study, while $X_{GG_{brain}}$ and $X_{GG_{subfields}}$ represent the selected features of GG_{brain} and $GG_{subfields}$ models, respectively. Finally, these *a posteriori* probabilities were used as the input of a linear classifier to make the final decision.

In addition of this new method, we also proposed a straightforward extension of our graph-based grading method. This approach results in the concatenation of GG_{brain} and $GG_{subfields}$ features into a single feature vector before the feature selection step.

3.8. Combination with cognitive tests

Previous works have shown that MRI-based biomarkers are complementary to cognitive assessments used in clinical routines (Tong et al., 2017a; Samper-Gonzalez et al., 2019). Therefore, in addition of studying the efficiency of our novel imaging-based biomarkers, a study of the complementarity of our proposed method with cognitive scores has also been conducted. In this work, we have considered different cognitive scores such as MMSE, CDR-SB, RAVLT, FAQ, ADAS11, and ADAS13 cognitive tests. The cognitive scores are concatenated into a vector of cognitive features and normalized by a z-score. Finally, a concatenation of normalized cognitive scores and graph-based features are used as inputs of the final classifier as illustrated in Fig. 2.

3.9. Details of implementation

A fast patch extraction scheme was used to find the most similar patches (Giraud et al., 2016). We used the grading method proposed in (Hett et al., 2018b), with the same parameters for the size of the patches and K_{x_i}. The effect of age was corrected using a linear regression estimated on CN population (Koikkalainen et al., 2012).

The elastic net feature selection was computed with the SLEP package (Liu et al., 2009). The two parameters λ and ρ were set

Table 2
Classification of sMCI versus pMCI. Results obtained by inter-subject similarity features (*i.e.*, vertices), intra-subject variability features (*i.e.*, edges) and a combination of both. The patch-based grading applied on the hippocampus is used as baseline. The experiment shows a slight superiority of the whole brain structures for AD prediction. All results are expressed in terms of percentage.

Methods	AUC	ACC	BACC	SEN	SPE
Hippocampus PBG	76.8 ± 0.2	70.3 ± 0.0	70.6 ± 0.0	69.0 ± 0.0	72.2 ± 0.0
Hipp. all vertices	73.9 ± 0.2	67.1 ± 0.0	67.9 ± 0.0	72.2 ± 0.0	63.5 ± 0.0
Hipp. selected vertices	77.1 ± 0.2	71.1 ± 0.4	71.4 ± 0.4	69.5 ± 0.6	**73.2 ± 0.5**
Hipp. all edges	66.7 ± 0.2	61.1 ± 0.4	62.0 ± 0.4	68.0 ± 0.4	56.0 ± 0.4
Hipp. selected edges	67.9 ± 0.2	63.0 ± 0.4	63.8 ± 0.4	68.9 ± 0.4	58.7 ± 0.4
$GG_{subfields}$	78.2 ± 0.2	74.7 ± 0.4	74.3 ± 0.5	77.1 ± 0.5	71.4 ± 0.9
Brain all vertices	68.2 ± 0.2	65.3 ± 0.4	66.7 ± 0.5	68.6 ± 0.5	62.2 ± 0.5
Brain selected vertices	77.2 ± 0.2	70.1 ± 0.4	71.1 ± 0.5	**77.8 ± 0.5**	64.4 ± 0.5
Brain all edges	67.1 ± 0.2	65.7 ± 0.2	64.8 ± 0.7	69.4 ± 0.2	60.5 ± 0.2
Brain selected edges	76.9 ± 0.2	72.2 ± 0.4	71.9 ± 0.5	73.8 ± 0.5	70.0 ± 0.5
GG_{brain}	**79.4 ± 0.2**	**75.5 ± 0.4**	**75.1 ± 0.5**	77.6 ± 0.5	72.6 ± 0.5

Table 3
Comparisons of the different PBG approaches for the task of classifying sMCI versus pMCI. PBG computed over the hippocampus is provided as a baseline. The results show that the *MGG* approach improves performance in terms of AUC, ACC, BACC, SEN and SPE. All results are expressed in terms of percentages. Non-parametric permutation tests were conducted to assess the differences between mean accuracies of each investigated method.

Methods	AUC	ACC	BACC	SEN	SPE
Hippocampus PBG[1]	76.8 ± 0.2	70.3 ± 0.0	70.6 ± 0.0	69.0 ± 0.0	72.2 ± 0.0
Graph of hippocampal subfields ($GG_{subfields}$)[2]	78.2 ± 0.2	74.7 ± 0.4	74.3 ± 0.4	77.1 ± 0.4	71.4 ± 0.4 $p_{1,2}$=0.0001
Graph of brain structures (GG_{brain})[3]	79.4 ± 0.2	75.5 ± 0.4	75.2 ± 0.4	77.6 ± 0.4	72.6 ± 0.4 $p_{2,3}$=0.0001
Graph of hipp. sub. + brain (GG_{+})[4]	79.6 ± 0.2	74.5 ± 0.4	73.9 ± 0.4	77.3 ± 0.4	70.6 ± 0.4 $p_{2,4}$=0.8
Multi-scale graph-based grading* (*MGG*)[5]	**80.6 ± 0.2**	**76.0 ± 0.4**	**75.7 ± 0.4**	**77.8 ± 0.4**	**73.6 ± 0.4** $p_{3,5}$=0.0001

* Method illustrated in Fig. 2

up with a grid search method conducted within a nested 10-folds cross validation procedure using CN and AD data. Then, the optimal parameters were directly applied for sMCI vs. pMCI classification without further tuning. The classifications based on the two different anatomical scales (*i.e.*, whole brain structures and hippocampal subfields) were obtained using a random forest (RF)[5]. In our experiments, we used the Gini index as impurity criterion. RF has also two parameters, the numbers of three N and the number of randomly selected features T, which were set to $N_{tree} = 500$, and $T = \lfloor log_2(N_{features}) \rfloor$ (Breiman, 2001). A linear discriminant analysis (LDA) classifier was used to compute the final decision for the combination of a posteriori probabilities from graph of brain structures and hippocampal subfields. In addition, a random forest classifier replaces linear classifier for the multimodal experiment to deal with the non-linear nature of feature boundaries resulting from the concatenation of image-based features and cognitive scores. All features were normalized using z-scores before the selection and classification steps.

In our experiments, we performed sMCI versus pMCI and CN versus AD classifications. For sMCI versus pMCI classification, the elastic net feature selection and the classifiers were trained with CN and AD. Indeed, as shown in Tong et al. (2017a), the use of CN and AD to train the feature selection method and the classifier enables to better discrimination between sMCI and pMCI subjects. Furthermore, this technique also limits bias and the overfitting problem. Finally, to estimate the variability of the classification performance, 100 runs were performed. A stratified 10-folds cross-validation procedure was conducted for the comparison of CN versus AD. Mean area under curve (AUC), accuracy (ACC), balanced accuracy (BACC), sensitivity (SEN), and specificity (SPE) are provided for each experiment. A non-parametric permutation test based on Fisher's technique was used to statistically estimate the improvement of classification performance of each anatomical scale and their combinations.

[5] http://code.google.com/p/randomforest-matlab.

4. Results

To evaluate the performance of the graph-based grading method, we first compare the prediction accuracy of the different graph components. Afterwards, we apply our method within the hippocampal subfields and the whole brain structures (see Table 2), and evaluate the proposed approach to combine different anatomical scales (see Table 3). Then, we evaluate the complementarity of our image-based biomarker and the cognitive scores that are usually used in clinical routines (see Table 4). Finally, we compare the performance of our method with state-of-the-art methods for early detection of Alzheimer's disease (see Tables 5 and 6).

4.1. Graph of hippocampal subfields

First, we compared each element of our graph of structure grading within the hippocampal subfields (see Table 2). As previously proposed in (Hett et al., 2018b), the PBG applied within the whole hippocampus is used as baseline for this experiment.

PBG based on the whole hippocampus structure obtains 76.8% of AUC, 70.3% of ACC and is more specific than sensitive. Although PBG values of all hippocampal subfields (see "all" in the Table 2) do not improve prediction performances, PBG values within selected vertices (*i.e.*, subiculum, CA1-SP, and CA1-SRLM) obtain 77.1% of AUC, 71.1% of ACC (see "selected" in the Table 2), and improve the specificity in comparison to hippocampus grading. Thus, the use of

Table 4
Comparison of our graph-based approach with cognitive test scores and combination of both for AD prediction (*i.e.*, sMCI versus pMCI comparison). Although our MSGG obtains better results in terms of AUC, ACC, BACC, and SPE, the results of this comparison demonstrate the complementarity of our imaging-based method with cognitive scores. All results are expressed as percentage.

Methods	AUC	ACC	BACC	SEN	SPE
Cognitive score	78.8 ± 0.2	74.5 ± 0.4	72.4 ± 0.4	84.9 ± 0.4	60.0 ± 0.4
MGG	80.6 ± 0.2	76.0 ± 0.4	75.7 ± 0.4	77.8 ± 0.4	**73.6 ± 0.4**
MGG + Cognitive score	**85.5 ± 0.2**	**80.6 ± 0.4**	**79.2 ± 0.4**	**87.3 ± 0.4**	71.1 ± 0.4

Table 5
Comparison with state-of-the-arts methods for Alzheimer's disease classification using similar ADNI1 dataset. In addition to sMCI versus pMCI, we provided results of CN versus AD classification. All results are expressed in percentage of accuracy (ACC) and balanced accuracy (BACC). Best balanced accuracy for each comparison is presented in bold font.

Methods	Subjects				CN vs. AD		sMCI vs. pMCI	
	CN	sMCI	pMCI	AD	ACC	BACC	ACC	BACC
Patch-based grading (Coupé et al., 2012b)	231	238	167	198	88.0	87.5	71.0	71.0
Sparse ensemble grading (Liu et al., 2012)	229	*n.a*	*n.a*	198	90.8	90.5	*n.a*	*n.a*
Voxel-based morphometry (Moradi et al., 2015)	231	100	164	200	*n.a*	*n.a*	74.7	70.2
Sparse-based grading (Tong et al., 2017a)	229	129	171	191	*n.a*	*n.a*	75.0	*n.a*
Multiple ensemble learning (Tong et al., 2014)	231	238	167	198	89.0	89.5	70.4	71.5
Deep ensemble learning (Suk et al., 2017)	226	226	167	186	91.0	91.3	74.8	74.9
Hierarchical network (Lian et al., 2018)	229	226	167	199	90.3	89.4	**80.9**	69.0
Deep neural network (Basaia et al., 2018)	352	510	253	295	**99.2**	**99.2**	75.1	75.0
Cortical graph network (Wee et al., 2019)	242	*n.a*	*n.a*	355	85.8	85.5	*n.a*	*n.a*
Proposed method	213	90	126	130	91.6	91.4	76.0	**75.7**

Table 6
Comparison of the different combination of different imaging biomarkers CSF, and demographic data used in clinical routines for the prediction of MCI conversion (*i.e.*, sMCI versus pMCI comparison). All results are expressed as percentage. Best AUC is expressed in bold font.

Methods	Source	AUC	ACC
Latent feature representation (Suk et al., 2015)	MRI + PET + CSF	n.a	**83.3**
Combined sparse-based grading (Tong et al., 2017a)	MRI + Cognitive scores[a]	87.0	80.7
Voxel-wise approach (Samper-Gonzalez et al., 2019)	MRI + FDG-PET + Cognitive score[b]	**88.5**	80.9
Multimodal deep learning approach (Lee et al., 2019)	MRI + CSF + Cognitive scores[c]	n.a	76.0
Proposed	MRI + Cognitive scores[d]	85.5	80.6

[a] FAge, MMSE, CDR-sb, RAVLT, ADAS.
[b] Gender, MMSE, Education level, CDR-sb, RAVLT, ADAS.
[c] ADNI-EF, ADNI-MEM.
[d] MMSE, CDR-sb, RAVLT, FAQ, ADAS11, ADAS13.

hippocampal subfields selected with the elastic net method slightly increases the prediction performance of AD compared to the union of all subfields or the whole hippocampus. Furthermore, the edges selected by the elastic net do not improve the prediction performance compared to other hippocampal features. Finally, the proposed method combining edges and vertices improves the AUC by 1.4 percent points and the accuracy 4.4 percent points compared to the global hippocampus grading. Our graph-based method also improves the AUC by 1.1 percent points and the accuracy by 3.6 percent points when compared to the use of the most discriminant hippocampal subfields. Moreover, in both cases, our proposed graph-based method has a higher sensitivity.

Fig. 3-B illustrates the contribution (*i.e.*, the number of selection by the elastic net) for each hippocampal subfield in the graph-based features vectors after the feature selection step. The experiments have shown that the most discriminant hippocampal subfields selected are the subiculum, and the two subfields representing the CA1. This is particularly interesting because hippocampal subfields selected by the elastic net regression method are in line with previous studies, which have shown that the CA1 and subiculum are the subfields with the most significant atrophy. These founding were shown in studies that analyses patients in late stages of AD (Kerchner et al., 2012; Trujillo-Estrada et al., 2014), and studies that analyze the hippocampal subfields at the early stage of the disease (Hett et al., 2019; Parker et al., 2019).

4.2. Graph of brain structures

We also conducted an evaluation of graph-based grading over the whole brain similar to that of the hippocampus where we individually estimated the performance obtained by each type of feature separately (see Table 2). The use of all vertices (*i.e.*, the averages of PBG values computed within each brain structure) decreases the prediction performance compared to the use of only the hippocampus (65.3% compared to 70.3% of accuracy). A selection of the most discriminating vertices obtains similar results to those of the hippocampus only with an accuracy of 70.1%. Contrary to the hippocampal subfields where vertices were most efficient than edges, the use of edge features performs similarly to the vertices.

As shown with the hippocampal subfields, the combination of both features, edges and vertices, that capture the inter-subjects' similarities and intra-subject variability enables an important increase of prediction performance. Our method applied using the brain structures obtains 75.5% accuracy and 79.4% AUC. Moreover, the experiments also show a sensitivity similar to using only selected vertices and a higher specificity than using only selected edges.

Fig. 3-A illustrates the most selected brain structures during the feature selection step. The experiments have shown that the most frequently selected brain structures are the temporal lobe, the

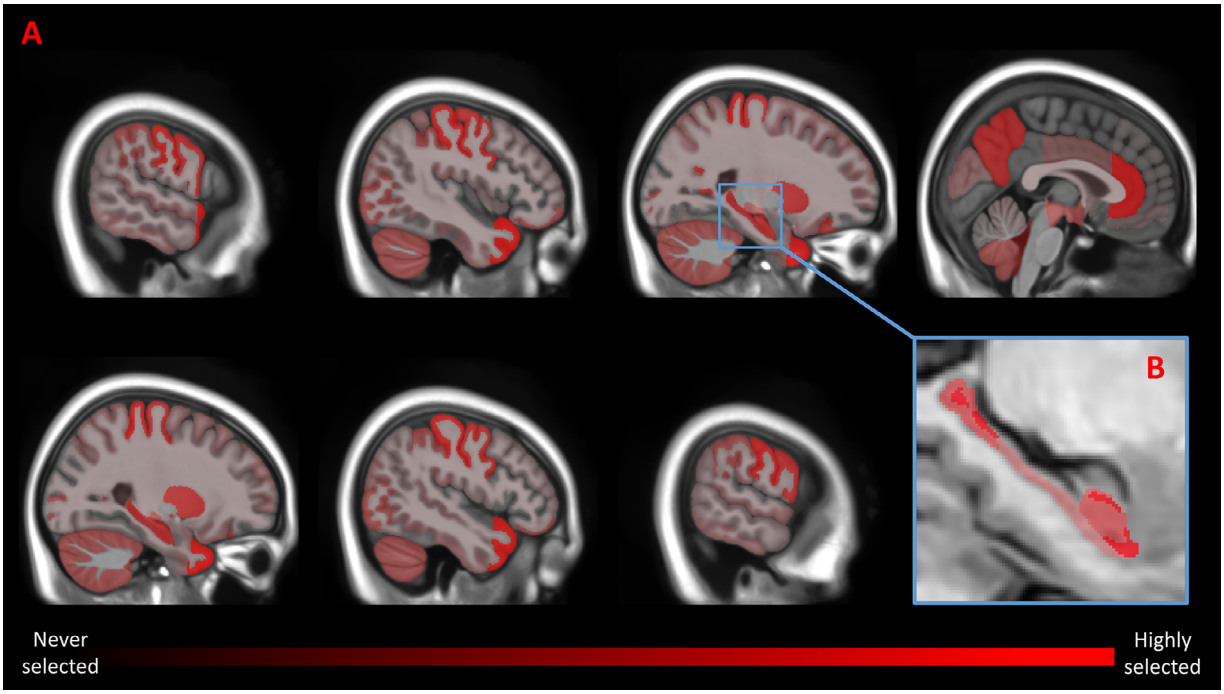

Fig. 3. Representation of the most selected structures. The brain structures and hippocampal subfields are selected separately with the elastic net method. Frequently selected structures are colored using opaque red to transparent for structures never selected. (A) the most frequently selected brain structures are the temporal lobe, the postcentral gyrus, the anterior cingulate gyrus, the hippocampus and the precuneus. (B) the most frequently selected hippocampal subfields are the CA1-SP, the CA1-SRLM, and the subiculum. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

postcentral gyrus, the anterior cingulate gyrus, the hippocampus and the precuneus. It is also interesting, as the results obtained from the hippocampal subfields, the most selected brain structures are in line with clinical studies that show a relationship between the atrophy of specific brain structures (Busatto et al., 2003; Hyman et al., 1984; Kogure et al., 2000; Frisoni et al., 2002; Apostolova et al., 2007).

4.3. Multiscale graph-based grading

Table 3 provides a comparison of prediction performances obtained with our graph-based grading method applied in each anatomical scale independently and the combination of both. In this experiment, two approaches were investigated.

First, the results of this comparison confirm that for sMCI versus pMCI classification, whole-brain analysis enables better performance than analysis of the hippocampus subfields (p=0.0001). Indeed, GG_{brain} (whole brain) obtains 79.4% of AUC and 75.5% of accuracy while $GG_{subfields}$ (hippocampus subfields) obtains 78.2% of AUC and 74.7% of accuracy.

Second, we compare the two approaches of combining both anatomical scales (*i.e.*, simple concatenation or cascade of classifiers). These results suggest that the straightforward concatenation of the feature vectors from GG_{brain} and $GG_{subfields}$ methods does not improve the performance compared to GG_{brain} and $GG_{subfields}$. Indeed, the concatenation of the feature vectors obtains 79.6% of AUC and 74.5% of ACC, which is lower than the results obtained from the use of whole brain structures. However, the multi-scale graph-based approach (*MGG*) method (see Fig. 2) shows increased performance for each considered measure of classification (p=0.0001). This last method obtains 80.6% of AUC and 76% of accuracy. This indicates that the analysis of hippocampal subfields and whole brain structures are complementary. Therefore, in the rest of the experiments, we only consider the *MGG* method in our comparisons.

4.4. Complementarity with cognitive tests

Table 4 presents a comparison of the results obtained using features derived from cognitive tests, our imaging-based method, and the combination of both. This comparison demonstrates that our imaging-based method obtains better results than using cognitive scores. Indeed, *MGG* improves the sMCI versus pMCI classification by 1.8 percent point of AUC and 1.5 percent point of accuracy compared to using cognitive scores only.

Moreover, the results of the experiment indicate the complementarity of imaging-based and cognitive assessments for AD prediction. Thus, the combination of cognitive scores and *MGG* features obtains 85.5% AUC and 80.6% accuracy which improves AUC by 4.9% and improves accuracy by 4.6% when compared to the *MGG* method.

4.5. Comparison with state-of-the-art methods

To compare to state-of-the-art methods, we evaluate our findings against those obtained using MRI-based methods in similar ADNI datasets (see Table 5), and to a set of multi-modal methods (see Table 6). Besides cognitive assessments, the presented methods involved cerebral spinal fluid biomarkers (CSF), positron emission tomography (PET), and fluorodeoxyglucose PET (FDG-PET).

Firstly, *MGG* is compared with state-of-the-art methods using a similar ADNI1 dataset. Our graph-based method is compared with the original PBG method (Coupé et al., 2012b), a graph-based grading method (Tong et al., 2014), an ensemble grading method (Liu et al., 2012), a sparse-based grading method (Tong et al., 2017a), a VBM method (Moradi et al., 2015) and advanced approaches based on deep ensemble learning technique (Suk et al., 2017; Lian et al., 2018; Basaia et al., 2018; Wee et al., 2019).

The results of these comparisons demonstrate the competitive performance of our *MGG* method for CN versus AD and sMCI versus pMCI classifications. Indeed, our method obtains state-of-the-art results with 91.6% of accuracy for CN versus AD which are comparable to the most recent method based on deep-learning techniques. Furthermore, our method also obtains state-of-the-art performances for sMCI versus pMCI classification with 76.0% of accuracy. These results are competitive with recent approaches based on deep-learning methods (Suk et al., 2017; Lian et al., 2018; Basaia et al., 2018; Wee et al., 2019). Moreover, our multi-scale graph-based method improves accuracy by 3.6 and 5 percentage points of the original PBG method for CN versus AD and sMCI versus pMCI classification, respectively (Coupé et al., 2012b).

As presented in Table 6, the combination of *MGG* and cognitive scores was compared with state-of-the-art multimodal approaches. This comparison includes a method combining structural MRI and cognitive scores that obtains 80.7% of accuracy (Tong et al., 2017a), a method combining MRI, PET scans and CSF that obtains 83.3% of accuracy (Suk et al., 2015), a voxel-wise approach that combines MRI, FDG-PET and cognitive scores that obtains 80.9% of ACC (Samper-Gonzalez et al., 2019), and a recent multimodal deep-learning approach combining MRI, CSF and cognitive scores that obtains 76% of accuracy (Lee et al., 2019). This demonstrates the competitive performance of our graph-based approach that obtains state-of-the-art results with only the use of MRI-based and cognitive score features.

5. Discussion

The first contribution of this paper is the development of a new graph-based grading approach that combines inter-subject similarities and intra-subject variability efficiently. We validated this new method with two different anatomical scales: the hippocampal subfields and the whole brain structures. The second contribution is the development of an anatomical scale fusion based on a cascade of classifiers approach. We applied this multi-scale graph-based grading framework to the hippocampal subfields and a parcellation of the entire brain structures. To validate our new multi-scale graph-based grading framework, we compared each component of our graph at each anatomical scale. Then, we compared the results obtained in our experiments with the results of state-of-the-art methods proposed in the literature. Finally, we compared the results obtained with our imaging-based biomarker with a bank of cognitive scores that are used in clinical routines.

5.1. Graph of hippocampal subfields

Postmortem and *in-vivo* studies have suggested that the first regions of the brain which are changed in typical disease progression are the entorhinal cortex (EC) and the hippocampus (Jack et al., 1992; Braak and Braak, 1995; Bobinski et al., 1999). Neuroimaging studies have further shown that the hippocampus undergoes the most significant alterations in the early stage of AD (Frisoni et al., 2010; Schwarz et al., 2016). However, recent methods applied to the hippocampus have shown limited performances for AD prediction (Hett et al., 2017; Tong et al., 2017a). This limitation could stem from global analysis of the hippocampus, which is divided into heterogeneous subfields. The terminology differs across segmentation protocols (Yushkevich et al., 2015) but the most recognized definition (Lorente de Nó, 1934) divides the hippocampus into the subiculum, the cornu ammonis (CA1/2/3/4), and the dentrate gyrus (DG). Studies have shown that hippocampal subfields are not equally impacted by AD (Braak and Braak, 1997; Braak et al., 2006; Apostolova et al., 2006; La Joie et al., 2013; Kerchner et al., 2010; 2012). Specifically, postmortem, animal-based and recent *in-vivo* imaging studies showed that the CA1 and the subiculum are the subfields impacted by the most discriminant atrophy in the last stage of AD (Apostolova et al., 2006; La Joie et al., 2013; Kerchner et al., 2012; Li et al., 2013; Trujillo-Estrada et al., 2014; Hett et al., 2019). This indicates that the analysis of the hippocampus with a global measure could limit prediction performance and that better modeling of the structural alterations within the hippocampal subfields could improve prediction performance.

Consequently, we proposed to better model hippocampus alterations with the application of our novel graph-based framework within the hippocampal subfields. First, we studied the efficiency of a straightforward approach that computes the average of grading values in each hippocampus subfield separately instead of the whole hippocampus structure as it is usually done. This results in poorer performance compared to the average of grading values within the whole hippocampus. However, the grading values within the most discriminant hippocampal subfields (*i.e.*, subiculum and the two definitions of CA1) obtain similar performances to the average of grading values within the whole hippocampus. This is possibly be due to the fact that the subiculum and CA1 represent the major part of the hippocampus.

The related hippocampal subfield features selected by the elastic net are consistent with previous *in-vivo* imaging studies, which are based on 3T MRI and ultra-high field MRI at 7T. These studies analyzed the atrophy of each hippocampal subfield at an advanced stage of AD. These studies showed that CA1 is the subfield with the most severe atrophy (Apostolova et al., 2006; Mueller et al., 2007; La Joie et al., 2013; Carlesimo et al., 2015; Hett et al., 2019), and also indicate that CA1SR-L-M is the subfield with the greatest atrophy at advanced stages of AD (Kerchner et al., 2010; 2012). It is interesting to note that the results of our experiments are also in accordance with previous postmortem, animal-based, and *in-vivo* studies combining volume and diffusivity MRI. These last studies demonstrated that the subiculum is the earliest hippocampal region affected by AD (Trujillo-Estrada et al., 2014; Li et al., 2013).

Finally, the great improvement obtained with the combination of inter-subject similarities and intra-subject variability shows that this information is complementary. It also confirms this combination enables the obtention of results similar to methods based on whole brain analysis with only the use of the hippocampus.

5.2. Graph of brain structures

Next, we investigated our method at the whole brain scale. The comparison of hippocampus PBG and the most discriminant vertices indicate that the straightforward combination of other discriminant brain structures does not increase the prediction performance compared to using only the hippocampus. Moreover, when the edges and the vertices are combined, our experiments show that the edges are the most discriminant selected elements.

Our experiments indicate that the most selected brain structures are the postcentral gyrus, the anterior cingulate gyrus, the hippocampus, and the precuneus (see Fig. 3), which align with current literature. First, it is interesting to note that the most discriminant features obtained by the sparse selection method show the importance of the temporal lobe and the hippocampus. Indeed, studies have shown a significant loss of gray matter within the temporal lobe (Killiany et al., 1993; Busatto et al., 2003), while the hippocampus has long been known as the structure with the earliest alterations (Hyman et al., 1984; West et al., 1994; Braak and Braak, 1995; Ledig et al., 2018). Second, VBM and perfusion studies have shown that the precuneus suffers from a noticeable atrophy and a bilateral decrease of regional cerebral blood flow compared to control subjects (Kogure et al., 2000; Karas et al., 2007). Studies have shown a significant reduction in volume of the anterior cingulate gyrus compared to control (Frisoni et al., 2002; Jones et al., 2006). Additionally, a study showed that the volume of the

anterior cingulate gyrus is correlated with apathy which is symptomatic of AD (Apostolova et al., 2007). However, the importance of the postcentral gyrus was unexpected since it has been shown that this structure seems unaffected by AD process (Halliday et al., 2003). These elements seem to indicate that the structural pattern of AD is composed of both highly impacted and healthy brain structures.

Finally, the good performance of the graph-based grading method demonstrates that the combination of both features enables a better discrimination of subjects who convert to dementia in the years following their first visits. The good results within hippocampal subfield and brain structure parcellation indicates that our framework can be applied with different anatomical representation.

5.3. Multi-scale graph-based grading

Afterwards, we compared the results of our multi-scale (*MGG*) approach with the previously described GG_{brain} and $GG_{subfields}$. First, the conducted experiments show that our graph of structure grading applied within hippocampal subfields improves prediction of conversion to Alzheimer's disease compared to the PBG applied within the hippocampus.

The results obtained by the straightforward extension of the graph of structure grading to combine whole brain structure and hippocampal subfields did not demonstrate an improvement in AD conversion prediction compared to the single use of GG_{brain} and $GG_{subfields}$. The main limitation might come from the fact that the straightforward combination of different anatomical representations suffer from a substantial augmentation of feature dimensionality. To address these limitations, we proposed the *MGG* method that is based on a cascade of classifiers. This method alleviates the dimensionality issue by estimating an intermediate conversion probability for each anatomical scale considered. This results in an increase in AD prediction performances compared to GG_{brain} and $GG_{subfields}$ methods.

5.4. Comparison with state-of-the-art methods

In this last decade, many improvements in computer-aided diagnosis methods were proposed to better capture structural alterations using anatomical MRI (see (Rathore et al., 2017) for a review). Two main approaches were proposed: methods based on inter-subject similarity (Coupé et al., 2012b; Moradi et al., 2015; Tong et al., 2017a) and methods based on intra-subject variability (Tong et al., 2014; Suk et al., 2014). Consequently, the first contribution of our work was to combine inter-subject similarity – using the PBG framework – and the intra-subject variability – with the integration of PBG into a graph-based model. Indeed, our graph-based grading can obtain competitive results with different anatomical representations.

Another difference with state-of-the-art methods comes from the proposition of a multi-anatomical scale analysis of AD alterations. In contrast to previous methods which analyzed changes at a unique anatomical scale (*i.e.*, cortical cortex, whole brain structures, hippocampus, or hippocampal subfields,...), we proposed combining whole brain structures parcellation with a representation of hippocampal subfields. This combination has resulted in performances competitive with state-of-the-art methods.

Finally, the comparison with state-of-the-art approaches using similar ADNI1 subset has shown that our multi-scale graph-based grading method obtains competitive results for both AD detection and prediction. The high performances of the methods proposed in (Basaia et al., 2018) and (Lian et al., 2018) have to be moderated. The high accuracy for AD detection obtained by Basaia et al. (2018) raises two concerns. First, it has been reported that the majority of deep-learning methods that obtain almost perfect detection performances suffer from data-leakage problems (Wen et al., 2019a; 2019b). Second, and more importantly, the diagnoses provided by ADNI were established using a set of cognitive-scores and functional performances, which results in diagnosis error (Beach et al., 2012; Matias-Guiu et al., 2017) since AD can only be diagnosed for sure after a postmortem analysis revealing the two cerebral hallmarks of the disease (*i.e.*, beta amyloid deposition and neurofibrillary tangles) (Cairns et al., 2010). This last element raises a question about the clinical relevance of such results.

Finally, the excellent result of Lian et al. (2018) for AD prediction mainly takes advantage its unbalanced nature (*i.e.*, 52.9% of sensitivity and 85.9% of specificity). This results in overly high classification accuracy, which is not representative of the overall performance provided by this method. Consequently, the balanced accuracy was used to fairly compare the classification performance with other methods (see Table 5).

5.5. Complementarity with cognitive tests

Finally, an analysis of the complementarity of our imaging-based method with scores resulting from cognitive assessments was carried out. These experiments enabled the comparison of the performance of cognitive scores and our imaging biomarker for AD prediction.

The conducted experiments demonstrate that our graph-based grading approach using T1 weighted MRI obtains substantially better results for the prediction of AD than the single use of cognitive scores. Moreover, as shown in many works listed in Table 6 (Suk et al., 2015; Tong et al., 2017a; Samper-Gonzalez et al., 2019; Lee et al., 2019), MRI-based biomarkers and cognitive assessments are complementary, and their combination improves classification performances. Thus, the combination of our graph-based grading technique and cognitive assessments demonstrates a great improvement in performance compared to the use of each method separately. This improvement is comparable to studies based on multi-modality frameworks, which use more expensive biomarkers (*i.e.*, PET, FDG-PET,...), and can be invasive (*i.e.*, CSF). These elements complicate the implementation of such multi-modal features in clinical routines.

5.6. Strength and limitations

In addition to the strengths and limitations of the patch-based grading framework, the major strength of the proposed method comes from the ability to efficiently combine intra-subject variability and inter-subject similarity in a common model that can be applied at different anatomical scales. Nonetheless, we acknowledge that the proposed multi-scale graph-based grading framework is not without potential limitations. The main limitation come from the dependence of our method to the quality of segmentation maps that are used to aggregate patch-based grading and estimate abnormality of each structure.

6. Conclusion

Improved modeling of AD alterations is a great challenge that could lead to earlier predictions of conversion. Therefore, in this work, we developed a new method to better model AD signature. Our proposed method models the pattern of AD alterations by combining inter-subject similarity and intra-subject variability. The conducted experiments have shown that our framework can be applied with different anatomical representations. Consequently, we proposed a multi-anatomical scale graph-based grading method to

combine the alterations at different anatomical scales. In addition, we conducted the first joint analysis of the hippocampus subfields and brain structure changes in the same framework. The results show state-of-the-art-performance, confirming the complementarity of hippocampal subfields and whole brain analysis, and the complementarity of inter-subject similarity and intra-subject variability.

Authorship contributions

Kilian Hett: Conceptualization, Validation, Formal analysis, Writing - original draft, Writing - review and editing:. **José V. Manjon**: Conceptualization, Validation, Formal analysis, Writing - original draft, Writing - review and editing, Supervision. **Vinh-Thong Ta**: Conceptualization, Validation, Formal analysis, Writing - original draft, Writing - review and editing, Supervision. **Pierrick Coupé**: Conceptualization, Validation, Formal analysis, Writing - original draft, Writing - review and editing, Supervision, Funding acquisition. **Ipek Oguz:** Supervision, Writing - review and editing, Funding acquisition. The data used in this manuscript is obtained from Alzheimer's Disease Neuroimaging Initiative (ADNI). 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.

Declaration of Competing Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could result as a potential conflict of interest.

Acknowledgement

This work benefited from the support of the project DeepvolBrain of the French National Research Agency (ANR-18-CE45-0013). This study was achieved within the context of the Laboratory of Excellence TRAIL ANR-10-LABX-57 for the BigDataBrain project. Moreover, we thank the Investments for the future Program IdEx Bordeaux (ANR-10-IDEX-03-02, HL-MRI Project), Cluster of excellence CPU and the CNRS. Finally, this work was also supported by the NIH grants R01–NS094456 and U01–NS106845. 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 by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Biogen; Bristol-Myes Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffman-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Pharmaceutical Research & Development LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical 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 of Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

References

Apostolova, L.G., Akopyan, G.G., Partiali, N., Steiner, C.A., Dutton, R.A., Hayashi, K.M., Dinov, I.D., Toga, A.W., Cummings, J.L., Thompson, P.M., 2007. Structural correlates of apathy in alzheimer's disease. Dement Geriatr Cogn Disord 24, 91.

Apostolova, L.G., Dutton, R.A., Dinov, I.D., Hayashi, K.M., Toga, A.W., Cummings, J.L., Thompson, P.M., 2006. Conversion of mild cognitive impairment to alzheimer disease predicted by hippocampal atrophy maps. Arch. Neurol. 63, 693–699.

Arbabshirani, M.R., Plis, S., Sui, J., Calhoun, V.D., 2017. Single subject prediction of brain disorders in neuroimaging: promises and pitfalls. Neuroimage 145, 137–165.

Ashburner, J., Friston, K.J., 2000. Voxel-based morphometry—the methods. Neuroimage 11, 805–821.

Alzheimer's Association, 2015. 2015 Alzheimer's disease facts and figures. Alzheimer's & dementia: the journal of the Alzheimer's Association 11, 332.

Avants, B.B., Tustison, N.J., Song, G., Cook, P.A., Klein, A., Gee, J.C., 2011. A reproducible evaluation of ANTs similarity metric performance in brain image registration. Neuroimage 54, 2033–2044.

Basaia, S., Agosta, F., Wagner, L., Canu, E., Magnani, G., Santangelo, R., Filippi, M., Initiative, A.D.N., et al., 2018. Automated classification of alzheimer's disease and mild cognitive impairment using a single mri and deep neural networks. NeuroImage: Clinical 101645.

Beach, T.G., Monsell, S.E., Phillips, L.E., Kukull, W., 2012. Accuracy of the clinical diagnosis of alzheimer disease at national institute on aging alzheimer disease centers, 2005–2010. J. Neuropathol. Exp. Neurol. 71, 266–273.

Bobinski, M., De Leon, M.J., Convit, A., De Santi, S., Wegiel, J., Tarshish, C.Y., Saint Louis, L., Wisniewski, H.M., 1999. Mri of entorhinal cortex in mild alzheimer's disease. The Lancet 353, 38–40.

Braak, E., Braak, H., 1997. Alzheimer's disease: transiently developing dendritic changes in pyramidal cells of sector CA1 of the Ammon's horn. Acta Neuropathol. 93, 323–325.

Braak, H., Alafuzoff, I., Arzberger, T., Kretzschmar, H., Del Tredici, K., 2006. Staging of Alzheimer disease-associated neurofibrillary pathology using paraffin sections and immunocytochemistry. Acta Neuropathol. 112, 389–404.

Braak, H., Braak, E., 1995. Staging of alzheimer's disease-related neurofibrillary changes. Neurobiol. Aging 16, 271–278.

Breiman, L., 2001. Random forests. Mach Learn 45, 5–32.

Bron, E.E., Smits, M., Van Der Flier, W.M., Vrenken, H., Barkhof, F., Scheltens, P., Papma, J.M., Steketee, R.M., Orellana, C.M., Meijboom, R., 2015. Standardized evaluation of algorithms for computer-aided diagnosis of dementia based on structural MRI: the CADDementia challenge. Neuroimage 111, 562–579.

Busatto, G.F., Garrido, G.E., Almeida, O.P., Castro, C.C., Camargo, C.H., Cid, C.G., Buchpiguel, C.A., Furuie, S., Bottino, C.M., 2003. A voxel-based morphometry study of temporal lobe gray matter reductions in alzheimer's disease. Neurobiol. Aging 24, 221–231.

Cairns, N.J., Taylor-Reinwald, L., Morris, J.C., Initiative, A.D.N., et al., 2010. Autopsy consent, brain collection, and standardized neuropathologic assessment of adni participants: the essential role of the neuropathology core. Alzheimer's & Dementia 6, 274–279.

Carlesimo, G.A., Piras, F., Orfei, M.D., Iorio, M., Caltagirone, C., Spalletta, G., 2015. Atrophy of presubiculum and subiculum is the earliest hippocampal anatomical marker of alzheimer's disease. Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring 1, 24–32.

Coupé, P., Eskildsen, S.F., Manjón, J.V., Fonov, V.S., Collins, D.L., disease Neuroimaging Initiative, A., et al., 2012. Simultaneous segmentation and grading of anatomical structures for patient's classification: application to alzheimer's disease. Neuroimage 59, 3736–3747.

Coupé, P., Eskildsen, S.F., Manjón, J.V., Fonov, V.S., Pruessner, J.C., Allard, M., Collins, D.L., Initiative, A.D.N., 2012. Scoring by nonlocal image patch estimator for early detection of Alzheimer's disease. NeuroImage: clinical 1, 141–152.

Coupé, P., Fonov, V.S., Bernard, C., Zandifar, A., Eskildsen, S.F., Helmer, C., Manjón, J.V., Amieva, H., Dartigues, J.-F., Allard, M., 2015. Detection of Alzheimer's disease signature in MR images seven years before conversion to dementia: toward an early individual prognosis. Hum Brain Mapp 36, 4758–4770.

Coupé, P., Manjón, J.V., Chamberland, M., Descoteaux, M., Hiba, B., 2013. Collaborative patch-based super-resolution for diffusion-weighted images. Neuroimage 83, 245–261.

Coupé, P., Manjón, J.V., Lanuza, E., Catheline, G., 2019. Lifespan changes of the human brain in Alzheimer's disease. Sci Rep 9, 3998.

Cuingnet, R., Glaunès, J.A., Chupin, M., Benali, H., Colliot, O., 2013. Spatial and anatomical regularization of svm: a general framework for neuroimaging data. IEEE Trans Pattern Anal Mach Intell 35, 682–696.

DeCarli, C., 2003. Mild cognitive impairment: prevalence, prognosis, aetiology, and treatment. The Lancet Neurology 2, 15–21.

Duthey, B., 2013. Background paper 6.11: Alzheimer disease and other dementias. A Public Health Approach to Innovation 1–74.

Frisoni, G., Testa, C., Zorzan, A., Sabattoli, F., Beltramello, A., Soininen, H., Laakso, M., 2002. Detection of grey matter loss in mild Alzheimer's disease with voxel based morphometry. Journal of Neurology, Neurosurgery & Psychiatry 73, 657–664.

Frisoni, G.B., Fox, N.C., Jack, C.R., Scheltens, P., Thompson, P.M., 2010. The clinical use of structural MRI in Alzheimer disease. Nature Reviews Neurology 6, 67–77.

Giraud, R., Ta, V.-T., Papadakis, N., Manjón, J.V., Collins, D.L., Coupé, P., Initiative, A.D.N., 2016. An optimized patchmatch for multi-scale and multi-feature label fusion. Neuroimage 124, 770–782.

Halliday, G., Double, K., Macdonald, V., Kril, J., 2003. Identifying severely atrophic cortical subregions in Alzheimer's disease. Neurobiol. Aging 24, 797–806.

Hardy, J., 2006. Alzheimer'S disease: the amyloid cascade hypothesis: an update and reappraisal. J. Alzheimers Dis. 9, 151–153.

Hett, K., Ta, V.-T., Catheline, G., Tourdias, T., Manjón, J.V., Coupe, P., 2019. Multimodal hippocampal subfield grading for Alzheimer's disease classification. Sci Rep 9, 1–16.

Hett, K., Ta, V.-T., Manjón, J. V., Coupé, P., 2018a. Graph of hippocampal subfields grading for Alzheimer's disease prediction. Springer. International Workshop on Machine Learning in Medical Imaging, 259–266.

Hett, K., Ta, V.-T., Manjón, J. V., Coupé, P., Initiative, A. D. N., 2017. Adaptive fusion of texture-based grading: Application to Alzheimer's disease detection. Springer. International Workshop on Patch-based Techniques in Medical Imaging, 82–89.

Hett, K., Ta, V.-T., Manjón, J.V., Coupé, P., Initiative, A.D.N., et al., 2018. Adaptive fusion of texture-based grading for alzheimer's disease classification. Computerized Medical Imaging and Graphics 70, 8–16.

Hett, K., Ta, V.-T., Manjón, J. V., Coupé, P., Initiative, A. D. N., et al., 2018c. Graph of brain structures grading for early detection of Alzheimer's disease. Springer. International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 429–436,

Hyman, B.T., Van Hoesen, G.W., Damasio, A.R., Barnes, C.L., 1984. Alzheimer's disease: cell-specific pathology isolates the hippocampal formation. Science 225, 1168–1171.

Jack, C.R., Petersen, R.C., O'brien, P.C., Tangalos, E.G., 1992. Mr-based hippocampal volumetry in the diagnosis of Alzheimer's disease. Neurology 42. 183–183

Jack Jr, C.R., Bennett, D.A., Blennow, K., Carrillo, M.C., Dunn, B., Haeberlein, S.B., Holtzman, D.M., Jagust, W., Jessen, F., Karlawish, J., et al., 2018. Nia-aa research framework: toward a biological definition of Alzheimer's disease. Alzheimer's & Dementia 14, 535–562.

Jones, B.F., Barnes, J., Uylings, H.B., Fox, N.C., Frost, C., Witter, M.P., Scheltens, P., 2006. Differential regional atrophy of the cingulate gyrus in alzheimer disease: a volumetric MRI study. Cerebral Cortex 16, 1701–1708.

Karas, G., Scheltens, P., Rombouts, S., Van Schijndel, R., Klein, M., Jones, B., Van Der Flier, W., Vrenken, H., Barkhof, F., 2007. Precuneus atrophy in early-onset Alzheimer's disease: a morphometric structural MRI study. Neuroradiology 49, 967–976.

Kerchner, G., Hess, C., Hammond-Rosenbluth, K., Xu, D., Rabinovici, G., Kelley, D., Vigneron, D., Nelson, S., Miller, B., 2010. Hippocampal CA1 apical neuropil atrophy in mild Alzheimer disease visualized with 7-t MRI. Neurology 75, 1381–1387.

Kerchner, G.A., Deutsch, G.K., Zeineh, M., Dougherty, R.F., Saranathan, M., Rutt, B.K., 2012. Hippocampal CA1 apical neuropil atrophy and memory performance in Alzheimer's disease. Neuroimage 63, 194–202.

Killiany, R.J., Moss, M.B., Albert, M.S., Sandor, T., Tieman, J., Jolesz, F., 1993. Temporal lobe regions on magnetic resonance imaging identify patients with early alzheimer's disease. Arch. Neurol. 50, 949–954.

Kogure, D., Matsuda, H., Ohnishi, T., Asada, T., Uno, M., Kunihiro, T., Nakano, S., Takasaki, M., 2000. Longitudinal evaluation of early alzheimer's disease using brain perfusion spect. J. Nucl. Med. 41, 1155–1162.

Koikkalainen, J., Pölönen, H., Mattila, J., Van Gils, M., Soininen, H., Lötjönen, J., Initiative, A.D.N., et al., 2012. Improved classification of alzheimer's disease data via removal of nuisance variability. PLoS ONE 7, e31112.

La Joie, R., Perrotin, A., De La Sayette, V., Egret, S., Doeuvre, L., Belliard, S., Eustache, F., Desgranges, B., Chételat, G., 2013. Hippocampal subfield volumetry in mild cognitive impairment, Alzheimer's disease and semantic dementia. NeuroImage: Clinical 3, 155–162.

Ledig, C., Schuh, A., Guerrero, R., Heckemann, R.A., Rueckert, D., 2018. Structural brain imaging in alzheimer's disease and mild cognitive impairment: biomarker analysis and shared morphometry database. Sci Rep 8, 11258.

Lee, G., Nho, K., Kang, B., Sohn, K.-A., Kim, D., 2019. Predicting alzheimer's disease progression using multi-modal deep learning approach. Sci Rep 9, 1952.

Li, Y.-D., Dong, H.-B., Xie, G.-M., Zhang, L.j., 2013. Discriminative analysis of mild Alzheimer's disease and normal aging using volume of hippocampal subfields and hippocampal mean diffusivity: an in vivo magnetic resonance imaging study. American Journal of Alzheimer's Disease & Other Dementias 28, 627–633.

Lian, C., Liu, M., Zhang, J., Shen, D., 2018. Hierarchical fully convolutional network for joint atrophy localization and Alzheimer's disease diagnosis using structural mri. IEEE Trans Pattern Anal Mach Intell.

Liu, J., Ji, S., Ye, J., et al., 2009. SLEP: Sparse learning with efficient projections. Arizona State University 6, 7.

Liu, M., Zhang, D., Shen, D., Initiative, A.D.N., 2012. Ensemble sparse classification of Alzheimer's disease. Neuroimage 60, 1106–1116.

Manjón, J., Eskildsen, S., Coupé, P., Romero, J., Collins, L., Robles, M., 2014. NICE: Non-local intracranial cavity extraction. Int J Biomed Imaging.

Manjón, J.V., Coupé, P., 2016. volBrain: an online MRI brain volumetry system. Front Neuroinform 10.

Manjón, J.V., Coupé, P., Martí-Bonmatí, L., Collins, D.L., Robles, M., 2010. Adaptive non-local means denoising of MR images with spatially varying noise levels. J. Magn. Reson. Imaging 31, 192–203.

Matias-Guiu, J.A., Valles-Salgado, M., Rognoni, T., Hamre-Gil, F., Moreno-Ramos, T., Matias-Guiu, J., 2017. Comparative diagnostic accuracy of the ace-iii, mis, mmse, moca, and rudas for screening of Alzheimer disease. Dement Geriatr Cogn Disord 43, 237–246.

Moradi, E., Pepe, A., Gaser, C., Huttunen, H., Tohka, J., Initiative, A. D. N. et al. (2015). Machine learning framework for early MRI-based Alzheimer's conversion prediction in MCI subjects. Neuroimage, 104, 398–412.

Mueller, S., Stables, L., Du, A., Schuff, N., Truran, D., Cashdollar, N., & Weiner, M. (2007). Measurement of hippocampal subfields and age-related changes with high resolution MRI at 4T. Neurobiology of aging, 28, 719–726.

Lorente de Nó, R., 1934. Studies on the structure of the cerebral cortex. ii. continuation of the study of the ammonic system. Journal für Psychologie und Neurologie.

Parisot, S., Ktena, S.I., Ferrante, E., Lee, M., Guerrero, R., Glocker, B., Rueckert, D., 2018. Disease prediction using graph convolutional networks: application to autism spectrum disorder and alzheimer's disease. Med Image Anal 48, 117–130.

Parker, T.D., Cash, D.M., Lane, C.A., Lu, K., Malone, I.B., Nicholas, J.M., James, S.-N., Keshavan, A., Murray-Smith, H., Wong, A., et al., 2019. Hippocampal subfield volumes and pre-clinical alzheimer's disease in 408 cognitively normal adults born in 1946. PLoS ONE 14.

Petersen, R.C., Smith, G.E., Waring, S.C., Ivnik, R.J., Tangalos, E.G., Kokmen, E., 1999. Mild cognitive impairment: clinical characterization and outcome. Arch. Neurol. 56, 303–308.

Petrella, J.R., Coleman, R.E., Doraiswamy, P.M., 2003. Neuroimaging and early diagnosis of alzheimer disease: a look to the future. Radiology 226, 315–336.

Rathore, S., Habes, M., Iftikhar, M.A., Shacklett, A., Davatzikos, C., 2017. A review on neuroimaging-based classification studies and associated feature extraction methods for alzheimer's disease and its prodromal stages. Neuroimage 155, 530–548.

Romero, J.E., Coupe, P., Manjon, J.V., 2017. Hips: a new hippocampus subfield segmentation method. Neuroimage 163, 286–295.

Rubner, Y., Tomasi, C., Guibas, L.J., 2000. The earth mover's distance as a metric for image retrieval. Int J Comput Vis 40, 99–121.

Samper-Gonzalez, J., Burgos, N., Bottani, S., Habert, M.-O., Evgeniou, T., Epelbaum, S., Colliot, O., 2019. Reproducible evaluation of methods for predicting progression to Alzheimer's disease from clinical and neuroimaging Data. SPIE Medical Imaging 2019.

Schwarz, C.G., Gunter, J.L., Wiste, H.J., Przybelski, S.A., Weigand, S.D., Ward, C.P., Senjem, M.L., Vemuri, P., Murray, M.E., Dickson, D.W., et al., 2016. A large-scale comparison of cortical thickness and volume methods for measuring alzheimer's disease severity. NeuroImage: Clinical 11, 802–812.

Sturges, H.A., 1926. The choice of a class interval. J Am Stat Assoc 21, 65–66.

Suk, H.-I., Lee, S.-W., Shen, D., 2017. Deep ensemble learning of sparse regression models for brain disease diagnosis. Med Image Anal 37, 101–113.

Suk, H.-I., Lee, S.-W., Shen, D., Initiative, A.D.N., et al., 2014. Hierarchical feature representation and multimodal fusion with deep learning for ad/mci diagnosis. Neuroimage 101, 569–582.

Suk, H.-I., Lee, S.-W., Shen, D., Initiative, A.D.N., et al., 2015. Latent feature representation with stacked auto-encoder for ad/mci diagnosis. Brain Structure and Function 220, 841–859.

Tong, T., Gao, Q., Guerrero, R., Ledig, C., Chen, L., Rueckert, D., Initiative, A.D.N., 2017. A novel grading biomarker for the prediction of conversion from mild cognitive impairment to alzheimer's disease. IEEE Trans. Biomed. Eng. 64, 155–165.

Tong, T., Gray, K., Gao, Q., Chen, L., Rueckert, D., Initiative, A.D.N., et al., 2017. Multi–modal classification of alzheimer's disease using nonlinear graph fusion. Pattern Recognit 63, 171–181.

Tong, T., Wolz, R., Gao, Q., Guerrero, R., Hajnal, J.V., Rueckert, D., Initiative, A.D.N., 2014. Multiple instance learning for classification of dementia in brain MRI. Med Image Anal 18, 808–818.

Trujillo-Estrada, L., Dávila, J.C., Sánchez-Mejias, E., Sánchez-Varo, R., Gomez-Arboledas, A., Vizuete, M., Vitorica, J., Gutiérrez, A., 2014. Early neuronal loss and axonal/presynaptic damage is associated with accelerated amyloid-β accumulation in aβPP/PS1 alzheimer's disease mice subiculum. J. Alzheimers Dis. 42, 521–541.

Tustison, N.J., Avants, B.B., Cook, P.A., Zheng, Y., Egan, A., Yushkevich, P.A., Gee, J.C., 2010. N4ITK: improved N3 bias correction. IEEE Trans Med Imaging 29, 1310–1320.

Wang, H., Das, S.R., Suh, J.W., Altinay, M., Pluta, J., Craige, C., Avants, B., Yushkevich, P.A., Initiative, A.D.N., et al., 2011. A learning-based wrapper method to correct systematic errors in automatic image segmentation: consistently improved performance in hippocampus, cortex and brain segmentation. Neuroimage 55, 968–985.

Wee, C.-Y., Liu, C., Lee, A., Poh, J.S., Ji, H., Qiu, A., Initiative, A.D.N., et al., 2019. Cortical graph neural network for ad and mci diagnosis and transfer learning across populations. NeuroImage: Clinical 101929.

Wee, C.-Y., Yap, P.-T., Shen, D., Initiative, A.D.N., 2013. Prediction of Alzheimer's disease and mild cognitive impairment using cortical morphological patterns. Hum Brain Mapp 34, 3411–3425.

Wen, J., Thibeau, E., Samper-González, J., Routier, A., Bottani, S., Dormont, D., Durrleman, S., Colliot, O., Burgos, N., et al., 2019. How serious is data leakage in Deep learning studies on Alzheimer's disease classification? OHBM Annual meeting - Organization for Human Brain Mapping.

Wen, J., Thibeau-Sutre, E., Samper-Gonzalez, J., Routier, A., Bottani, S., Durrleman, S., Burgos, N., Colliot, O., 2019. Convolutional neural networks for classification of alzheimer's disease: overview and reproducible evaluation. arXiv preprint arXiv:1904.07773.

West, M.J., Coleman, P.D., Flood, D.G., Troncoso, J.C., 1994. Differences in the pattern of hippocampal neuronal loss in normal ageing and Alzheimer's disease. The Lancet 344, 769–772.

Winterburn, J.L., Pruessner, J.C., Chavez, S., Schira, M.M., Lobaugh, N.J., Voineskos, A.N., Chakravarty, M.M., 2013. A novel in vivo atlas of human hippocampal subfields using high-resolution 3T magnetic resonance imaging. Neuroimage 74, 254–265.

Wolz, R., Julkunen, V., Koikkalainen, J., Niskanen, E., Zhang, D.P., Rueckert, D., Soininen, H., Lötjönen, J., Initiative, A.D.N., 2011. Multi-method analysis of MRI images in early diagnostics of Alzheimer's disease. PLoS ONE 6, e25446.

Wyman, B.T., Harvey, D.J., Crawford, K., Bernstein, M.A., Carmichael, O., Cole, P.E., Crane, P.K., DeCarli, C., Fox, N.C., Gunter, J.L., et al., 2013. Standardization of analysis sets for reporting results from adni mri data. Alzheimer's & Dementia 9, 332–337.

Yushkevich, P.A., Amaral, R.S., Augustinack, J.C., Bender, A.R., Bernstein, J.D., Boccardi, M., Bocchetta, M., Burggren, A.C., Carr, V.A., Chakravarty, M.M., 2015. Quantitative comparison of 21 protocols for labeling hippocampal subfields and parahippocampal subregions in vivo MRI: towards a harmonized segmentation protocol. Neuroimage 111, 526–541.

Zhou, L., Wang, Y., Li, Y., Yap, P.-T., Shen, D., (ADNI, A.D.N.I., et al., 2011. Hierarchical anatomical brain networks for MCI prediction: revisiting volumetric measures. PLoS ONE 6, e21935.

Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67, 301–320.