



HAL
open science

Automatic classification of patients with Alzheimer's disease from structural MRI: a comparison of ten methods using the ADNI database

Rémi Cuingnet, Emilie Gerardin, Jérôme Tessieras, Guillaume Auzias, Stéphane Lehéricy, Marie-Odile Habert, Marie Chupin, Habib Benali, Olivier Colliot

► To cite this version:

Rémi Cuingnet, Emilie Gerardin, Jérôme Tessieras, Guillaume Auzias, Stéphane Lehéricy, et al. Automatic classification of patients with Alzheimer's disease from structural MRI: a comparison of ten methods using the ADNI database. *NeuroImage*, 2011, 56 (2), pp.766-81. 10.1016/j.neuroimage.2010.06.013. hal-00795739

HAL Id: hal-00795739

<https://inria.hal.science/hal-00795739v1>

Submitted on 29 Dec 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

1 **Automatic classification of patients with Alzheimer’s disease from structural**
2 **MRI: a comparison of ten methods using the ADNI database**

3
4 Rémi Cuingnet^{1,2,3,4}, MSc; Emilie Gérardin^{1,2,3}, MSc; Jérôme Tessieras^{1,2,3}, MSc; Guillaume
5 Auzias^{1,2,3}, MSc; Stéphane Lehéricy^{1,2,3,5}, MD, PhD; Marie-Odile Habert^{4,6}, MD, PhD; Marie
6 Chupin^{1,2,3}, PhD; Habib Benali⁴, PhD; Olivier Colliot^{1,2,3}, PhD and the Alzheimer’s Disease
7 Neuroimaging Initiative*

8
9 ¹*UPMC Université Paris 6, UMR 7225, UMR_S 975, Centre de Recherche de l’Institut du Cerveau*
10 *et de la Moelle épinière (CRICM), Paris, F-75013, France*

11 ²*CNRS, UMR 7225, CRICM, Paris, F-75013, France*

12 ³*Inserm, UMR_S 975, CRICM, Paris, F-75013, France*

13 ⁴*Inserm, UMR_S 678, LIF, Paris, F-75013, France*

14 ⁵*Centre for NeuroImaging Research – CENIR, Department of Neuroradiology, Groupe hospitalier*
15 *Pitié-Salpêtrière, Paris, F-75013, France*

16 ⁶*AP-HP, Department of Nuclear Medicine, Groupe hospitalier Pitié-Salpêtrière, Paris, F-75013,*
17 *France*

18
19
20
21
22 **Key words:** Alzheimer’s disease, AD, MCI, converter, prodromal, classification, magnetic
23 resonance imaging, support vector machines,
24

25
26
27 **Type of manuscript:** Review
28

29
30
31
32
33 **Correspondence to:**

34 Rémi Cuingnet
35 CRICM – Equipe Cogimage (ex LENA)
36 Hôpital de la Pitié-Salpêtrière
37 47, boulevard de l’hôpital
38 75651 Paris Cedex 13
39 France
40 Phone: +33 1 42 17 85 24
41 e-mail: remi.cuingnet@gmail.com
42

* Data used in the preparation of this article were obtained from the Alzheimer’s Disease Neuroimaging Initiative (ADNI) database (www.loni.ucla.edu/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. ADNI investigators include (complete listing available at www.loni.ucla.edu/ADNI/Collaboration/ADNI_Authorship_list.pdf).

1 **Abstract**

2 Recently, several high dimensional classification methods have been proposed to automatically
3 discriminate between patients with Alzheimer's disease (AD) or mild cognitive impairment (MCI)
4 and elderly controls (CN) based on T1-weighted MRI. However, these methods were assessed on
5 different populations, making it difficult to compare their performance. In this paper, we evaluated
6 the performance of ten approaches (five voxel-based methods, three methods based on cortical
7 thickness and two methods based on the hippocampus) using 509 subjects from the ADNI database.
8 Three classification experiments were performed: CN *vs* AD, CN *vs* MCIc (MCI who had converted
9 to AD within 18 months, MCI converters - MCIc) and MCIc *vs* MCInc (MCI who had not
10 converted to AD within 18 months, MCI non-converters - MCInc). Data from 81 CN, 67 MCInc, 39
11 MCIc and 69 AD were used for training and hyperparameters optimization. The remaining
12 independent samples of 81 CN, 67 MCInc, 37 MCIc and 68 AD were used to obtain an unbiased
13 estimate of the performance of the methods. For AD *vs* CN, whole-brain methods (voxel-based or
14 cortical thickness-based) achieved high accuracies (up to 81% sensitivity and 95% specificity). For
15 the detection of prodromal AD (CN *vs* MCIc), the sensitivity was substantially lower. For the
16 prediction of conversion, no classifier obtained significantly better results than chance. We also
17 compared the results obtained using the DARTEL registration to that using SPM5 unified
18 segmentation. DARTEL significantly improved six out of 20 classification experiments and led to
19 lower results in only two cases. Overall, the use of feature selection did not improve the
20 performance but substantially increased the computation times.

21

1 **Introduction**

2 Alzheimer's disease (AD) is the most frequent neurodegenerative dementia and a growing health
3 problem. Definite diagnosis can only be made postmortem, and requires histopathological
4 confirmation of amyloid plaques and neurofibrillary tangles. Early and accurate diagnosis of
5 Alzheimer's Disease (AD) is not only challenging, but is crucial in the perspective of future
6 treatments. Clinical diagnostic criteria are currently based on the clinical examination and
7 neuropsychological assessment, with the identification of dementia and then of the Alzheimer's
8 phenotype (Blennow et al., 2006). Patients suffering from AD at a prodromal stage are, mostly,
9 clinically classified as amnesic mild cognitive impairment (MCI) (Petersen et al., 1999; Dubois and
10 Albert, 2004), but not all patients with amnesic MCI will develop AD. Recently, more precise
11 research criteria were proposed for the early diagnostic of AD at the prodromal stage of the disease
12 (Dubois et al., 2007). These criteria are based on a clinical core of early episodic memory
13 impairment and the presence of at least one additional supportive feature including abnormal MRI
14 and PET neuroimaging or abnormal cerebrospinal fluid amyloid and tau biomarkers (Dubois et al.,
15 2007). Neuroimaging therefore adds a positive predictive value to the diagnosis and includes
16 measurements using structural MRI to assess medial temporal lobe atrophy and positron emission
17 tomography using fluorodeoxyglucose (FDG) or amyloid markers (Fox et al., 2004; Jagust, 2006)

18 Many groups studies based on volumetric measurements of regions of interest (ROI) (Convit
19 et al., 1997, 2000; Jack et al., 1997,1998; Juottonen et al., 1998; Laakso et al., 1998, 2000; Busatto
20 et al., 2003; Xu et al., 2000; Good et al., 2002; Chételat and Baron, 2003; Rusinek et al., 2004;
21 Tapiola et al., 2008), voxel-based morphometry (Good et al., 2002; Busatto et al., 2003; Karas et al.,
22 2003, 2004; Chételat et al., 2005; Whitwell et al., 2007, 2008) or group comparison of cortical
23 thickness (Thompson et al., 2001, 2003, 2004; Lerch et al., 2005, 2008; Bakkour et al., 2009,
24 Dickerson et al., 2009; Hua et al., 2009; McDonald et al., 2009) have shown that brain atrophy in
25 AD and prodromal AD is spatially distributed over many brain regions including the entorhinal
26 cortex, the hippocampus, lateral and inferior temporal structures, anterior and posterior cingulate.

27 However these analyses measure group differences and thus are of limited value for individual
28 diagnosis.

29 Advances in statistical learning with the development of new machine learning algorithms
30 capable of dealing with high dimensional data, such as the support vector machine (SVM) (Vapnik,
31 1995; Shawe-Taylor and Cristianini, 2000; Schölkopf and Smola, 2001), enables the development
32 of new diagnostic tools based on T1-weighted MRI. Recently, several approaches have been
33 proposed to automatically classify patients with AD and/or MCI from anatomical MRI (Fan et al.,
34 2005, 2007, 2008a, 2008b; Colliot et al., 2008; Davatzikos et al., 2008a, 2008b; Klöppel et al.,
35 2008; Vemuri et al., 2008; Chupin et al., 2009a, 2009b; Desikan et al., 2009; Gerardin et al., 2009;
36 Hinrichs et al., 2009; Magnin et al., 2009; Misra et al., 2009; Querbes et al., 2009). These
37 approaches could have the potential to assist in the early diagnosis of AD. These approaches can
38 roughly be grouped into three different categories, depending on the type of features extracted from
39 the MRI (voxel-based, vertex-based or ROI-based). In the first category, the features are defined at
40 the level of the MRI voxel. Specifically, the features are the probability of the different tissue
41 classes (grey matter, white matter and cerebrospinal fluid) in a given voxel (Lao et al., 2004; Fan et
42 al., 2007, 2008a, 2008b; Davatzikos et al., 2008a, 2008b; Klöppel et al., 2008; Vemuri et al., 2008;
43 Hinrichs et al., 2009; Magnin et al., 2009; Misra et al., 2009). Klöppel et al. (2008) directly
44 classified these features with an SVM. All other methods first reduce the dimensionality of the
45 feature space relying on different types of features extraction, agglomeration and/or selection
46 methods. Vemuri et al. (2008) used smoothing, voxel-downsampling, and then a feature selection
47 step. Another solution is to group voxels into anatomical regions through the registration of a
48 labeled atlas (Lao et al., 2004; Ye et al., 2008; Magnin et al., 2009). However, this anatomical
49 parcellation may not be adapted to the pathology. In order to overcome this limitation, Fan et al.
50 (2007) have proposed an adaptive parcellation approach in which the image space is divided into
51 the most discriminative regions. This method has been used in several studies (Davatzikos et al.,
52 2008a, 2008b; Fan et al., 2008a, 2008b; Misra et al., 2009). In the second category, the features are

53 defined at the vertex-level on the cortical surface (Desikan et al., 2009; Querbes et al., 2009). The
54 methods of the third category include only the hippocampus. Their approach is based on the
55 analysis of the volume and/or shape of the hippocampus (Colliot et al., 2008, Chupin et al., 2009a,
56 2009b; Gerardin et al., 2009).

57 These approaches achieve high accuracy (over 84%). However, they were evaluated on
58 different study populations, making it difficult to compare their respective discriminative power.
59 Indeed, many factors such as degree of impairment, age, gender, genotype, educational level and
60 MR image quality perceptibly affect the evaluation of the prediction accuracy. This variability
61 between evaluations is increased for statistical reasons when the number of subjects is small.
62 Therefore a meta-analysis would be of limited value to compare the prediction accuracies of
63 different methods.

64 The goal of this paper was to compare different methods for the classification of patients
65 with AD based on anatomical MRI, using the same study population. To that purpose, we used the
66 Alzheimer's Disease Neuroimaging Initiative (ADNI) database. Ten methods were evaluated. We
67 tested five voxel-based approaches: a direct approach (Klöppel et al., 2008), an approach based on a
68 volume of interest (Klöppel et al., 2008), an atlas-based approach (Magnin et al., 2009) and the
69 approaches proposed by Vemuri et al. (2008) and Fan et al. (2008a, 2008b) respectively. In order to
70 assess the influence of the registration step and the features used on the classification accuracies,
71 these latter methods were tested with two different registration steps: SPM5 (Ashburner and
72 Friston, 2005) and DARTEL (Ashburner, 2007) and also with either only the grey matter (GM)
73 probability maps or all the tissues probability maps including also white matter (WM) and
74 cerebrospinal fluid (CSF). Three cortical approaches were evaluated as well: a direct one similar to
75 (Klöppel et al., 2008), an atlas based one and an approach using only the regions found in (Desikan
76 et al., 2009). Two methods respectively based on the volume (Colliot et al., 2008, Chupin et al.,
77 2009a, 2009b) and the shape (Gerardin et al., 2009) of the hippocampus were also tested.

78

79

80 **Materials**

81 **Data**

82 Data used in the preparation of this article were obtained from the Alzheimer's disease
83 Neuroimaging Initiative (ADNI) database (www.loni.ucla.edu/ADNI). The ADNI was launched in
84 2003 by the National Institute on Aging (NIA), the National Institute of Biomedical Imaging and
85 Bioengineering (NIBIB), the Food and Drug Administration (FDA), private pharmaceutical
86 companies and non-profit organizations, as a \$60 million, 5-year public-private partnership. The
87 primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron
88 emission tomography (PET), other biological markers, and clinical and neuropsychological
89 assessment can be combined to measure the progression of mild cognitive impairment (MCI) and
90 early Alzheimer's disease (AD). Determination of sensitive and specific markers of very early AD
91 progression is intended to aid researchers and clinicians to develop new treatments and monitor
92 their effectiveness, as well as lessen the time and cost of clinical trials.

93

94 **MRI acquisition**

95 The MR scans are T1-weighted MR images. MRI acquisition had been done according to the ADNI
96 acquisition protocol in (Jack et al., 2008). For each subject, we used the MRI scan from the baseline
97 visit when available and from the screening visit otherwise. We only used images acquired at 1.5 T.
98 To enhance standardization across sites and platforms of images acquired in the ADNI study, pre-
99 processed images that have undergone some post-acquisition correction of certain image artifacts
100 are available (Jack et al., 2008). We used those corrected in image geometry for gradient
101 nonlinearity and corrected for intensity non-uniformity due to non-uniform receiver coil sensitivity.
102 The image geometry correction was the 3D gradwarp correction (Hajnal et al., 2001; Jovicich et al.
103 2006). The B1 non-uniformity correction is detailed in (Narayana et al., 1988). These two
104 preprocessing steps can be performed directly on the MRI console and thus seem feasible in clinical
105 routine. All subjects were scanned twice at each visit. As explained in (Jack et al., 2008), MR scans

106 were graded qualitatively by the ADNI investigators of the ADNI MRI quality control center at the
107 Mayo Clinic for artifacts and general image quality. Each scan was graded on several separate
108 criteria: blurring/ghosting, flow artifact, intensity and homogeneity, signal-to-noise ratio (SNR),
109 susceptibility artifacts, and gray-white/cerebrospinal fluid contrast. For each subject, we used the
110 MRI scan which was considered as the “best” quality scan by the ADNI investigators. In the
111 description of the ADNI methods (http://www.loni.ucla.edu/ADNI/Data/ADNI_Data.shtml), the
112 “best” quality image is the one which was used for the complete pre-processing steps. We thus used
113 the images which had been selected for the complete pre-processing pipeline. No other exclusion
114 criteria based on image quality were applied. The identification numbers of the images used in this
115 study are reported in Tables S2 to S9.

116

117 **Participants**

118 The criteria used for the inclusion of participants were those defined in the ADNI protocol
119 (described in details at <http://www.adni-info.org/Scientists/AboutADNI.aspx#>). Enrolled subjects
120 were between 55-90 (inclusive) years of age, had a study partner able to provide an independent
121 evaluation of functioning, and spoke either English or Spanish. All subjects were willing and able to
122 undergo all test procedures including neuroimaging and agreed to longitudinal follow up. Specific
123 psychoactive medications were excluded. General inclusion/exclusion criteria were as follows:
124 Control subjects (CN) had MMSE scores between 24 and 30 (inclusive), a CDR (Clinical Dementia
125 Rating) (Morris, 1993) of zero. They were non-depressed, non MCI, and non-demented. MCI
126 subjects had MMSE scores between 24 and 30 (inclusive), a memory complaint, had objective
127 memory loss measured by education adjusted scores on Wechsler Memory Scale Logical Memory II
128 (Wechsler, 1987), a CDR of 0.5, absence of significant levels of impairment in other cognitive
129 domains, essentially preserved activities of daily living, and an absence of dementia. AD patients
130 had MMSE scores between 20 and 26 (inclusive), CDR of 0.5 or 1.0, and met NINCDS/ADRDA
131 criteria for probable AD (McKhann et al., 1984).

132 We selected all the subjects for whom preprocessed images were available. The
133 identification numbers of the subjects used in this study are reported in Tables S2 to S9. As a result,
134 509 subjects were selected: 162 cognitively normal elderly controls (CN) (76 males, 86 females,
135 age \pm SD = 76.3 \pm 5.4 years, range = 60–90 years, mini-mental score (MMS) = 29.2 \pm 1.0, range =
136 25–30), 137 patients with AD (67 males, 70 females, age \pm SD = 76.0 \pm 7.3 years, range = 55–91
137 years, MMS = 23.2 \pm 2.0, range = 18–27), 76 patients with MCI who had converted to AD within
138 18 months (MCIC) (43 males, 33 females, age \pm SD = 74.8 \pm 7.4 years, range = 55–88 years, MMS
139 = 26.5 \pm 1.9, range = 23–30) and 134 patients with MCI who had not converted to AD within 18
140 months (MCINC) (84 males, 50 females, age \pm SD = 74.5 \pm 7.2 years, range = 58–88 years, MMS =
141 27.2 \pm 1.7, range = 24–30). We did not consider MCI patients who had been followed less than 18
142 months and had not converted within this time frame. The 509 images came from 41 different
143 centers.

144 To assess differences in demographic and clinical characteristics between groups, we used
145 Student's t-test for age and MMS and Pearson's chi-square test for gender. Significance level was
146 set at 0.05. No significant differences were found except for the MMS between controls and
147 patients (AD or MCIC, $p < 0.0001$).

148 In order to obtain unbiased estimates of the performances, the set of participants was then
149 randomly split up into two groups of the same size: a training set and a testing set. The algorithms
150 were trained on a training set and the measures of the diagnostic sensitivity and specificity were
151 carried out with an independent test set. The division process preserved the age and sex distribution.

152 Demographic characteristics of the studied population selected from the ADNI database are
153 presented in Table 1.

1 **Methods**

2 **Classification experiments**

3 Three classification experiments were performed to compare the different approaches. The first one
4 is the classification between CN subjects and patients with probable AD and is referred to as “CN *vs*
5 AD” in the following. The second one is the classification between CN subjects and MCI
6 converters and is referred to as “CN *vs* MCIc”. It corresponds to the detection of patients with
7 prodromal AD as defined by (Dubois and Albert, 2004). Indeed, MCI patients who will convert to
8 AD are, at baseline, patients with incipient AD but non-demented, i.e. patients with prodromal AD.
9 The third one is the classification MCIinc versus MCIc and is referred to as “MCIinc *vs* MCIc”. It
10 corresponds to the prediction of conversion in MCI patients.

11

12 **Classification methods**

13 The different approaches we compared can be grouped into three categories with respect to the
14 features used for the classification. In the first category, the features are defined at the level of the
15 MRI voxel. Specifically, the features are the probability of the different tissue classes (GM, WM
16 and CSF) in a given voxel. In the second category, the features are defined at the vertex-level on the
17 cortical surface. Specifically, the features are the cortical thickness at each vertex of the cortex. The
18 methods of the third category include only the hippocampus.

19 These methods are summarized in Table 2 and briefly presented in the following paragraphs.

20

21 **First category: voxel-based segmented tissue probability maps**

22 The features of the methods of the first category were computed as follows. All T1-weighted MR
23 images were spatially normalized and segmented into GM, WM and CSF using the SPM5
24 (Statistical Parametric Mapping, London, UK) unified segmentation routine (Ashburner and
25 Friston, 2005) with the default parameters. These maps constitute a first set of tissue probability
26 maps and will be referred to respectively as SPM5_GM, SPM5_WM and SPM5_CSF.

27 To evaluate the impact of the registration step on the classification accuracy, the GM and
28 WM probability maps in native space segmented by the SPM5 unified segmentation routine were
29 also normalized to the population template generated from all the images, using the DARTEL
30 diffeomorphic registration algorithm (Ashburner, 2007) with the default parameters. The obtained
31 transformations were applied to the GM, WM and CSF tissue maps. These maps compose a second
32 set of tissue probability maps and will be referred to respectively as DARTEL_GM, DARTEL_WM
33 and DARTEL_CSF. Some papers used only GM maps while others included all three tissues. In our
34 experiments, we systematically evaluated the added value of WM and CSF maps by comparing the
35 classification obtained with only GM to that obtained with all three classes. All maps were then
36 modulated to ensure that the overall tissue amount remains constant. No spatial smoothing was
37 performed, unless when otherwise specified.

38 The different methods of this category differ by the way the features are extracted and/or
39 selected from the voxel probability maps. This is detailed in the following paragraphs.

40

41 *Direct*

42 The simplest approach consists in considering the voxels of the tissue probability maps directly as
43 features in the classification. This type of approach is referred to as “*Voxel-Direct*” in the following.
44 Such an approach was proposed by Klöppel et al. (2008) with two different versions: one is based
45 on whole brain datasets and the other includes only data from a volume of interest (VOI) located in
46 the anterior medial temporal lobe, including part of the hippocampus. This volume of interest was
47 defined as two rectangular cuboids centered on $x = -17, y = -8, z = -18$ and $x = 16, y = -9, z = -18$ in
48 the MNI space. Their dimensions were 12 mm, 16 mm and 12 mm in the x, y and z directions
49 respectively. The latter method will be referred to as “*Voxel-Direct_VOP*”. In their paper, they used
50 only DARTEL_GM maps. Here, we will test all approaches with the following sets of probability
51 maps: SPM5_GM only, SPM5_GM and SPM5_WM and SPM5_CSF, DARTEL_GM only,
52 DARTEL_GM, and DARTEL_WM and DARTEL_CSF.

53

54 *STAND-score*

55 Vemuri et al. (2008) proposed an approach called the STAND score, in which the dimensionality is
56 reduced by a sequence of feature aggregation and selection steps. First, the tissue probability maps
57 were smoothed and down-sampled by averaging. Then, voxels that contained less than 10% tissue
58 density values and CSF in half or more of the images were not considered for further analysis. A
59 feature selection step was then carried out. First, a linear SVM was applied for each tissue class,
60 which attributes a weight to each feature. Only features of which weights are consistent with
61 increased neurodegeneration in the pathological group were kept. Then a second feature selection
62 step was performed on the remaining features. To ensure spatial consistency, neighboring voxels of
63 the voxels selected so far were also selected. The features from the different tissue classes were
64 concatenated and then used in the classification. This approach is referred to as “*Voxel-STAND*” in
65 the following. In their paper, the features used for this approach were the GM, WM and CSF tissue
66 probability maps segmented and registered with the SPM5 unified segmentation routine using a
67 customized tissue probability maps. Thus we also tested the classification with customized tissue
68 probability maps.

69

70 *Atlas based*

71 Another approach consists in grouping the voxels into anatomical regions using a labeled atlas. This
72 type of approach is used in (Lao et al., 2004; Magnin et al., 2009). Each tissue probability map in
73 the stereotaxic space was parceled into 116 regions of interest (ROI) using the AAL (Automatic
74 Anatomical Labeling) atlas (Tzourio-Mazoyer et al., 2002). In each ROI, we computed the mean
75 tissue probability and used these values as features in the classification. Such an approach will be
76 referred to as “*Voxel-Atlas*”. Note that the AAL is a predefined anatomical atlas, which has not been
77 specifically designed for studying patients with AD; its areas thus do not necessarily represent
78 pathologically homogeneous regions.

79

80 *COMPARE*

81 Instead of using a predefined atlas, Fan et al. (2007, 2008a, 2008b) proposed a parcellation that is
82 adapted to the pathology. The thorough explanation of the method is in (Fan et al., 2007). Very
83 briefly, the concept of COMPARE is to create homogeneously discriminative regions. In these
84 regions, the voxel values are aggregated to form the features of the classification. Feature selection
85 steps are then performed to identify the most discriminative regions. In the following, we refer to
86 this approach as “*Voxel-COMPARE*”. We used the COMPARE software freely available on request
87 for download online (<https://www.rad.upenn.edu/sbia/software/index.html>).

88

89 **Second category: cortical thickness**

90 In this second category, the features are the cortical thickness values at each vertex of the cortical
91 surface. Cortical thickness represents a direct index of atrophy and thus is a potentially powerful
92 candidate to assist in the diagnosis of AD (Thompson et al., 2001, 2003, 2004; Lerch et al., 2005,
93 2008; Bakkour et al., 2009; Dickerson et al., 2009; Hua et al., 2009; McDonald et al., 2009).
94 Cortical thickness measures were performed with the FreeSurfer image analysis suite
95 (Massachusetts General Hospital, Boston, MA), which is documented and freely available for
96 download online (<http://surfer.nmr.mgh.harvard.edu/>). The technical details of this procedure are
97 described in (Sled et al., 1998; Dale et al., 1999; Fischl et al., 1999a, 1999b; Fischl and Dale, 2000).
98 All the cortical thickness maps were registered onto the default FreeSurfer common template. Four
99 subjects were not successfully processed by the FreeSurfer pipeline. Those subjects are marked by
100 an asterisk in Tables S2 to S9. They could thus not be classified with the SVM and were excluded
101 from the training set. For the testing set, the subjects were considered as 50% misclassified.

102

103 *Direct*

104 As in (Klöppel et al., 2008) for voxel-based maps, the simplest way consists in considering cortical

105 thickness values at every vertices directly as features in the classification with no other
106 preprocessing step. This approach is referred to as “*Thickness-Direct*” in the following.

107

108 *Atlas based*

109 As in the voxel-based case, we also tested an approach where vertices are grouped into anatomical
110 regions using an atlas. Such approach is used in (Querbes et al., 2009, Desikan et al., 2009). The
111 cortical parcellation was carried out with the cortical atlas of (Desikan et al., 2006). The atlas is
112 composed of 68 gyral based ROIs. In each ROI, we computed the mean cortical thickness and used
113 these values as features in the classification. This approach is referred to as “*Thickness-Atlas*” in the
114 following.

115

116 *ROI*

117 Desikan et al. (2009) parcellated the brain into neocortical and non-neocortical ROIs by wrapping
118 an anatomical atlas (Desikan et al., 2006). They studied the discriminative power for CN versus
119 MCIc of the mean thickness (neocortical regions) and the volume (both neocortical and non-
120 neocortical regions). For their analysis, the mean thickness and the volumes of the right and the left
121 hemispheres, for each ROI, were added together. The volumes were corrected using estimate of the
122 total intracranial volume.

123 Their study was carried out on a cohort of 97 participants selected from the Open Access Series of
124 Imaging Studies (OASIS) database (Marcus et al., 2007). They found out that, with a logistic
125 regression analysis, the best set of discriminator was: the entorhinal cortex thickness, the
126 supramarginal gyrus thickness and the hippocampal volume. They used these features with a
127 logistic regression to classify CN and MCIc and to classify CN and AD. Therefore, in this approach,
128 we used only these three features for the classification. This approach is referred to as “*Thickness-*
129 *ROI*” in the following.

130

131 **Third category: hippocampus**

132 Finally, we tested the discriminative power of methods which consider only the hippocampus and
133 not the whole brain or the whole cortex as in the two first categories. The hippocampus is affected
134 at the earliest stages of the disease and has thus been used as a marker of early AD in a vast number
135 of studies.

136 Here, the segmentation of the hippocampus was performed using SACHA, a fully automatic
137 method we previously developed (Chupin et al., 2007, 2009a). This approach has been shown to be
138 competitive with manual tracing for the discrimination of patients with AD and MCI (Colliot et al.,
139 2008; Chupin et al., 2009b). This approach segments both the hippocampus and the amygdala
140 simultaneously based on competitive region-growing between these two structures. It includes prior
141 knowledge on the location of the hippocampus and the amygdala derived from a probabilistic atlas
142 and on the relative positions of these structures with respect to anatomical landmarks which are
143 automatically identified.

144 We also evaluated the hippocampal volume obtained with the FreeSurfer image analysis
145 suite.

146

147 *Volume*

148 We first tested the classification accuracy obtained when the only feature is the hippocampal
149 volume. For each subject, we computed the volume of the hippocampi. Volumes were normalized
150 by the total intracranial volume (TIV) computed by summing SPM5 segmentation maps of grey
151 matter, white matter, and cerebrospinal fluid (CSF), inside a bounding box defined in standard
152 space to obtain a systematic inferior limit. For more robustness with respect to segmentation errors,
153 left and right volumes were averaged. The thorough explanation of the method is in (Chupin et al.,
154 2007, 2009a, 2009b). This approach is referred to as “*Hippo-Volume-S*” in the following.

155 We also evaluated this approach with the hippocampal volume obtained with the FreeSurfer
156 image analysis suite and corrected with the total intracranial volume also obtained with obtained
157 with FreeSurfer. This approach will be referred to as “*Hippo-Volume-F*”.

158

159 *Shape*

160 We then tested an approach in which the features describe the hippocampal shape (Gerardin et al.,
161 2009). Each hippocampus was described by a series of spherical harmonics (SPHARM) to model
162 the shape of the segmented hippocampi. The classification features were based on the SPHARM
163 coefficients. Specifically, each subject was represented by two sets (one for each hippocampus) of
164 three-dimensional SPHARM coefficients. The SPHARM coefficients were computed using the
165 SPHARM-PDM (Spherical Harmonics-Point Distribution Model) software developed by the
166 University of North Carolina and the National Alliance for Medical Imaging Computing
167 (http://www.namic.org/Wiki/index.php/Algorithm:UNC:Shape_Analysis). In the original paper by
168 our group describing this method (Gerardin et al., 2009), we used a feature selection step because
169 the subjects groups were much smaller (less than 30 subjects in each group). When the number of
170 subjects is small, the classifier can be more sensitive to uninformative features. In the present study,
171 the number of subjects was larger and thus a feature selection step is less necessary and increases
172 the risk of overfitting. We thus chose to avoid this selection step. We also tested the procedure with
173 the selection step but it did not lead to further improvement in this study. Moreover, the degree of
174 the SPHARM decomposition was set at four. Four subjects were not successfully processed by the
175 SPHARM pipeline. Those subjects are marked by a dagger in Tables S2 to S9. They could thus not
176 be classified with the SVM and were excluded from the training set. For the testing set, those
177 subjects were considered as 50% misclassified. This approach is referred to as “*Hippo-Shape*” in
178 the following.

179

180

181

182 **Classification using SVM**

183 *Classifiers*

184 A support vector machine is a supervised learning method. In brief: given a training set of size K :
185 $(x_k, y_k)_{k=1\dots K}$, where x_k in \mathbf{R}^d are observations, and y_k in $\{-1, 1\}$ are corresponding labels, SVMs
186 search for the optimal margin hyperplane (OMH) separating groups, i.e. the hyperplane for which
187 the margin between groups is maximal. More details on SVM can be found in (Vapnik 1995;
188 Shawe-Taylor and Cristianini, 2000, 2004; Schölkopf and Smola, 2001). We used a linear C-SVM
189 for all the approaches except COMPARE (Fan et al., 2007) for which a non-linear C-SVM with a
190 Gaussian kernel was used. The SVM implementation relied on the LIBSVM Library (Chang and
191 Lin, 2001).

192 The dimension of the features of the approach *Hippo-Volume* is only one. Therefore a much
193 simpler classifier can be used with no hyperparameter: each participant is assigned to the closest
194 group. Specifically, if S_1 and S_2 are two groups of participants with respective centers of mass
195 defined as m_1 and m_2 , a new individual with hippocampus volume x is assigned to the closest group
196 according to its Euclidean distance to the center of mass. This is a Parzen window classifier with the
197 linear kernel and assuming a prevalence of 50% (Shawe-Taylor and Cristianini, 2004).

198 As in (Desikan et al., 2009) a logistic regression is used instead of a SVM, the classification
199 step of *Thickness-ROI* was also based on a logistic regression.

200

201 *Evaluation*

202 In order to obtain unbiased estimates of the performances, the set of participants was randomly split
203 into two groups of the same size: a training set and a testing set. The division process preserves the
204 age and sex distribution. The training set was used to determine the optimal values of the
205 hyperparameters of each method and to train the classifier. The testing set was then only used to
206 evaluate the classification performances. The training and testing sets were identical for all

207 methods, except for those four cases for which the cortical thickness pipeline failed and those other
208 four for which the SPHARM pipeline failed. For the SPHARM and the cortical thickness methods,
209 the subjects for whom the corresponding pipeline failed could not be classified with the SVM and
210 were therefore excluded from the training set. As for the testing set, since those subjects were
211 neither misclassified nor correctly classified, they were considered as 50% misclassified. This
212 approach was chosen because a failure of the pipeline is a weakness of the methods.

213 On the training set, cross-validation (CV) was used to estimate the optimal values of
214 hyperparameters. In general, there is only one hyperparameter which is the cost parameter C of the
215 linear C -SVM. In *Voxel-STAND*, there is a second parameter which is the threshold t of feature
216 selection. In *Voxel-COMPARE*, a second parameter is the size σ of the Gaussian kernel and the third
217 parameter is the number n of selected features. In *Hippo-Volume*, there is no hyperparameter. The
218 optimal parameters values were determined using a grid-search and leave-one-out cross validation
219 (LOOCV) on the training set. The grid search was performed over the ranges $C = 10^{-5}, 10^{-4.5}, \dots,$
220 $10^{2.5}, 10^3, t = 0.06, 0.08, \dots, 0.98, \sigma = 100, 200, \dots, 1000$ and $n = 1, 2, \dots, 150$ (except for *Voxel-*
221 *COMPARE* were $C=10^0, 10^1, 10^{1.5}, 10^2, 10^{2.5}$).

222 For each approach, the optimized set of hyperparameters was then used to train the classifier
223 using the training group; the performance of the resulting classifier was then evaluated on the
224 testing set. In this way, we achieved unbiased estimates of the performances of each method.

225 For each method, we computed the number of true positives TP (i.e. the number of diseased
226 individuals which were correctly identified by the classifier), the number of true negatives TN (i.e.
227 the number of healthy individuals which were correctly identified by the classifier), the number of
228 false positives FP (i.e. the number of healthy individuals which were not correctly identified by the
229 classifier), the number of false negatives FN (i.e. the number of diseased individuals which were
230 not correctly identified by the classifier). We then computed the sensitivity defined as $TP/(TP+FN)$,
231 the specificity defined as $TN/(TN+FP)$, the positive predictive value defined as $PPV=TP/(TP+FP)$,
232 the negative predictive value defined as $NPV=TN/(TN+FN)$. Finally it should be noted that the

233 number of subjects in each group is not the same. The classification accuracy does not enable to
234 compare the performances between the different classification experiments. Thus we considered
235 both the specificity and the sensitivity instead.

236 To assess whether each method performs significantly better than a random classifier, we
237 used McNemar's chi square tests. Significance level was set at 0.05. We also used McNemar's chi
238 square tests to assess differences between DARTEL and SPM5 registrations and between
239 classification results obtained using only GM and using all three maps. The McNemar test
240 investigates the difference between proportions in paired observations. We used it to assess the
241 difference between proportions of correctly classified subjects, i.e. accuracy. The corresponding
242 contingency table is presented in Table 3.

243

1 **Results**

2 **Classification results**

3 The results of the classification experiments are summarized in Tables 4, 5 and 6 respectively for
4 CN vs AD, CN vs MCIC and CN vs MCInc. The classification results of CN vs AD and CN vs MCIC
5 are also represented in Figure 1. In each table, the different methods are referred to either by their
6 abbreviation or by their number defined in Table 2.

7

8 *CN vs AD*

9 The classification results for CN vs AD are summarized in Table 4 and in Figure 1. All methods
10 performed significantly better than chance ($p < 0.05$). The four *Voxel* methods (*Voxel-Direct*, *Voxel-*
11 *STAND*, *Voxel-Atlas*, *Voxel-COMPARE*) classified AD from CN with very high specificity (over
12 89%) and high sensitivity: 75% for *Voxel-STAND* and over 81% for the other three methods.
13 Methods based on the cortical thickness led to similar results with at least 90% specificity and 69%,
14 74% and 79% respectively for *Thickness-ROI*, *Thickness-Direct* and *Thickness-Atlas*. The
15 hippocampus-based strategies were as sensitive but less specific: between 63% for *Hippo-Volume*
16 and 84% for *Hippo-Shape*.

17

18 *CN vs MCIC*

19 Classification results for CN vs MCIC are summarized in Table 5 and in Figure 1. Most methods
20 were substantially less sensitive than for AD vs CN classification. All methods except *Voxel-*
21 *COMPARE* and the *Hippo* methods obtained significantly better results than a random classifier
22 ($p < 0.05$). There was no substantial difference between the results obtained with *Voxel-Direct*, *Voxel-*
23 *Atlas* and *Voxel-STAND*. All those methods reached a high specificity (over 85%) but a sensitivity
24 ranging between 51% (*Voxel-COMPARE*) and 73% (*Voxel-STAND*). The methods based on cortical
25 thickness behave as well as the previous ones. *Hippo-Volume* was slightly less specific but as
26 sensitive as for the AD vs CN classification.

27 *MCInc vs MCIC*

28 The classification results for MCInc vs MCIC are summarized in Table 5. Only four methods
29 managed to predict conversion slightly more accurately than a random classifier but none of them
30 got significantly better results ($p>0.05$). *Thickness-Direct* reached 32% sensitivity and 91%
31 specificity. *Voxel-STAND* reached 57% sensitivity and 78% specificity, *Voxel-COMPARE* reached
32 62% sensitivity and 67% specificity. *Hippo-Volume* distinguished MCIC from MCInc with 62%
33 sensitivity and 69% specificity.

34

35 **Influence of the preprocessing**

36 To evaluate the impact of the registration step, we tested both the registration using SPM5 unified
37 segmentation and the registration DARTEL as described in the previous section. The influence of
38 the registration step on the classification results is illustrated on Figures 2 and 3. The performances
39 obtained for the MCInc vs MCIC experiment were too low to be used to evaluate the impact of the
40 registration step. Therefore we did not take them into account for this comparison. The use of the
41 diffeomorphic registration algorithm DARTEL significantly improved the results of six out of 20
42 classification experiments ($p<0.05$). On the other hand, it led to significantly worse results in two
43 cases. According to the results in Tables 4, 5, 6, the use of customized tissue probability maps for
44 the registration with SPM5 unified segmentation did not improve the results of *Voxel-STAND*.

45 We also compared the classification obtained with only the GM maps to those with GM,
46 WM and CSF maps. Results are presented on Figures 2 and 3. The use of all three maps led to
47 significantly worse results ($p<0.05$) for two out of 20 classification experiments (*Voxel-*
48 *Direct_VOI-S* and *Voxel-COMPARE-D*). It never led to significantly better results.

49

50 **Complementariness of the methods**

51 The different approaches tested tackle the classification problem with various angles and could thus
52 be complementary. In order to quantify their similarity, we used the Jaccard similarity coefficient

53 (Jaccard, 1901; Shattuck et al., 2001). In this case, the Jaccard index of two methods is the number
54 of subjects correctly classified by both methods divided by the number of subjects correctly
55 classified by at least one of the two methods. Results are presented on Figures S1 and S2. All
56 methods are in at least substantial agreement (Jaccard over 0.6) and most of them are in strong
57 agreement. The most different results were obtained with the methods relying on the hippocampus.

58 We tested the combination of three approaches, one of each strategy: *Voxel-Direct-D-gm*,
59 *Thickness-Atlas* and *Hippo-Volume-S*. A convenient approach to combine different SVM-based
60 methods is to consider that the resulting classifier is a SVM which kernel is a linear convex
61 combination of the kernels of each method. The problem of learning both the coefficients of the best
62 convex linear combination of kernels and the optimal margin hyperplane (OMH) is known as the
63 multiple kernel learning (MKL) problem (Lanckriet et al., 2004; Bach et al., 2004; Sonnenburg et
64 al., 2006). We used the SimpleMKL toolbox (Rakotomamonjy et al., 2008). All four possible
65 combinations have been tested. The kernels are normalized with the trace of the Gram matrix of the
66 training set. Note that for *Hippo-Volume-S*, the Parzen window classifier is replaced by a linear
67 SVM.

68 None of these four combinations improved the accuracy in the CN vs AD experiment. Only
69 the combination of *Hippo-Volume-S* and *Thickness-Atlas* improved only slightly the accuracy for
70 the CN vs MCIc and the MCIc vs MCIc experiments. It distinguished MCIc from CN with 76%
71 sensitivity and 85% specificity. The optimal coefficients of the linear combination were 0.057 and
72 0.943 for the kernels of *Hippo-Volume-S* and *Thickness-Atlas*, respectively. This combination
73 classified MCIc and MCIc with 43% sensitivity and 83% specificity. The optimal coefficients of
74 the linear combination were 0.030 and 0.970 respectively.

75

76 **Influence of age and gender on classification results**

77 We investigated whether the age of the subjects influences the classification results. We thus
78 computed the average age of true positives, false positives, true negatives and false negatives.

79 Overall, we found that the false positives were often older than the true negatives, meaning that the
80 oldest controls were more often misclassified. Specifically, this was the case for 25 methods over 28
81 for CN *vs* AD and 24/28 for CN *vs* MCIc. Conversely, false negatives were often younger than the
82 true positives, meaning that the youngest patients were more often misclassified. Specifically, this
83 was the case for 26 methods over 28 for CN *vs* AD and 28/28 for CN *vs* MCIc. The number of
84 misclassified subjects was too small to test for statistical significance of these differences. However,
85 the fact that this difference was present for the vast majority of method suggests that it may not be
86 due to chance. We also investigated the influence of gender but did not find any difference.

87

88 **Computation time**

89 The computations were carried out with a processor running at 3.6 GHz with 2GB of RAM. Table 7
90 presents, for each method, the order of magnitude of the computation time (i.e. minutes, hours,
91 days, weeks). For each method, we report the computation time of its three main phases: the feature
92 computation step (segmentation, registration), the building of the classifier (including the grid
93 search for the optimization of the hyperparameters and the learning of the classifier), and the
94 classification of a new subject.

95 The order of magnitude of the computation time for the tissue segmentation and the
96 registration step per subject is respectively about ten minutes and an hour with SPM5 and
97 DARTEL. The cortical thickness computation and the registration of a single subject with
98 FreeSurfer take roughly a day. The segmentation of the hippocampi of a subject lasts a few minutes
99 and the shape analysis process with the SPHARM decomposition about one hour. The tuning of
100 parameters and learning phase took from a few minutes to several weeks for the *Voxel-STAND* and
101 *Voxel-COMPARE* methods. Once the hyperparameters are set and the learning is done, it takes at
102 most minutes to classify a new subject.

103

104 **Optimal margin hyperplanes**

105 The classification function obtained with a linear SVM is the sign of the inner product of the
106 features with w , a vector orthogonal to the optimal margin hyperplane (OMH) (Vapnik 1995;
107 Shawe-Taylor and Cristianini, 2000, 2004; Schölkopf and Smola, 2001). Therefore if the i^{th}
108 component w_i of the vector w is small, the i^{th} feature will have a small influence on the
109 classification. Conversely, if w_i is large, the i^{th} feature will play an important role in the classifier.
110 When the input features are the voxels of the image, each component of w also corresponds to a
111 voxel. One can thus represent the values of w in the image space. Similarly, for the *Thickness*
112 methods, the values of w can be represented on the cortical surface. The values of the optimal
113 margin hyperplanes for the different methods are presented on Figures from 4 to 7. This allows a
114 qualitative comparison of the features used in the classifier. Our aim was not to perform a statistical
115 analysis of differences between groups - for example using permutation tests on the coefficients
116 (Mourão-Miranda et al., 2005).

117 Figures 4 and 5 show the OMH for CN vs AD and CN vs MCIc respectively for the *Voxel*
118 methods. Overall, the spatial patterns corresponding to CN vs AD and CN vs MCIc are similar. For
119 *Voxel-Direct-D-gm*, the main regions were the medial temporal lobe (hippocampus, amygdala and
120 the parahippocampal gyrus), the inferior and middle temporal gyri, the posterior cingulate gyrus and
121 the posterior middle frontal gyrus. To a lesser extent, the OMH also included the inferior parietal
122 lobule, the supramarginal gyrus, fusiform gyrus, the middle cingulate gyrus and in the thalamus.
123 When all three tissue maps were used, the CSF maps mirrored the GM map (the enlargement of the
124 ventricle mirroring GM reduction). This was also the case for part of the WM map, in particular in
125 the hippocampal region. When using SPM5 unified segmentation instead of DARTEL, voxels were
126 much more scattered and not grouped into anatomical regions except in the medial temporal lobe.
127 For the AAL atlas, regions included the hippocampus, the amygdala, the parahippocampal gyrus,
128 the cingulum, the middle and inferior temporal gyri and the superior and inferior frontal gyri. The
129 regions were very similar for the surface Atlas as shown on Figure 6. Regions corresponding to
130 *Thickness-Direct* (Figure 7) were more restricted: the entorhinal cortex, the parahippocampal gyrus

131 and to a lesser extent the lateral temporal lobe, the inferior parietal lobule and some prefrontal areas.

132

133

134 **Optimal parameters of the classifiers**

135 For each approach, the optimal values of the hyperparameters are summarized in Table S1. One

136 should note that the *Hippo-Volume* method has no hyperparameter.

137 **Discussion**

138 In this paper, we compared different methods for the classification of patients with AD and MCI
139 based on anatomical T1-weighted MRI. To evaluate and compare the performances of each method,
140 three classification experiments were performed: CN *vs* AD, CN *vs* MCIc and CN *vs* MCIinc. The
141 set of participants was randomly split up into two groups of the same size: a training set and a
142 testing set. For each approach, the optimal parameter values had been determined using a grid-
143 search and LOOCV on the training set. Those values were then used to train the classifier using the
144 training group; the performance of the resulting classifier was then evaluated on the testing set. In
145 this way, we obtained unbiased estimates of the performances of each method.

146

147 **Classification methods discriminate AD from normal aging**

148 All the classification methods that we tested in this paper achieved accuracies significantly better
149 than chance for the discrimination of patients with AD from normal aging. All methods except
150 *Voxel-COMPARE* and *Hippo* methods performed significantly better than chance for the
151 discrimination of patients with prodromal AD (MCIc) from normal aging. For AD *vs* CN, most
152 methods achieved high sensitivity and specificity. However, at the prodromal stage, their sensitivity
153 was substantially lower.

154 The classification results we obtained for AD *vs* CN with *Atlas* and *COMPARE* methods are
155 lower than those reported in the respective papers: 94% accuracy for the *COMPARE* method in (Fan
156 et al., 2008a) and 92% sensitivity and 97% specificity for the *Atlas* in Magnin et al. (2009). These
157 differences can be explained by several factors. First, in the original papers, the hyperparameters
158 were optimized on the testing set. This may lead to overfitting the testing set and thus to
159 overestimate the sensitivity and specificity. On the contrary, in our evaluation, the learning step as
160 well as the optimization of the hyperparameters had been carried out on a training set and the
161 evaluation of the performance on a completely separated testing set. Thus our evaluation was
162 unbiased. Another explanation may stem from differences between studied populations (sample

163 size, stage of the disease). In particular, the ADNI population includes a large number of subjects
164 with vascular lesions, which was not the case in Magnin et al. (2009). Finally, the image
165 preprocessing step may also have an impact on the classification results. Davatzikos et al. (2008b)
166 and Fan et al. (2008b) used the RAVENS maps (Goldszal et al., 1998), thus the registration and the
167 segmentation step was different and might lead to different classification results. However, the aim
168 of the present paper was to compare different classification strategies and it was thus necessary to
169 use the same preprocessing for all methods. Since most of them relied on SPM, we chose to use this
170 preprocessing for all methods. It is possible that using other registration approaches such as
171 HAMMER would increase the classification performance but this is beyond the scope of this paper.

172 For the *Voxel-STAND* and *Voxel-Direct* methods, our results were similar to those reported in
173 the original papers by Vemuri et al. in (2008) and Klöppel et al. (2008). This can probably be
174 explained by the fact that Vemuri et al.'s (2008) evaluation procedure is also based on independent
175 testing group and that Klöppel et al. (2008) did not mention any optimization of the
176 hyperparameters. As for the *Thickness-ROI*, the results (69% sensitivity and 94% specificity) were
177 lower than those obtained by Desikan et al. (2009) (100% specificity and sensitivity). A possible
178 explanation is that in their study the classifier was trained on a different population (patients with
179 CDR=0.5) selected from a different database (the OASIS database).

180 The results obtained with Hippo-Volume were similar to those that we previously reported
181 for the ADNI database (Chupin et al., 2009b). The sensitivities and specificities were however
182 lower than those found in our previous study on a different population (Colliot et al., 2008) (84%
183 sensitivity and specificity for CN vs AD). This can be explained by several factors (Chupin et al.,
184 2009b). First ADNI is a multi-site database whereas the data in the previous study came from a
185 single scanner. Moreover the population included a large number of subjects with vascular lesions.
186 The slight difference between the results obtained in (Chupin et al., 2009b) and the present results
187 mostly comes from the difference in the accuracy estimation: two separate groups instead of a
188 LOOCV procedure. As for the *Hippo-Shape* method the results were substantially lower than our

189 results reported in (Gerardin et al., 2009) (86% for CN vs AD). This may result from the relatively
190 small number of subjects used in our previous study. Besides, the estimation was carried out with a
191 LOOCV. Moreover, this can also be due to that fact that all subjects were considered without taking
192 into consideration the quality control (Chupin et al., 2009b) of the hippocampus segmentation.

193 To our knowledge, the classification CN vs MCIc has only been addressed by Desikan et al.
194 (2009). Davatzikos et al. (2008a) and Fan et al. (2008a, 2008b) have performed the classification
195 CN vs MCI with no distinction between converters and non-converters. The MCI group did not
196 include only prodromal AD, hence the classification experiment cannot be compared to CN vs
197 MCIc. Desikan et al. (2009) classified CN and MCI who converted within two years after baseline
198 with 91% accuracy. This is substantially higher than the results obtained in our paper with the same
199 method *Thickness-ROI* (65% sensitivity and 94% specificity).

200

201 **Prediction of conversion in MCI patients**

202 No method was able to predict conversion better than chance. The three most accurate methods
203 were: *Voxel-STAND* (57% sensitivity and 78% specificity), *Voxel-COMPARE* (62% sensitivity and
204 67% specificity) and *Hippo-Volume* (62% sensitivity and 69% specificity). These three methods
205 restricted their search to a portion of the brain. In *Voxel-STAND* and *Voxel-COMPARE*, this was
206 done using feature selection: the selected regions are mainly in the medial temporal structures. In
207 *Hippo-Volume*, this was done by considering only the hippocampus.

208 Even for these three methods, the performances remained particularly low. The main reason
209 is certainly that MCI non converters are a very heterogeneous group: some patients would convert
210 shortly after the end of the follow-up and are thus in fact prodromal AD patients while others would
211 remain stable for a long period of time. We thus advocate that classification methods should be
212 focused on the detection of prodromal AD (i.e. MCI converters) which is a much better defined
213 entity.

214 To our knowledge, the classification MCIc vs MCIc has only been addressed by Misra et

215 al. (2009) and Querbes et al. (2009). Misra et al. (2009) considered the conversion within 12
216 months and Querbes et al. (2009) within 24 months. They obtained substantially higher accuracy:
217 respectively 81.5% and 76% accuracy. Misra et al. (2009) used the COMPARE (Fan et al., 2007)
218 classification methods. The differences may result from the same reasons as explained in the
219 previous paragraph: the use of separate training and testing sets and differing preprocessing steps.
220 Querbes et al. (2009) used a feature selection step, which may explain the slightly higher accuracy.

221

222 **Whole brain or hippocampus?**

223 For CN vs AD, methods using the whole brain (or the whole cortex) reached substantially higher
224 specificity (over 90%) than those based on the hippocampus (from 63% to 84%). For the detection
225 of prodromal AD, hippocampal-based approaches remained competitive with whole-brain methods.
226 It thus seems that considering the whole brain is advantageous mostly at the most advanced stages.
227 Indeed, at these more advanced stages, the atrophy is much more widespread. Moreover, it should
228 be noted that many subjects included in the ADNI have vascular lesions which may be, at least
229 partially, captured by whole brain methods. For intermediate stages, an alternative would be to
230 consider a set of selected regions instead of the whole brain or the hippocampus alone. For
231 example, *Thickness-ROI* performs at least as well as whole brain approaches for the detection of
232 prodromal AD. Even though they achieve lower accuracies, hippocampal-based methods may still
233 be of interest to the clinician because they provide a direct and easily interpretable index to the
234 clinician (the hippocampal volume) while the whole-brain approaches base their classification on a
235 complex combination of different regions.

236 All methods presented substantial agreement (Jaccard index over 0.6). The most different
237 results were obtained between hippocampal and whole brain methods. However, combining them
238 through multiple kernel learning did not improve the classification results.

239

240

241 **The registration step: is a fully deformable method advantageous?**

242 The use of DARTEL significantly improved the classification results in six cases, while it led to
243 lower results in only two cases. This is in line with other studies which reported that DARTEL led
244 to higher overlap values (Klein et al., 2009; Yassa and Stark, 2009) and higher sensitivity for voxel-
245 based morphometry (Bergouignan et al., 2009). In particular, the use of a fully deformable method
246 was advantageous for the medial temporal lobe as shown in (Yassa and Stark, 2009; Bergouignan et
247 al., 2009). Since the hippocampus is highly affected in AD, we expected that using a method which
248 registers the hippocampus better, would result in higher classification accuracy.

249

250 **Does adding WM and CSF maps increase the performance of the classifiers?**

251 In their original description, some of the tested methods used the three tissue (GM, WM and CSF)
252 maps (e.g. Vemuri et al., 2008, Fan et al., 2007, Magnin et al., 2009) while others used only the
253 GM maps (e.g. Klöppel et al., 2008). In this paper, we systematically tested whether the compared
254 methods performed better with the three maps or with only the GM maps. It should be noted that
255 this does not aim at assessing the diagnostic value of WM or CSF in general but only to test if
256 including all tissue maps is more effective for these particular classification approaches under study.
257 On the whole, adding the WM and the CSF probability maps did not improve the classification
258 performances. Adding WM and CSF maps increases the dimensionality of the feature space which
259 can make the classifier unstable and lead to overfitting the data. This problem is well-known in
260 machine learning as the curse of dimensionality. Besides, elder subjects are likely to have WM
261 structural abnormalities caused by leucoaraiosis or other diseases. Therefore adding WM tissues
262 may add noise in the features. Even if WM structural abnormalities alter (Levy-Cooperman et al.,
263 2008) the tissue segmentation step, GM probability maps are more robust features than WM tissues
264 probability maps.

265 Adding the WM and the CSF in the features may improve the results in two instances. The
266 first one is when the method encloses a feature selection step. Methods including feature selection

267 steps are more able to keep only the added value and avoid considering the noise but, overall, the
268 improvement is not substantial. Adding WM and CSF may also improve the results of methods
269 grouping the voxels into ROIs via wrapping a labelled atlas. It may make up for the parcellation
270 error due to the registration step but, again, the improvement is not substantial.

271

272 **Is it worth performing feature selection?**

273 The main objectives of the feature selection step are to keep only informative features and to reduce
274 the dimensionality of the feature space. In our evaluation, two methods included a feature selection
275 step: *Voxel-STAND* and *Voxel-COMPARE*. Overall, these methods did not perform substantially
276 better than simpler ones. In particular, their results might be more sensitive to the training set.
277 Indeed, feature selection can be regarded as a learning step. In such a case, the feature selection step
278 increases the class of all possible classification functions, which could lead to overfitting the data. A
279 more robust way to decrease the dimensionality of the features way would be to use more prior
280 knowledge of the disease.

281 Besides features selection can be time consuming as it adds new hyperparameters and thus
282 makes the gridsearch less tractable. Compared to *Voxel-Direct* and *Voxel-Atlas*, *Voxel-STAND* and
283 *Voxel-COMPARE* are time consuming (up to weeks), mostly because of the number of
284 hyperparameters to be tuned.

285 Nevertheless, feature selection proved useful in two specific cases. First, these methods
286 proved less sensitive when increasing the dimensionality of the feature space by adding WM and
287 CSF maps. They also tended to be more accurate for the MCIc vs MCInc experiment, where only a
288 few brain regions are informative.

289

290 **Does age influence the classification accuracy?**

291 Overall, we found that the oldest controls and the youngest patients were more often misclassified.
292 This may results from different causes. Normal aging is associated with atrophy of the grey and

293 white matter and increase of the CSF (Good et al., 2001; Salat et al., 2004). Moreover, aging is also
294 associated with alterations in tissue intensity and contrast, which can disrupt the segmentation step
295 and thus artificially increase the measured atrophy (Salat et al., 2009). Besides, elderly subjects are
296 more likely to have structural abnormalities of the white matter, which can also impede the tissue
297 segmentation step (Levy-Cooperman et al., 2008) and increase the measured atrophy. In addition,
298 elderly subjects have a propensity to suffer from mixed dementia (Zekry et al., 2002).

299

300 **Optimal margin hyperplanes**

301 In a linear SVM, the OMH can be easily represented. The OMH provides information about the
302 regions of the brain which were used by the classifier. It should be noted that this only provides
303 qualitative information on the hyperplanes, and that no statistical analysis of the OMH coefficients
304 was performed.

305 With *Voxel-Direct-D*, *Voxel-Atlas* and *Thickness-Atlas*, the regions in which atrophy
306 increased the likelihood of being classified as AD or MCIc were largely consistent with the pattern
307 of atrophy demonstrated in previous morphometric studies. These regions included the medial
308 temporal lobe, the inferior and middle temporal gyri (Chételat and Baron, 2003; Good et al., 2002;
309 Busatto et al., 2003; Rusinek et al., 2004; Tapiola et al., 2008), the posterior cingulate gyrus (Karas
310 et al., 2004; Chételat et al., 2005; Laakso et al., 1998) and the posterior middle frontal gyrus
311 (Whitwell et al., 2007), the fusiform gyrus, the thalamus (Karas et al., 2003, 2004; Chételat et al.,
312 2005). As for the cortical methods, the main regions in the medial temporal, middle and inferior
313 lateral temporal, inferior parietal, and posterior cingulated cortices and with a lesser extent parietal,
314 frontal, and lateral occipital cortices, which is consistent with the previous group studies based on
315 cortical thickness (Thompson et al., 2004; Lerch et al., 2005, 2008; McDonald et al., 2009).

316

317 In conclusion, we compared different automatic classification methods to assist in the early
318 diagnosis of Alzheimer's disease using the ADNI database. Most of them classify AD and CN with

319 high accuracy. However, at the prodromal stage, their sensitivity was substantially lower.
320 Combinations with other markers and/or more sophisticated prior knowledge seem necessary to be
321 able to detect prodromal AD with high accuracy.

322

323 **Acknowledgements**

324 Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging
325 Initiative (ADNI; Principal Investigator: Michael Weiner; NIH grant U01 AG024904). ADNI is
326 funded by the National Institute on Aging, the National Institute of Biomedical Imaging and
327 Bioengineering (NIBIB), and through generous contributions from the following: Pfizer Inc., Wyeth
328 Research, Bristol-Myers Squibb, Eli Lilly and Company, GlaxoSmithKline, Merck & Co. Inc.,
329 AstraZeneca AB, Novartis Pharmaceuticals Corporation, Alzheimer's Association, Eisai Global
330 Clinical Development, Elan Corporation plc, Forest Laboratories, and the Institute for the Study of
331 Aging, with participation from the U.S. Food and Drug Administration. Industry partnerships are
332 coordinated through the Foundation for the National Institutes of Health. The grantee organization
333 is the Northern California Institute for Research and Education, and the study is coordinated by the
334 Alzheimer's disease Cooperative Study at the University of California, San Diego. ADNI data are
335 disseminated by the Laboratory of Neuro Imaging at the University of California, Los Angeles.

1 **References**

- 2 Ashburner, J., Friston, K.J., 2005. Unified segmentation. *NeuroImage* 26 (3), 839-51.
- 3 Ashburner, J., 2007. A fast diffeomorphic image registration algorithm. *NeuroImage* 38 (1), 95-113.
- 4 Bach, F., Lanckriet, G., Jordan, M.I., 2004. Multiple kernel learning, conic duality, and the SMO
5 algorithm. In *Proceedings of the 21st International Conference on Machine Learning*, 41-8.
- 6 Bakkour, A., Morris, J.C., Dickerson, B.C., 2009. The cortical signature of prodromal AD: regional
7 thinning predicts mild AD dementia. *Neurology* 72 (12), 1048-55.
- 8 Bergouignan, L., Chupin, M., Czechowska, Y., Kinkingnéhun, S., Lemogne, C., Le Bastard, G.,
9 Lepage, M., Garnero, L., Colliot, O., Fossati P., 2009. Can voxel based morphometry, manual
10 segmentation and automated segmentation equally detect hippocampal volume differences in acute
11 depression? *NeuroImage* 45 (1), 29-37.
- 12 Blennow, K., de Leon, M.J., Zetterberg, H., 2006. Alzheimer's disease. *Lancet* 368 (9533), 387-403.
- 13 Busatto, G.F., Garrido, G.E., Almeida, O.P., Castro, C.C., Camargo, C.H., Cid, C.G., Buchpiguel,
14 C.A., Furuie, S., Bottino, C.M., 2003. A voxel-based morphometry study of temporal lobe gray
15 matter reductions in Alzheimer's disease. *Neurobiology of Aging* 24 (2), 221-31.
- 16 Chang, C.-C., Lin C.-J., 2001. LIBSVM: a library for support vector machines. Software available
17 at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.
- 18 Chételat, G., Baron, J.C., 2003. Early diagnosis of Alzheimer's disease: contribution of structural
19 neuroimaging. *NeuroImage* 18 (2), 525-41.
- 20 Chételat, G., Landeau, B., Eustache, F., Mézenge, F., Viader, F., de la Sayette, V., Desgranges, B.,
21 Baron, J.C., 2005. Using voxel-based morphometry to map the structural changes associated with
22 rapid conversion in MCI: a longitudinal MRI study. *NeuroImage* 27(4), 934-46.
- 23 Chupin, M., Mukuna-Bantumbakulu, A.R., Hasboun, D., Bardinet, E., Baillet, S., Kinkingnéhun, S.,
24 Lemieux, L., Dubois, B., Garnero, L., 2007. Automated segmentation of the hippocampus and the
25 amygdala driven by competition and anatomical priors: Method and validation on healthy subjects

26 and patients with Alzheimer's disease. *NeuroImage* 34, 996-1019.

27 Chupin, M., Hammers, A., Liu, R.S., Colliot, O., Burdett, J., Bardinnet, E., Duncan, J.S., Garnero, L.,
28 Lemieux, L., 2009a. Automatic segmentation of the hippocampus and the amygdala driven by
29 hybrid constraints: method and validation. *NeuroImage* 46 (3), 749-61.

30 Chupin, M., Gérardin, E., Cuingnet, R., Boutet, C., Lemieux, L., Lehericy, S., Benali, H., Garnero,
31 L., Colliot, O., Alzheimer's Disease Neuroimaging Initiative, 2009b. Fully automatic hippocampus
32 segmentation and classification in Alzheimer's disease and mild cognitive impairment applied on
33 data from ADNI. *Hippocampus* 19 (6), 579-87.

34 Colliot, O., Chételat, G., Chupin, M., Desgranges, B., Magnin, B. Benali, H., Dubois, B., Garnero,
35 L., Eustache, F., Lehericy, S., 2008. Discrimination between Alzheimer disease, mild cognitive
36 impairment, and normal aging by using automated segmentation of the hippocampus. *Radiology*
37 248 (1), 194-201.

38 Convit, A., De Leon, M.J., Tarshish, C., De Santi, S., Tsui, W., Rusinek, H., George, A., 1997.
39 Specific hippocampal volume reductions in individuals at risk for Alzheimer's disease.
40 *Neurobiology of Aging* 18 (2), 131-8.

41 Convit, A., de Asis, J., de Leon, M.J., Tarshish, C.Y., De Santi, S., Rusinek, H., 2000. Atrophy of
42 the medial occipitotemporal, inferior, and middle temporal gyri in non-demented elderly predict
43 decline to Alzheimer's disease. *Neurobiology of Aging* 21 (1), 19-26.

44 Dale, A.M., Fischl, B., Sereno, M.I., 1999. Cortical surface-based analysis. I. Segmentation and
45 surface reconstruction. *NeuroImage* 9 (2), 179-94.

46 Davatzikos, C., Fan, Y., Wu, X., Shen, D., Resnick, S.M., 2008a. Detection of prodromal
47 Alzheimer's disease via pattern classification of magnetic resonance imaging. *Neurobiology of*
48 *Aging* 29 (4), 514-23.

49 Davatzikos, C., Resnick, S.M., Wu, X., Parnpi, P., Clark, C.M., 2008b. Individual patient diagnosis
50 of AD and FTD via high-dimensional pattern classification of MRI. *NeuroImage* 41 (4), 1220-7.

51 Desikan, R.S., Segonne, F., Fischl, B., Quinn, B.T., Dickerson, B.C., Blacker, D., Buckner, R.L.,

52 Dale, A.M., Maguire, R.P., Hyman, B.T., Albert, M.S., Killiany, R.J., 2006. An automated labeling
53 system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest.
54 *NeuroImage* 31 (3), 968-80.

55 Desikan, R.S., Cabral, H.J., Hess, C.P., Dillon, W.P., Glastonbury, C.M., Weiner, M.W.,
56 Schmansky, N.J., Greve, D.N., Salat, D.H., Buckner, R.L., Fischl, B. and Alzheimer's Disease
57 Neuroimaging Initiative, 2009. Automated MRI measures identify individuals with mild cognitive
58 impairment and Alzheimer's disease. *Brain* 132 (8), 2048-57.

59 Dickerson, B.C., Bakkour, A., Salat, D.H., Feczko, E., Pacheco, J., Greve, D.N., Grodstein, F.,
60 Wright, C.I., Blacker, D., Rosas, H.D., Sperling, R.A., Atri, A., Growdon, J.H., Hyman, B.T.,
61 Morris, J.C., Fischl, B., Buckner, R.L., 2009. The cortical signature of Alzheimer's disease:
62 regionally specific cortical thinning relates to symptom severity in very mild to mild AD dementia
63 and is detectable in asymptomatic amyloid-positive individuals. *Cerebral Cortex* 19 (3), 497-510.

64 Dubois B., Albert M.L., 2004. Amnestic MCI or prodromal Alzheimer's disease? *The Lancet*
65 *Neurology* 3 (4), 246-48.

66 Dubois B., Feldman H.H., Jacova C., DeKosky S.T., Barberger-Gateau P., Cummings J., Delacourte
67 A., Galasko D., Gauthier S., Jicha G., Meguro K., O'Brien J., Pasquier F., Robert P., Rossor M.,
68 Salloway S., Stern Y., Visser P.J., Scheltens P., 2007. Research criteria for the diagnosis of
69 Alzheimer's disease: revising the NINCDS-ADRDA criteria. *The Lancet Neurology* 6 (8), 734-46.

70 Fan, Y., Shen, D., Davatzikos, C., 2005. Classification of structural images via high-dimensional
71 image warping, robust feature extraction, and SVM. In *Proceedings of the 8th International*
72 *Conference on Medical Image Computing and Computer-Assisted Intervention* 8 (Pt 1), 1-8.

73 Fan, Y., Shen, D., Gur, R.C., Davatzikosa, C., 2007. COMPARE: classification of morphological
74 patterns using adaptive regional elements. *IEEE Transactions on Medical Imaging* 26 (1), 93-105.

75 Fan, Y., Batmanghelich, N., Clark, C.M., Davatzikos, C., and the Alzheimer's Disease
76 Neuroimaging Initiative, 2008a. Spatial patterns of brain atrophy in MCI patients, identified via
77 high-dimensional pattern classification, predict subsequent cognitive decline. *NeuroImage* 39 (4),

78 1731-43.

79 Fan, Y., Resnick, S.M., Wu, X., Davatzikos, C., 2008b. Structural and functional biomarkers of
80 prodromal Alzheimer's disease: a high-dimensional pattern classification study. *NeuroImage* 41 (2),
81 277-85.

82 Fischl, B., Sereno, M.I., Dale, A.M., 1999a. Cortical surface-based analysis. II: Inflation, flattening,
83 and a surface based coordinate system. *NeuroImage* 9 (2), 195-207.

84 Fischl, B., Sereno, M.I., Tootell, R.B., Dale, A.M., 1999b. High-resolution intersubject averaging
85 and a coordinate system for the cortical surface. *Human Brain Mapping* 8 (4), 272-84.

86 Fischl, B., Dale, A.M., 2000. Measuring the thickness of the human cerebral cortex from magnetic
87 resonance images. *Proceedings of the National Academy of Sciences of the United States of*
88 *America* 97, 11050-11055.

89 Fox, N.C., Schott, J.M., 2004. Imaging cerebral atrophy: normal ageing to Alzheimer's disease. *The*
90 *Lancet* 363 (9406), 392-4.

91 Gerardin, E., Chételat, G., Chupin, M., Cuingnet, R., Desgranges, B., Kim, H.-S., Niethammer, M.,
92 Dubois, B., Lehericy, S., Garnero, L., Francis, E., Colliot, O., 2009. Multidimensional classification
93 of hippocampal shape features discriminates Alzheimer's disease and mild cognitive impairment
94 from normal aging. *NeuroImage* 47 (4), 1476-86.

95 Goldszal, A.F., Davatzikos, C., Pham, D.L., Yan, M.X., Bryan, R.N., Resnick, S.M., 1998. An
96 image-processing system for qualitative and quantitative volumetric analysis of brain images.
97 *Journal of Computer Assisted Tomography* 22 (5), 827-37.

98 Good, C.D., Johnsrude, I.S., Ashburner, J., Henson, R.N., Friston, K.J., Frackowiak, R.S., 2001. A
99 voxel-based morphometric study of ageing in 465 normal adult human brains. *NeuroImage* 14 (1),
100 21-36.

101 Good, C.D., Scahill, R.I., Fox, N.C., Ashburner, J., Friston, K.J., Chan, D., Crum, W.R., Rossor,
102 M.N., Frackowiak, R.S., 2002. Automatic differentiation of anatomical patterns in the human brain:
103 validation with studies of degenerative dementias. *NeuroImage* 17 (1), 29-46.

104 Hajnal J.V., Hill D.L.G., Hawkes D.J., 2001. Medical image registration. New York: CRC Press.

105 Hinrichs, C., Singh, V., Mukherjee, L., Xu, G., Chung, M.K., Johnson, S.C., and the Alzheimer's
106 Disease Neuroimaging Initiative, 2009. Spatially augmented LPboosting for AD classification with
107 evaluations on the ADNI dataset. *NeuroImage* 48 (1), 138-49.

108 Hua, X., Lee, S., Yanovsky, I., Leow, A.D., Chou, Y.Y., Ho, A.J., Gutman, B., Toga, A.W., Jack,
109 C.R.Jr, Bernstein, M.A., Reiman, E.M., Harvey, D.J., Kornak, J., Schuff, N., Alexander, G.E.,
110 Weiner, M.W., Thompson, P.M., and the Alzheimer's Disease Neuroimaging Initiative, 2009.
111 Optimizing power to track brain degeneration in Alzheimer's disease and mild cognitive impairment
112 with tensor-based morphometry: an ADNI study of 515 subjects. *NeuroImage* 48 (4), 668-81.

113 Jaccard P., 1901. Etude Comparative de la Distribution Florale dans une Portion des Alpes et du
114 Jura, *Bulletin de la Société vaudoise des Sciences Naturelles* 37, 547-79.

115 Jack, C.R.Jr, Petersen, R.C., Xu, Y.C., Waring, S.C., O'Brien P.C., Tangalos, E.G., Smith, G.E.,
116 Ivnik, R.J., Kokmen, E., 1997. Medial temporal atrophy on MRI in normal aging and very mild
117 Alzheimer's disease. *Neurology* 49 (3), 786-94.

118 Jack, C.R.Jr., Petersen, R.C., Xu, Y., O'Brien, P.C., Smith, G.E., Ivnik, R.J., Tangalos, E.G.,
119 Kokmen, E., 1998. Rate of medial temporal lobe atrophy in typical aging and Alzheimer's disease.
120 *Neurology* 51(4), 993-9.

121 Jack, C.R.Jr, Bernstein, M.A., Fox, N.C., Thompson, P., Alexander, G., Harvey, D., Borowski, B.,
122 Britson, P.J., L Whitwell, J., Ward, C., Dale, A.M., Felmlee, J.P., Gunter, J.L., Hill, D.L., Killiany,
123 R., Schuff, N., Fox-Bosetti, S., Lin, C., Studholme, C., DeCarli, C.S., Krueger, G., Ward, H.A.,
124 Metzger, G.J., Scott, K.T., Mallozzi, R., Blezek, D., Levy, J., Debbins, J.P., Fleisher, A.S., Albert
125 M., Green, R., Bartzokis, G., Glover, G., Mugler, J., Weiner, M.W., 2008. The Alzheimer's Disease
126 Neuroimaging Initiative (ADNI): MRI methods. *Journal of Magnetic Resonance Imaging* 27(4),
127 685-91.

128 Jagust, W., 2006. Positron emission tomography and magnetic resonance imaging in the diagnosis
129 and prediction of dementia. *Alzheimers Dementia* 2, 36-42.

130 Jovicich J., Czanner S., Greve D., Haley E., van der Kouwe A., Gollub R., Kennedy D., Schmitt F.,
131 Brown G., MacFall J., Fischl B., Dale A., 2006. Reliability in multi-site structural MRI studies:
132 effects of gradient non-linearity correction on phantom and human data. *NeuroImage*, 30, 436-43.

133 Juottonen, K., Laakso, M.P., Insausti, R., Lehtovirta, M., Pitkänen, A., Partanen, K., Soininen, H.,
134 1998. Volumes of the entorhinal and perirhinal cortices in Alzheimer's disease. *Neurobiology of*
135 *Aging* 19 (1), 15-22.

136 Karas, G.B., Burton, E.J., Rombouts, S.A., van Schijndel, R.A., O'Brien, J.T., Scheltens, P.,
137 McKeith, I.G., Williams, D., Ballard, C., Barkhof, F., 2003. A comprehensive study of gray matter
138 loss in patients with Alzheimer's disease using optimized voxel-based morphometry. *NeuroImage*
139 18 (4), 895-907.

140 Karas, G.B., Scheltens, P., Rombouts, S.A., Visser, P.J., van Schijndel, R.A., Fox, N.C., Barkhof F.,
141 2004. Global and local gray matter loss in mild cognitive impairment and Alzheimer's disease.
142 *NeuroImage* 23 (2), 708-16.

143 Klein, A., Andersson, J., Ardekani, B.A., Ashburner, J., Avants, B., Chiang, M.C., Christensen,
144 G.E., Collins, D.L., Gee, J., Hellier, P., Song, J.H., Jenkinson, M., Lepage, C., Rueckert, D.,
145 Thompson, P., Vercauteren, T., Woods, R.P., Mann, J.J., Parsey, R.V., 2009. Evaluation of 14
146 nonlinear deformation algorithms applied to human brain MRI registration. *NeuroImage* 46 (3),
147 786-802.

148 Klöppel, S., Stonnington, C.M., Chu, C., Draganski, B., Scahill, R.I., Rohrer, J.D., Fox, N.C.,
149 Jack, C.R.Jr., Ashburner, J., Frackowiak, R.S.J., 2008. Automatic classification of MR scans in
150 Alzheimer's disease. *Brain* 131 (3), 681-9.

151 Laakso, M.P., Soininen, H., Partanen, K., Lehtovirta, M., Hallikainen, M., Hänninen, T., Helkala,
152 E.L., Vainio, P., Riekkinen, P.J.Sr., 1998. MRI of the hippocampus in Alzheimer's disease:
153 sensitivity, specificity, and analysis of the incorrectly classified subjects. *Neurobiology of Aging* 19
154 (1), 23-31.

155 Laakso, M.P., Frisoni, G.B., Könönen, M., Mikkonen, M., Beltramello, A., Geroldi, C., Bianchetti,

156 A., Trabucchi, M., Soininen, H., Aronen H.J., 2000. Hippocampus and entorhinal cortex in
157 frontotemporal dementia and Alzheimer's disease: a morphometric MRI study. *Biol Psychiatry* 47
158 (12), 1056-63.

159 Lanckriet, G.R.G., De Bie, T., Cristianini, N., Jordan, M.I., Noble, W.S., 2004. A statistical
160 framework for genomic data fusion. *Bioinformatics* 20 (16), 2626-35.

161 Lao Z., Shen D., Xue Z., Karacali B., Resnick S.M., Davatzikos C., 2004. Morphological
162 classification of brains via high-dimensional shape transformations and machine learning methods.
163 *NeuroImage* 21 (1), 46-57.

164 Lerch, J.P., Pruessner, J.C., Zijdenbos, A., Hampel, H., Teipel, S.J., Evans, A.C., 2005. Focal
165 decline of cortical thickness in Alzheimer's disease identified by computational neuroanatomy.
166 *Cerebral Cortex* 15 (7), 995-1001.

167 Lerch, J.P., Pruessner, J., Zijdenbos, A.P., Collins, D.L., Teipel, S.J., Hampel, H., Evans, A.C., 2008.
168 Automated cortical thickness measurements from MRI can accurately separate Alzheimer's patients
169 from normal elderly controls. *Neurobiology of Aging* 29 (1), 23-30.

170 Levy-Cooperman, N., Ramirez, J., Lobaugh, N.J., Black, S.E., 2008. Misclassified tissue volumes
171 in Alzheimer disease patients with white matter hyperintensities: importance of lesion segmentation
172 procedures for volumetric analysis. *Stroke* 39 (4), 1134-41.

173 Magnin, B., Mesrob, L., Kinkingnéhun, S., Péligrini-Issac, M., Colliot, O., Sarazin, M., Dubois, B.,
174 Lehericy, S., Benali, H., 2009. Support vector machine-based classification of Alzheimer's disease
175 from whole-brain anatomical MRI. *Neuroradiology* 51 (2), 73-83.

176 Marcus, D.S., Wang, T.H., Parker, J., Csernansky, J.G., Morris, J.C., Buckner, R.L., 2007. Open
177 Access Series of Imaging Studies (OASIS): cross-sectional MRI data in young, middle aged,
178 nondemented, and demented older adults. *Journal of Cognitive Neuroscience* 19 (9), 1498-507.

179 McDonald, C.R., McEvoy, L.K., Gharapetian, L., Fennema-Notestine, C., Hagler, D.J.Jr., Holland
180 D., Koyama, A., Brewer, J.B., Dale, A.M., and the Alzheimer's Disease Neuroimaging Initiative,
181 2009. Regional rates of neocortical atrophy from normal aging to early Alzheimer disease.

182 Neurology 73 (6), 457-65.

183 McKhann G., Drachman D., Folstein M., Katzman R., Price D., Stadlan E.M., 1984. Clinical
184 diagnosis of Alzheimer's disease: report of the NINCDS-ADRDA Work Group under the auspices
185 of Department of Health and Human Services Task Force on Alzheimer's Disease. Neurology
186 34 (7), 939-44.

187 Misra C., Fan Y., Davatzikos C., 2009. Baseline and longitudinal patterns of brain atrophy in MCI
188 patients, and their use in prediction of short-term conversion to AD: Results from ADNI.
189 NeuroImage 44 (4), 1415-22.

190 Morris, J.C., 1993. The Clinical Dementia Rating (CDR): current version and scoring rules.
191 Neurology 43, 2412-14.

192 Mourao-Miranda J., Bokde A.L.W., Born C., Hampel H., Stetter M., 2005. Classifying brain states
193 and determining the discriminating activation patterns: Support Vector Machine on functional MRI
194 data. NeuroImage 28 (4) 980-95.

195 Narayana P.A., Brey W.W., Kulkarni M.V., Sievenpiper C.L., 1988. Compensation for surface coil
196 sensitivity variation in magnetic resonance imaging. Magnetic Resonance Imaging 6, 271-74.

197 Petersen, R.C., Smith, G.E., Waring, S.C., Ivnik, R.J., Tangalos, E.G., Kokmen, E., 1999. Mild
198 Cognitive Impairment: Clinical Characterization and Outcome. Archives of Neurology 56, 303-308.

199 Querbes, O., Aubry, F., Pariente, J., Lotterie, J.-A., Démonet, J.-F., Duret, V., Puel, M., Berry, I.,
200 Fort, J.-C., Celsis, P., and the Alzheimer's Disease Neuroimaging Initiative, 2009. Early diagnosis
201 of Alzheimer's disease using cortical thickness: impact of cognitive reserve. Brain 32 (8), 2036-47.

202 Rakotomamonjy, A., Bach, F.R., Canu, S., Grandvalet, Y., 2008. SimpleMKL. Journal of Machine
203 Learning Research 9, 2491-2521.

204 Rusinek, H., Endo, Y., De Santi, S., Frid, D., Tsui, W.H., Segal, S., Convit, A., de Leon, M.J., 2004.
205 Atrophy rate in medial temporal lobe during progression of Alzheimer disease. Neurology 63 (12),
206 2354-9.

207 Salat, D.H., Buckner, R.L., Snyder, A.Z., Greve, D.N., Desikan, R.S., Busa, E., Morris, J.C., Dale,

208 A.M., Fischl, B., 2004. Thinning of the cerebral cortex in aging. *Cerebral Cortex* 14(7), 721-30

209 Salat, D.H., Lee, S.Y., van der Kouwe, A.J., Greve, D.N., Fischl, B., Rosas, H.D., 2009. Age-
210 associated alterations in cortical gray and white matter signal intensity and gray to white matter
211 contrast. *NeuroImage* 48 (1), 21-8.

212 Schölkopf, B., Smola, A. J., 2001. *Learning with Kernels*. MIT Press.

213 Shattuck D.W., Sandor-Leahy S.R., Schaper K.A., Rottenberg D.A., Leahy R.M., 2001.
214 M. Resonance Image Tissue Classification Using a Partial Volume Model. *NeuroImage* 13 (5), 856-
215 76

216 Shawe-Taylor, J., Cristianini, N., 2000. *Support Vector Machines and other kernel-based learning*
217 *methods*. Cambridge University Press.

218 Shawe-Taylor, J., Cristianini, N., 2004. *Kernel Methods for Pattern Analysis*. Cambridge
219 University Press.

220 Sled, J.G., Zijdenbos, A.P., Evans, A.C., 1998. A nonparametric method for automatic correction of
221 intensity nonuniformity in MRI data. *IEEE Transactions on Medical Imaging* 17, 87-97.

222 Sonnenburg, S., Rätsch, G., Schäfer, C., Schölkopf, B., 2006. Large Scale Multiple Kernel
223 Learning. *The Journal of Machine Learning Research archive* 7, 1531-65.

224 Tapiola, T., Pennanen, C., Tapiola, M., Tervo, S., Kivipelto, M., Hänninen, T., Pihlajamäki, M.,
225 Laakso, M.P., Hallikainen, M., Hämäläinen, A., Vanhanen, M., Helkala, E.L., Vanninen, R.,
226 Nissinen, A., Rossi, R., Frisoni, G.B., Soininen, H., 2008. MRI of hippocampus and entorhinal
227 cortex in mild cognitive impairment: a follow-up study. *Neurobiology of Aging* 29 (1), 31-8.

228 Thompson, P.M., Mega, M.S., Woods, R.P., Zoumalan, C.I., Lindshield, C.J., Blanton, R.E.,
229 Moussai, J., Holmes, C.J., Cummings, J.L., Toga, A.W., 2001. Cortical change in Alzheimer's
230 disease detected with a disease-specific population-based brain atlas. *Cerebral Cortex* 11(1), 1-16.

231 Thompson, P.M., Hayashi, K.M., de Zubicaray, G., Janke, A.L., Rose, S.E., Semple, J., Herman,
232 D., Hong, M.S., Dittmer, S.S., Doddrell, D.M., Toga, A.W., 2003. Dynamics of gray matter loss in
233 Alzheimer's disease. *The Journal of Neuroscience* 23 (3), 994-1005.

234 Thompson, P.M., Hayashi, K.M., Sowell, E.R., Gogtay, N., Giedd, J.N., Rapoport, J.L., de
235 Zubicaray, G.I., Janke, A.L., Rose, S.E., Semple, J., Doddrell, D.M., Wang, Y., van Erp, T.G.,
236 Cannon, T.D., Toga, A.W., 2004. Mapping cortical change in Alzheimer's disease, brain
237 development, and schizophrenia. *NeuroImage* 23 (Suppl 1), S2-18.

238 Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N.,
239 Mazoyer, B., Joliot, M., 2002. Automated anatomical labeling of activations in SPM using a
240 macroscopic anatomical parcellation of the MNI MRI single-subject brain. *NeuroImage* 15, 273-89.

241 Vapnik, V.N., 1995. *The Nature of Statistical Learning Theory*. Springer-Verlag New York Inc.

242 Vemuri, P., Gunter, J.L., Senjem, M.L., Whitwell, J.L., Kantarci, K., Knopman, D.S., Boeve, B.F.,
243 Petersen, R.C., Jack, C.R.Jr., 2008. Alzheimer's disease diagnosis in individual subjects using
244 structural MR images: validation studies. *NeuroImage* 39 (3), 1186-97.

245 Wechsler, D., 1987. *Manual for the Wechsler Memory Scale-Revised*. The Psychological
246 Corporation, San Antonio.

247 Whitwell, J.L., Przybelski, S.A., Weigand, S.D., Knopman, D.S., Boeve, B.F., Petersen, R.C., Jack,
248 C.R.Jr., 2007. *Brain* 130 (7), 1777-86.

249 Whitwell, J.L., Shiung, M.M., Przybelski, S.A., Weigand, S.D., Knopman, D.S., Boeve, B.F.,
250 Petersen, R.C., Jack, C.R.Jr., 2008. MRI patterns of atrophy associated with progression to AD in
251 amnesic mild cognitive impairment. *Neurology* 70 (7), 512-20.

252 Xu, Y., Jack, C.R.Jr, O'Brien, P.C., Kokmen, E., Smith, G.E., Ivnik, R.J., Boeve, B.F., Tangalos,
253 R.G., Petersen, R.C., 2000. Usefulness of MRI measures of entorhinal cortex versus hippocampus
254 in AD. *Neurology* 54 (9), 1760-7.

255 Yassa, M.A., Stark, C.E., 2009. A quantitative evaluation of cross-participant registration
256 techniques for MRI studies of the medial temporal lobe. *NeuroImage* 44 (2), 319-27.

257 Ye J., Chen K., Wu T., Li J., Zhao Z., Patel R., Bae M., Janardan R., Liu H., Alexander G., and
258 Reiman E., 2008. Heterogeneous data fusion for alzheimer's disease study. In *Proceeding of the*
259 *14th ACM SIGKDD international Conference on Knowledge Discovery and Data Mining KDD '08*,

260 1025-33.

261 Zekry, D., Hauw, J.J., Gold, G., 2002. Mixed dementia: epidemiology, diagnosis, and treatment.

262 Journal of the American Geriatrics Society 8, 1431-8.

263

264

265

266 **List of tables**

267 **Table 1.** Demographic characteristics of the studied population (from the ADNI database). Values
 268 are indicated as mean \pm standard-deviation [range].

269

Group	Diagnostic	Number	Age	Gender	MMS	# Centers
Whole set	CN	162	76.3 \pm 5.4 [60 - 90]	76 M / 86 F	29.2 \pm 1.0 [25 - 30]	40
	AD	137	76.0 \pm 7.3 [55 - 91]	67 M / 70 F	23.2 \pm 2.0 [18 - 27]	39
	MCIc	76	74.8 \pm 7.4 [55 - 88]	43 M / 33 F	26.5 \pm 1.9 [23 - 30]	30
	MCInc	134	74.5 \pm 7.2 [58 - 88]	84 M / 50 F	27.2 \pm 1.7 [24 - 30]	36
Training set	CN	81	76.1 \pm 5.6 [60 - 89]	38 M / 43 F	29.2 \pm 1.0 [25 - 30]	35
	AD	69	75.8 \pm 7.5 [55 - 89]	34 M / 35 F	23.3 \pm 1.9 [18 - 26]	32
	MCIc	39	74.7 \pm 7.8 [55 - 88]	22 M / 17 F	26.0 \pm 1.8 [23 - 30]	21
	MCInc	67	74.3 \pm 7.3 [58 - 87]	42 M / 25 F	27.1 \pm 1.8 [24 - 30]	30
Testing set	CN	81	76.5 \pm 5.2 [63 - 90]	38 M / 43 F	29.2 \pm 0.9 [26 - 30]	35
	AD	68	76.2 \pm 7.2 [57 - 91]	33 M / 35 F	23.2 \pm 2.1 [20 - 27]	33
	MCIc	37	74.9 \pm 7.0 [57 - 87]	21 M / 16 F	26.9 \pm 1.8 [24 - 30]	24
	MCInc	67	74.7 \pm 7.3 [58 - 88]	42 M / 25 F	27.3 \pm 1.7 [24 - 30]	31

270

271

272

273 **Table 2.** Summary of the approaches tested in this study.

274

FEATURES		SEGMENTATION REGISTRATION	TISSUES PROBABILITY MAPS	CLASSIFIER	METHOD #	METHOD'S NAME	
Voxel - segmented tissue probability maps	<i>Direct</i>	DARTEL	GM GM + WM + CSF	Linear SVM Linear SVM	1.1.1 a 1.1.1 b	<i>Voxel-Direct-D-gm</i> <i>Voxel-Direct-D-all</i>	
		SPM5	GM GM + WM + CSF	Linear SVM Linear SVM	1.1.2 a 1.1.2 b	<i>Voxel-Direct-S-gm</i> <i>Voxel-Direct-S-all</i>	
	<i>Direct VOI</i>	DARTEL	GM GM + WM + CSF	Linear SVM Linear SVM	1.2.1 a 1.2.1 b	<i>Voxel-Direct_VOI-D-gm</i> <i>Voxel-Direct_VOI-D-all</i>	
		SPM5	GM GM + WM + CSF	Linear SVM Linear SVM	1.2.2 a 1.2.2 b	<i>Voxel-Direct_VOI-S-gm</i> <i>Voxel-Direct_VOI-S-all</i>	
	<i>STAND- score</i>	DARTEL	GM GM + WM + CSF	Linear SVM Linear SVM	1.3.1 a 1.3.1 b	<i>Voxel-STAND-D-gm</i> <i>Voxel-STAND-D-all</i>	
		SPM5	GM GM + WM + CSF	Linear SVM Linear SVM	1.3.2 a 1.3.2 b	<i>Voxel-STAND-S-gm</i> <i>Voxel-STAND-S-all</i>	
		SPM5 custom template	GM GM + WM + CSF	Linear SVM Linear SVM	1.3.3 a 1.3.3 b	<i>Voxel-STAND-Sc-gm</i> <i>Voxel-STAND-Sc-all</i>	
	<i>Atlas</i>	DARTEL	GM GM + WM + CSF	Linear SVM Linear SVM	1.4.1 a 1.4.1 b	<i>Voxel-Atlas-D-gm</i> <i>Voxel-Atlas-D-all</i>	
		SPM5	GM GM + WM + CSF	Linear SVM Linear SVM	1.4.2 a 1.4.2 b	<i>Voxel-Atlas-S-gm</i> <i>Voxel-Atlas-S-all</i>	
	<i>COMPARE</i>	DARTEL	GM GM + WM + CSF	Linear SVM Linear SVM	1.5.1 a 1.5.1 b	<i>Voxel-COMPARE-D-gm</i> <i>Voxel-COMPARE-D-all</i>	
		SPM5	GM GM + WM + CSF	Gaussian SVM Gaussian SVM	1.5.2 a 1.5.2 b	<i>Voxel-COMPARE-S-gm</i> <i>Voxel-COMPARE-S-all</i>	
	Cortical thickness	<i>Direct</i>	Freesurfer	-	Linear SVM	2.1	<i>Thickness-Direct</i>
		<i>Atlas</i>	Freesurfer	-	Linear SVM	2.2	<i>Thickness-Atlas</i>
		<i>ROI</i>	Freesurfer	-	Logistic Reg.	2.3	<i>Thickness-ROI</i>
	Hippocampus	<i>Volume</i>	Freesurfer	-	Parzen	3.1.1	<i>Hippo-Volume-F</i>
		<i>Volume</i>	SACHA	-	Parzen	3.1.2	<i>Hippo-Volume-S</i>
<i>Shape</i>		Freesurfer	-	Linear SVM	3.2	<i>Hippo-Shape</i>	

275

276

277

278

279 **Table 3.** Contingency table for the McNemar test. *a*: number of subjects correctly classified by both
280 classifiers; *b*: number of subjects correctly classified by classifier 1 but misclassified by classifier 2;
281 *c*: number of subjects misclassified by classifier 1 but correctly classified by classifier 2; *d*: number
282 of subjects misclassified by both classifiers.

283

	Classifier 2: correctly classified	Classifier 2: misclassified
Classifier 1: correctly classified	a	b
Classifier 1: misclassified	c	d

284

285

METHOD #	METHOD'S NAME	CN vs AD				MCNEMAR TEST
		SENSITIVITY	SPECIFICITY	PPV	NPV	
1.1.1 a	<i>Voxel-Direct-D-gm</i>	81%	95%	93%	86%	p < 0.0001
1.1.1 b	<i>Voxel-Direct-D-all</i>	68%	98%	96%	78%	p < 0.0001
1.1.2 a	<i>Voxel-Direct-S-gm</i>	72%	89%	84%	79%	p < 0.0001
1.1.2 b	<i>Voxel-Direct-S-all</i>	65%	88%	81%	75%	p < 0.0001
1.2.1 a	<i>Voxel-Direct_VOI-D-gm</i>	71%	95%	92%	79%	p < 0.0001
1.2.1 b	<i>Voxel-Direct_VOI-D-all</i>	65%	95%	92%	76%	p < 0.0001
1.2.2 a	<i>Voxel-Direct_VOI-S-gm</i>	65%	91%	86%	76%	p < 0.0001
1.2.2 b	<i>Voxel-Direct_VOI-S-all</i>	59%	81%	73%	70%	p = 0.0012
1.3.1 a	<i>Voxel-STAND-D-gm</i>	69%	90%	85%	78%	p < 0.0001
1.3.1 b	<i>Voxel-STAND-D-all</i>	71%	91%	87%	79%	p < 0.0001
1.3.2 a	<i>Voxel-STAND-S-gm</i>	75%	91%	88%	81%	p < 0.0001
1.3.2 b	<i>Voxel-STAND-S-all</i>	75%	86%	82%	80%	p < 0.0001
1.3.3 a	<i>Voxel-STAND-Sc-gm</i>	72%	91%	88%	80%	p < 0.0001
1.3.3 b	<i>Voxel-STAND-Sc-all</i>	71%	91%	87%	79%	p < 0.0001
1.4.1 a	<i>Voxel-Atlas-D-gm</i>	78%	93%	90%	83%	p < 0.0001
1.4.1 b	<i>Voxel-Atlas-D-all</i>	81%	90%	87%	85%	p < 0.0001
1.4.2 a	<i>Voxel-Atlas-S-gm</i>	75%	93%	89%	82%	p < 0.0001
1.4.2 b	<i>Voxel-Atlas-S-all</i>	74%	93%	89%	81%	p < 0.0001
1.5.1 a	<i>Voxel-COMPARE-D-gm</i>	82%	89%	86%	86%	p < 0.0001
1.5.1 b	<i>Voxel-COMPARE-D-all</i>	69%	81%	76%	76%	p < 0.0001
1.5.2 a	<i>Voxel-COMPARE-S-gm</i>	66%	86%	80%	75%	p < 0.0001
1.5.2 b	<i>Voxel-COMPARE-S-all</i>	72%	91%	88%	80%	p < 0.0001
2.1	<i>Thickness-Direct</i>	74%	90%	86%	80%	p < 0.0001
2.2	<i>Thickness-Atlas</i>	79%	90%	87%	84%	p < 0.0001
2.3	<i>Thickness-ROI</i>	69%	94%	90%	78%	p < 0.0001
3.1.1	<i>Hippo-Volume-F</i>	63%	80%	73%	72%	p = 0.0007
3.1.2	<i>Hippo-Volume-S</i>	71%	77%	72%	76%	p = 0.0006
3.2	<i>Hippo-Shape</i>	69%	84%	78%	76%	p < 0.0001

291 **Table 5.** Classification results CN vs MCIC.

292

METHOD #	METHOD'S NAME	CN vs MCIC				MCNEMAR TEST
		SENSITIVITY	SPECIFICITY	PPV	NPV	
1.1.1 a	<i>Voxel-Direct-D-gm</i>	57%	96%	88%	83%	p = 0.00052
1.1.1 b	<i>Voxel-Direct-D-all</i>	49%	91%	72%	80%	p = 0.046
1.1.2 a	<i>Voxel-Direct-S-gm</i>	32%	96%	80%	76%	p = 0.039
1.1.2 b	<i>Voxel-Direct-S-all</i>	41%	94%	75%	78%	p = 0.044
1.2.1 a	<i>Voxel-Direct_VOI-D-gm</i>	54%	95%	83%	82%	p = 0.0022
1.2.1 b	<i>Voxel-Direct_VOI-D-all</i>	41%	96%	83%	78%	p = 0.0095
1.2.2 a	<i>Voxel-Direct_VOI-S-gm</i>	32%	88%	55%	74%	p = 0.83
1.2.2 b	<i>Voxel-Direct_VOI-S-all</i>	22%	99%	89%	73%	p = 0.046
1.3.1 a	<i>Voxel-STAND-D-gm</i>	73%	85%	69%	87%	p = 0.025
1.3.1 b	<i>Voxel-STAND-D-all</i>	65%	93%	80%	85%	p = 0.0019
1.3.2 a	<i>Voxel-STAND-S-gm</i>	59%	86%	67%	82%	p = 0.082
1.3.2 b	<i>Voxel-STAND-S-all</i>	49%	93%	75%	80%	p = 0.025
1.3.3 a	<i>Voxel-STAND-Sc-gm</i>	62%	85%	66%	83%	p = 0.091
1.3.3 b	<i>Voxel-STAND-Sc-all</i>	57%	90%	72%	82%	p = 0.026
1.4.1 a	<i>Voxel-Atlas-D-gm</i>	65%	80%	60%	83%	p = 0.27
1.4.1 b	<i>Voxel-Atlas-D-all</i>	54%	91%	74%	81%	p = 0.021
1.4.2 a	<i>Voxel-Atlas-S-gm</i>	68%	95%	86%	87%	p = 0.00020
1.4.2 b	<i>Voxel-Atlas-S-all</i>	59%	94%	81%	84%	p = 0.0021
1.5.1 a	<i>Voxel-COMPARE-D-gm</i>	49%	81%	55%	78%	p = 0.73
1.5.1 b	<i>Voxel-COMPARE-D-all</i>	51%	85%	61%	79%	p = 0.28
1.5.2 a	<i>Voxel-COMPARE-S-gm</i>	49%	78%	50%	77%	p = 0.87
1.5.2 b	<i>Voxel-COMPARE-S-all</i>	59%	78%	55%	81%	p = 0.64
2.1	<i>Thickness-Direct</i>	54%	96%	87%	82%	p = 0.00084
2.2	<i>Thickness-Atlas</i>	57%	93%	78%	82%	p = 0.0071
2.3	<i>Thickness-ROI</i>	65%	94%	83%	85%	p = 0.00083
3.1.1	<i>Hippo-Volume-F</i>	73%	74%	56%	86%	p = 0.47
3.1.2	<i>Hippo-Volume-S</i>	70%	73%	54%	84%	p = 0.67
3.2	<i>Hippo-Shape</i>	57%	88%	68%	82%	p = 0.072

293 PPV: positive predictive value; NPV: negative predictive value.

294

295

296

Table 6. Classification results MCInc vs MCIC.

METHOD #	METHOD'S NAME	MCInc vs MCIC				MCNEMAR TEST
		SENSITIVITY	SPECIFICITY	PPV	NPV	
1.1.1 a	<i>Voxel-Direct-D-gm</i>	0%	100%	-	64%	p = 1.0
1.1.1 b	<i>Voxel-Direct-D-all</i>	0%	100%	-	64%	p = 1.0
1.1.2 a	<i>Voxel-Direct-S-gm</i>	0%	100%	-	64%	p = 1.0
1.1.2 b	<i>Voxel-Direct-S-all</i>	0%	100%	-	64%	p = 1.0
1.2.1 a	<i>Voxel-Direct_VOI-D-gm</i>	43%	70%	44%	69%	p = 0.62
1.2.1 b	<i>Voxel-Direct_VOI-D-all</i>	0%	100%	-	64%	p = 1.0
1.2.2 a	<i>Voxel-Direct_VOI-S-gm</i>	0%	100%	-	64%	p = 1.0
1.2.2 b	<i>Voxel-Direct_VOI-S-all</i>	0%	100%	-	64%	p = 1.0
1.3.1 a	<i>Voxel-STAND-D-gm</i>	57%	78%	58%	76%	p = 0.40
1.3.1 b	<i>Voxel-STAND-D-all</i>	0%	100%	-	64%	p = 1.0
1.3.2 a	<i>Voxel-STAND-S-gm</i>	22%	91%	57%	68%	p = 0.79
1.3.2 b	<i>Voxel-STAND-S-all</i>	51%	79%	58%	75%	p = 0.49
1.3.3 a	<i>Voxel-STAND-Sc-gm</i>	35%	70%	39%	66%	p = 0.30
1.3.3 b	<i>Voxel-STAND-Sc-all</i>	41%	72%	44%	69%	p = 0.61
1.4.1 a	<i>Voxel-Atlas-D-gm</i>	0%	100%	-	64%	p = 1.0
1.4.1 b	<i>Voxel-Atlas-D-all</i>	0%	100%	-	64%	p = 1.0
1.4.2 a	<i>Voxel-Atlas-S-gm</i>	0%	100%	-	64%	p = 1.0
1.4.2 b	<i>Voxel-Atlas-S-all</i>	0%	100%	-	64%	p = 1.0
1.5.1 a	<i>Voxel-COMPARE-D-gm</i>	62%	67%	51%	76%	p = 1.0
1.5.1 b	<i>Voxel-COMPARE-D-all</i>	54%	78%	57%	75%	p = 0.50
1.5.2 a	<i>Voxel-COMPARE-S-gm</i>	32%	82%	50%	69%	p = 0.84
1.5.2 b	<i>Voxel-COMPARE-S-all</i>	51%	72%	50%	73%	p = 0.87
2.1	<i>Thickness-Direct</i>	32%	91%	67%	71%	p = 0.24
2.2	<i>Thickness-Atlas</i>	27%	85%	50%	68%	p = 0.82
2.3	<i>Thickness-ROI</i>	24%	82%	43%	66%	p = 0.66
3.1.1	<i>Hippo-Volume-F</i>	70%	61%	50%	79%	p = 0.89
3.1.2	<i>Hippo-Volume-S</i>	62%	69%	52%	77%	p = 0.88
3.2	<i>Hippo-Shape</i>	0%	100%	-	64%	p = 1.0

298

PPV: positive predictive value; NPV: negative predictive value.

299

300

301 **Table 7.** Order of magnitude of the computation time (i.e. minutes, hours, days, weeks) for each
302 method for its three main phases: feature computation step (segmentation, registration), building of
303 the classifier (including the grid search for the optimization of the hyperparameters and the learning
304 of the classifier), classification of a new subject. The computations have been carried out with a
305 processor running at 3.6 GHz with 2GB of RAM.

306

307

METHOD #	METHOD'S NAME	SEGMENTATION REGISTRATION	GRID SEARCH LEARNING	TESTING
1.1.1 a	<i>Voxel-Direct-D-gm</i>	Hour(s) per subject	Minute(s)	Minute(s)
1.1.1 b	<i>Voxel-Direct-D-all</i>	Hour(s) per subject	Minute(s)	Minute(s)
1.1.2 a	<i>Voxel-Direct-S-gm</i>	10 Minutes per subject	Minute(s)	Minute(s)
1.1.2 b	<i>Voxel-Direct-S-all</i>	10 Minutes per subject	Minute(s)	Minute(s)
1.2.1 a	<i>Voxel-Direct_VOI-D-gm</i>	Hour(s) per subject	Minute(s)	Minute(s)
1.2.1 b	<i>Voxel-Direct_VOI-D-all</i>	Hour(s) per subject	Minute(s)	Minute(s)
1.2.2 a	<i>Voxel-Direct_VOI-S-gm</i>	10 Minutes per subject	Minute(s)	Minute(s)
1.2.2 b	<i>Voxel-Direct_VOI-S-all</i>	10 Minutes per subject	Minute(s)	Minute(s)
1.3.1 a	<i>Voxel-STAND-D-gm</i>	Hour(s) per subject	Day(s)	Hour(s)
1.3.1 b	<i>Voxel-STAND-D-all</i>	Hour(s) per subject	Week(s)	Hour(s)
1.3.2 a	<i>Voxel-STAND-S-gm</i>	10 Minutes per subject	Day(s)	Hour(s)
1.3.2 b	<i>Voxel-STAND-S-all</i>	10 Minutes per subject	Week(s)	Hour(s)
1.3.3 a	<i>Voxel-STAND-Sc-gm</i>	20 Minutes per subject	Day(s)	Hour(s)
1.3.3 b	<i>Voxel-STAND-Sc-all</i>	20 Minutes per subject	Week(s)	Hour(s)
1.4.1 a	<i>Voxel-Atlas-D-gm</i>	Hour(s) per subject	Minute(s)	Minute(s)
1.4.1 b	<i>Voxel-Atlas-D-all</i>	Hour(s) per subject	Minute(s)	Minute(s)
1.4.2 a	<i>Voxel-Atlas-S-gm</i>	10 Minutes per subject	Minute(s)	Minute(s)
1.4.2 b	<i>Voxel-Atlas-S-all</i>	10 Minutes per subject	Minute(s)	Minute(s)
1.5.1 a	<i>Voxel-COMPARE-D-gm</i>	Hour(s) per subject	Week(s)	Hour(s)
1.5.1 b	<i>Voxel-COMPARE-D-all</i>	Hour(s) per subject	Week(s)	Hour(s)
1.5.2 a	<i>Voxel-COMPARE-S-gm</i>	10 Minutes per subject	Week(s)	Hour(s)
1.5.2 b	<i>Voxel-COMPARE-S-all</i>	10 Minutes per subject	Week(s)	Hour(s)
2.1	<i>Thickness-Direct</i>	Day(s) per subject	Minute(s)	Minute(s)
2.2	<i>Thickness-Atlas</i>	Day(s) per subject	Minute(s)	Minute(s)
2.3	<i>Thickness-ROI</i>	Day(s) per subject	Minute(s)	Seconds
3.1.1	<i>Hippo-Volume-F</i>	Day(s) per subject	Minute(s)	Seconds
3.1.2	<i>Hippo-Volume-S</i>	10 Minutes per subject	Minute(s)	Seconds
3.2	<i>Hippo-Shape</i>	Hour(s) per subject	Minute(s)	Seconds

308

309

310

311

312 **List of tables for Supplementary Materials**

313

314 **Table S1.** Optimal parameters values.

315

METHOD #	METHOD'S NAME	CN vs AD	CN vs MC1c	MC1nc vs MC1c
		OPTIMAL HYPERPARAMETERS	OPTIMAL HYPERPARAMETERS	OPTIMAL HYPERPARAMETERS
1.1.1 a	<i>Voxel-Direct-D-gm</i>	Log10(C) = -3.5	Log10(C) = -3.5	Log10(C) = -5.0
1.1.1 b	<i>Voxel-Direct-D-all</i>	Log10(C) = -4.5	Log10(C) = -4.0	Log10(C) = -5.0
1.1.2 a	<i>Voxel-Direct-S-gm</i>	Log10(C) = -3.5	Log10(C) = -3.5	Log10(C) = -5.0
1.1.2 b	<i>Voxel-Direct-S-all</i>	Log10(C) = -4.0	Log10(C) = -4.0	Log10(C) = -5.0
1.2.1 a	<i>Voxel-Direct_VOI-D-gm</i>	Log10(C) = -1.5	Log10(C) = -1.0	Log10(C) = 0.0
1.2.1 b	<i>Voxel-Direct_VOI-D-all</i>	Log10(C) = -2.0	Log10(C) = -1.5	Log10(C) = -5.0
1.2.2 a	<i>Voxel-Direct_VOI-S-gm</i>	Log10(C) = -1.5	Log10(C) = -1.0	Log10(C) = -5.0
1.2.2 b	<i>Voxel-Direct_VOI-S-all</i>	Log10(C) = -1.0	Log10(C) = -2.0	Log10(C) = -5.0
1.3.1 a	<i>Voxel-STAND-D-gm</i>	Log10(C) = 0.5; t = 0.88	Log10(C) = -1.5; t = 0.80	Log10(C) = -2.0; t = 0.72
1.3.1 b	<i>Voxel-STAND-D-all</i>	Log10(C) = -2.0; t = 0.90	Log10(C) = -2.5; t = 0.64	Log10(C) = -5.0; t = 0.10
1.3.2 a	<i>Voxel-STAND-S-gm</i>	Log10(C) = -1.5; t = 0.84	Log10(C) = -2.0; t = 0.44	Log10(C) = -1.0; t = 0.90
1.3.2 b	<i>Voxel-STAND-S-all</i>	Log10(C) = -0.5; t = 0.90	Log10(C) = -3.5; t = 0.22	Log10(C) = -3.5; t = 0.34
1.3.3 a	<i>Voxel-STAND-Sc-gm</i>	Log10(C) = -3.0; t = 0.54	Log10(C) = -1.5; t = 0.54	Log10(C) = -1.5; t = 0.86
1.3.3 b	<i>Voxel-STAND-Sc-all</i>	Log10(C) = -4.0; t = 0.06	Log10(C) = -3.5; t = 0.18	Log10(C) = -2.0; t = 0.64
1.4.1 a	<i>Voxel-Atlas-D-gm</i>	Log10(C) = 0.5	Log10(C) = 1.5	Log10(C) = -5.0
1.4.1 b	<i>Voxel-Atlas-D-all</i>	Log10(C) = 0.5	Log10(C) = 0.5	Log10(C) = -5.0
1.4.2 a	<i>Voxel-Atlas-S-gm</i>	Log10(C) = 1.0	Log10(C) = 1.0	Log10(C) = -5.0
1.4.2 b	<i>Voxel-Atlas-S-all</i>	Log10(C) = 0.5	Log10(C) = 0.5	Log10(C) = -5.0
1.5.1 a	<i>Voxel-COMPARE-D-gm</i>	Log10(C) = 2.5 $\sigma = 700$; n = 20	Log10(C) = 1.5 $\sigma = 500$; n = 5	Log10(C) = 0.0 $\sigma = 400$; n = 4
1.5.1 b	<i>Voxel-COMPARE-D-all</i>	Log10(C) = 2.0 $\sigma = 300$; n = 16	Log10(C) = 2.0 $\sigma = 300$; n = 20	Log10(C) = 2.5 $\sigma = 1000$; n = 89
1.5.2 a	<i>Voxel-COMPARE-S-gm</i>	Log10(C) = 1.5 $\sigma = 500$; n = 6	Log10(C) = 1.0 $\sigma = 200$; n = 117	Log10(C) = 1.5 $\sigma = 100$; n = 128
1.5.2 b	<i>Voxel-COMPARE-S-all</i>	Log10(C) = 2.0 $\sigma = 600$; n = 98	Log10(C) = 2.0 $\sigma = 1000$; n = 5	Log10(C) = 1.0 $\sigma = 100$; n = 104
2.1	<i>Thickness-Direct</i>	Log10(C) = -5.0	Log10(C) = -5.0	Log10C = -5.0
2.2	<i>Thickness-Atlas</i>	Log10(C) = 0.0	Log10(C) = -0.5	Log10(C) = -0.5
2.3	<i>Thickness-ROI</i>	-	-	-
3.1.1	<i>Hippo-Volume-F</i>	-	-	-
3.1.2	<i>Hippo-Volume-S</i>	-	-	-
3.2	<i>Hippo-Shape</i>	Log10(C) = -1.5	Log10(C) = -1.0	Log10C = -5.0

316

317

318

319

320 **Table S2.** List of subjects from the ADNI database that were included in the study. This table lists
 321 the IDs of the subjects and the IDs of their T1 images (UID) as well as the series identifier (SID) for
 322 the CN training set.

CN: TRAINING SET				CN: TRAINING SET			
CENTER ID	SUBJECT ID	SID	UID	CENTER ID	SUBJECT ID	SID	UID
2	295	13408	45109	36	1023	24338	65105
2	1261	26574	62378	41	1002	23705	65221
2	1280	26453	60057	57	643	15782	34726
3	907	19728	52782	57 *	779 *	18109 *	80630 *
3	981	20753	52777	57	934	19971	34735
5	602	15966	32674	67	19	9539	45229
6	484	18837	65567	73	312	15079	39888
6	498	15857	67058	73	386	13746	49681
6	681	18451	92306	82	363	13017	39731
7	68	10356	35791	82	640	16562	49480
7	1222	25402	60004	94	489	14121	47292
11	2	9107	35470	94	711	16553	39139
11	5	9136	32247	98	171	11460	65753
11	16	9253	32307	98	896	19439	56032
11	21	9581	32342	99	533	14938	38786
11	22	9617	32394	109	1014	24818	63480
11	23	8868	32410	114	173	11594	96322
13	502	17232	51139	114	601	15201	39851
13	575	17859	51161	126	405	14635	38828
13	1035	21984	51166	126	506	15652	38855
14	519	14488	39648	126	605	15665	38705
14	520	14473	39661	126	680	16099	38927
14	558	15789	39684	127 *	260 *	12419 *	34385 *
20	883	19459	60674	127	622	15473	34453
20	899	19329	60628	128	863	19272	98876
20	1288	26890	60665	130	232	11928	39110
22	14	9271	59376	130	969	20385	39277
22	130	11089	59525	130	1200	25090	63753
23	61	10312	31103	131	123	10960	63785
23	963	19740	89418	131	319	12322	47985
24	985	21112	62716	131	441	13681	48030
24	1063	21525	63393	133	433	13952	39947
27 *	118 *	11796 *	34115 *	133	488	14150	107935
27	403	13717	34182	136	86	13191	66414
29	824	18139	96285	136	196	13253	40261
29	843	18909	66998	941	1194	25323	63848
33	920	19288	42482	941	1195	26180	63866
33	1098	22792	42833	941	1197	25332	66463
35	48	10257	45184	941	1202	25680	63875
36	672	17131	36939	941	1203	25671	63880
36	813	18252	36980				

323 *: images for which the cortical thickness extraction pipeline failed.

324 †: images for which the SPHARM computation pipeline failed.

325

326

327 **Table S3.** IDs of the subjects and the IDs of their T1 images (UID and SID) for the CN testing set.

CN: TESTING SET				CN: TESTING SET			
CENTER ID	SUBJECT ID	SID	UID	CENTER ID	SUBJECT ID	SID	UID
2	413	13893	45118	52	1250	25829	62241
2	559	14875	40675	52	1251	26231	62283
2	685	16309	40684	62	578	15035	50460
3	931	20051	53391	62	768	17527	50507
3	1021	21771	73507	62	1099	22713	50558
5	553	15527	32646	67	56	8723	35894
5	610	15727	32669	67	59	10517	35903
6	731	18321	90849	67	177	12187	34807
7	70	10950	36629	67	257	14730	34825
7	1206	25173	59982	73	89	11161	49676
11	8	9195	32265	73	311	15069	39869
13	1276	27641	62681	82	304	12557	95654
14	548	14921	39666	82	761	18119	39792
16	359	13000	96222	82	1256	26812	63157
16	538	17545	40773	94	526	14559	63469
20	97	10858	64047	94	692	17207	47296
22	66	10271	59447	94	1267	28217	80719
22	96	11006	59457	98	172	11812	65758
23	31	9785	69613	99	40	8920	34608
23	58	10335	30551	99	90	10835	35842
23	81	10813	31126	99	352	12992	34538
23	926	19390	31548	99	534	14009	34579
23	1190	24847	46418	109	876	19132	82595
23	1306	26604	46436	109	967	21137	55045
27	74	10605	34317	109	1013	22585	66150
27	120	11535	34333	114	166	11584	39811
29	845	18829	64868	114	416	13556	39837
29	866	18917	65612	127	259	12137	34363
33	516	14818	42309	127	684	16759	34459
33	734	16942	42435	130	886	19561	39173
33	741	17006	42451	131	436	14710	48021
33	923	19544	42510	131	1301	26899	63794
33	1016	21817	42773	133	493	14156	107944
33	1086	23547	42786	133	525	14991	39981
35	156	11391	39534	136	184	11974	40180
35	555	15332	39602	136	186	11774	40202
36	576	15156	36904	141	717	18413	98889
41	125	10883	35672	141	767	18337	47308
41	898	20306	34699	141	810	20274	47315
51	1123	24099	58044	141	1094	23294	47733
52	951	20352	64171				

328 *: images for which the cortical thickness extraction pipeline failed.

329 †: images for which the SPHARM computation pipeline failed.

330

331

332 **Table S4.** IDs of the subjects and the IDs of their T1 images (UID and SID) for the AD training
 333 set.

AD: TRAINING SET				AD: TRAINING SET			
CENTER ID	SUBJECT ID	SID	UID	CENTER ID	SUBJECT ID	SID	UID
2	816	18402	40732	67	812	19629	38727
2	938	19852	40981	67	828	18532	65717
2	955	20004	40755	67	1185	24635	63105
3	1059	22300	52817	67	1253	27558	55034
3	1257	27340	52791	73	565	15762	39920
5	814	18390	74592	82	1377	28495	63171
7	1248	25568	59951	94	1102	22905	63187
7	1304	26475	59911	94	1164	23871	67224
11	3	9127	32238	94	1397	31011	95663
11	10	8800	32275	94	1402	32102	66079
13	592	18419	79145	98	884	24183	56027
13	699	18366	62651	99	470	14222	34571
13	1161	24399	51486	99	492	14944	38771
14	356	12857	39630	99	1144	24218	102041
14	1095	23323	45741	109	777	18676	82577
22	129	11485	59485	114	374	13031	39818
23 ^(†)	78 ^(†)	10619 ^(†)	52000 ^(†)	114	979	21933	39860
23	84	10764	31207	126	606	17191	38910
23	139	11079	31304	126	1221	25457	48977
27	1081	25357	47169	127	844	19874	34472
29	836	18151	65014	128	1409	33787	69401
29	1184	25463	67211	128 *	1430 *	39199 *	79858 *
33	889	19296	51630	130	1290	26038	63767
35	341	12952	45217	130	1337	27584	63776
36	577	14974	36915	131	457	13976	92407
36	760	18264	38653	131	497	15315	48039
57	474	13990	34721	131	691	17266	48048
57	1371	28667	62999	133	1170	24674	89958
57	1373	28698	63009	136	299	13839	40313
57 ^(†)	1379 ^(†)	28761 ^(†)	63015 ^(†)	136	300	14136	40329
62	535	14699	50427	136	426	14581	40357
62	690	16924	50469	141	1024	22699	47749
62	730	17062	50488	141	1137	24301	48582
67	29	9904	38718	141	1152	24487	48591
67	76	10468	35912				

334 *: images for which the cortical thickness extraction pipeline failed.

335 †: images for which the SPHARM computation pipeline failed.

336

337

338 **Table S5.** IDs of the subjects and the IDs of their T1 images (UID and SID) for the AD testing set.

AD: TESTING SET				AD: TESTING SET			
CENTER ID	SUBJECT ID	SID	UID	CENTER ID	SUBJECT ID	SID	UID
2	1018	23128	40818	33	1281	26136	54781
5	221	11958	72129	33	1283	26144	54786
5	929	19669	74610	33	1285	26128	51589
5	1341	27673	60418	33	1308	26600	54753
6	547	16033	67316	36	759	18094	36970
6	653	16073	67325	36	1001	22691	38662
7	316	12583	36574	41	1368	27512	65249
7	1339	27414	56320	41	1391	29116	62934
11	53	10064	35487	41	1435	39186	79637
11	183	12000	32004	51	1296	26431	58024
13	996	22240	51184	53	1044	21256	64204
13	1205	25024	51543	62	793	18189	50525
14	328	12402	39621	67	110	11177	35934
16	991	21737	40795	82	1079	22650	49491
16	1263	27303	64623	94	1027	21207	49528
20	213	12386	60601	94	1090	23375	63177
22	7	9024	59367	98	149	11021	89430
22	219	12375	59535	99	372	13672	34550
22	543	14849	59544	109	1157	24711	66159
23	83	10568	31144	109	1192	25056	63504
23	93	10736	31254	114	228	11697	49736
23	916	19228	31534	126	784	19752	39013
23	1262	26314	62434	126 ^(†)	891 ^(†)	19386 ^(†)	39055 ^(†)
23	1289	26374	89939	127	431	14595	34444
24	1171	24659	63407	127	754	18515	80761
24	1307	27061	63416	127	1382	28261	66311
27	404	13866	34205	130	956	20667	39187
27	850	18554	48997	130	1201	25082	63758
27	1254	25764	47229	133	1055	22386	40029
27	1385	28133	47575	136	194	13178	40240
29	999	23248	64899	141	696	18373	82739
29	1056	22977	60742	141	790	18766	91254
33	724	17337	42401	141	852	19395	47745
33	733	16932	42426	141	853	18348	112293

339 *: images for which the cortical thickness extraction pipeline failed.

340 †: images for which the SPHARM computation pipeline failed.

341

342

343

344 **Table S6.** IDs of the subjects and the IDs of their T1 images (UID and SID) for the MCIc training

345 set.

MCIc: TRAINING SET				MCIc: TRAINING SET			
CENTER ID	SUBJECT ID	SID	UID	CENTER ID	SUBJECT ID	SID	UID
2	954	19979	40745	33	906	19314	42469
2	1070	23120	40832	33	922	19341	42494
5	222	11754	54689	35	204	11661	39543
7	41	9994	35735	35	997	23184	62909
7	128	10936	36641	51	1331	29664	64153
7	344	12631	36580	52	952	20364	89953
11	856	19031	89409	52	1054	22955	62235
13	240	12308	51152	53	507	14483	80200
13	860	19237	51534	62	1299	26794	50585
22	750	17695	59553	67	243	12030	34820
22	1394	34317	68083	67	336	14023	34858
23	42	8852	31085	94	1015	21457	40764
23	388	13076	31438	94	1398	31771	63228
23	604	15182	31456	127	394	14603	34399
23	855	18561	31510	133	638	16608	67532
23	887	19087	31527	136	195	12523	40453
23	1247	25741	48858	141	982	22644	47704
27	461	15192	34232	941	1311	27408	97328
33	723	16845	42385	941	1363	28008	63898
33	725	17092	42410				

346 *: images for which the cortical thickness extraction pipeline failed.

347 †: images for which the SPHARM computation pipeline failed.

348

349

350 **Table S7.** IDs of the subjects and the IDs of their T1 images (UID and SID) for the MCIc testing

351 set.

MCIc: TESTING SET				MCIc: TESTING SET			
CENTER ID	SUBJECT ID	SID	UID	CENTER ID	SUBJECT ID	SID	UID
2	729	16874	40709	57	941	19985	34748
5	572	15709	32659	57	1217	25854	62985
6	1130	23457	55973	67	45	10185	35889
7	249	11911	36531	67	77	11136	68121
11	241	12088	32021	94	434	13570	39124
11	861	19476	35514	98	269	11964	65258
11	1282	26225	62637	99	54	10329	35826
13	325	13524	54666	99	111	10933	35850
14	658	17481	39702	126	1077	23496	48881
23	30	9441	31632	127	1427	37933	91127
23	625	15820	31496	128	947	19859	69074
27	179	11781	34137	130	423	15030	39134
27	256	12357	34151	133	727	18620	66358
27	1213	25492	47224	133 ^(†)	913 ^(†)	24646 ^(†)	63820 ^(†)
27	1387	28123	67202	136	695	19019	70925
33	567	14572	42371	141	915	20504	48573
41	549	15488	39507	141	1244	26845	92647
41	1412	34022	72220	941	1295	26290	63889
41	1423	36902	72233				

352 *: images for which the cortical thickness extraction pipeline failed.

353 †: images for which the SPHARM computation pipeline failed.

354

355

356

357 **Table S8.** IDs of the subjects and the IDs of their T1 images (UID and SID) for the MCInc
 358 training set.

MCInc: TRAINING SET				MCInc: TRAINING SET			
CENTER ID	SUBJECT ID	SID	UID	CENTER ID	SUBJECT ID	SID	UID
3	1122	23542	52800	41	314	12492	34681
5	324	12599	32893	41	679	17077	40047
7	414	14826	36600	41	1010	23880	65231
11	326	12342	89391	41	1260	25806	65240
11	362	12678	89405	51	1072	22884	58012
11	1080	23159	35592	52	671	16062	64162
16	702	17341	40782	52	989	22476	64180
16	1028	22058	40800	52	1168	23688	65668
16	1138	24779	86046	53	621	15442	64190
22	1097	23337	59611	53	919	20422	65690
22	1351	28484	59616	57	464	14736	34708
23	126	11525	31272	57	1007	21339	47766
23	217	11731	31347	73	909	19716	75486
23	1046	22199	46397	94	1330	27038	63207
27	408	14231	37549	98	160	11224	65740
27	644	15630	34241	99	60	10478	35834
27	835	18760	35667	99	291	12065	34525
29	878	18986	64877	109	950	21165	97201
29	1038	22851	60733	114	378	12760	95689
29	1073	22828	65023	126	865	22295	39034
29	1215	25348	60747	126	1187	25143	48960
29	1218	25478	67390	127	112	11194	98858
33	511	15101	42241	127	1140	24278	63638
33	513	14673	42259	130	285	12462	39119
33	514	14663	42277	130	289	12111	39114
33	1279	26694	54758	130	449	22250	80931
33	1284	26686	54800	130	783	18023	39154
35	33	10396	45167	131	384	12790	48003
35	292	12408	39570	133	771	18575	92287
36	656	16286	36925	133	912	19884	40001
36	673	17157	36950	136	1227	26399	63839
36	748	17705	36960	141	697	18466	91236
36	869	21421	36994	141	851	19364	47871
36	976	23852	65092				

359 *: images for which the cortical thickness extraction pipeline failed.
 360 †: images for which the SPHARM computation pipeline failed.

361
 362

363

364 **Table S9.** IDs of the subjects and the IDs of their T1 images (UID and SID) for the MCInc testing
 365 set.

366

MCInc: TESTING SET				MCInc: TESTING SET			
CENTER ID	SUBJECT ID	SID	UID	CENTER ID	SUBJECT ID	SID	UID
2	782	17835	40718	62	1182	25166	50567
2	1155	24144	40846	73	746	23286	63123
3	908	32516	62590	82	1119	23733	63148
3	1074	23536	53396	94	531	14554	49667
5	448	14032	32877	94	921	19582	49511
5	546	15566	32683	94	1293	27865	64375
5	1224	25412	60407	94	1314	27488	63197
7	101	10679	36727	98	667	15980	89496
7	293	12193	36550	99	51	10325	35820
7	698	16403	36614	99	1034	21759	47954
13	1186	25689	62657	109	1114	24702	63490
14	169	11565	40859	109	1183	24718	66168
14	557	15094	39675	114	410	13289	39825
14	563	16079	39693	114	458	14111	39846
16	769	17720	48868	114	1103	23239	49758
22	4	9234	64632	114	1106	22859	49911
22	544	14679	64673	114	1118	23803	49769
22	961	20711	59602	126	708	16897	38945
23	331	12428	31374	126	709	17326	38965
23	376	12652	31385	127	393	13200	34394
27	116	11442	34326	127	925	21560	67267
27	307	13072	34160	127	1032	22259	63633
27	485	14208	37558	128	1043	22562	69092
27	1045	22174	47211	130	102	10746	39461
33	1116	22799	42845	130	505	17292	39144
33	1309	26195	51606	133	629	15915	40955
36	945	20971	37000	133	792	18306	66374
36	1135	24501	65110	133	1031	21552	40020
36	1240	26423	65119	136	107	11707	40446
41	282	13496	39487	136	429	15534	40388
41	598	15604	40038	136	579	15952	40411
51	1131	24089	62944	136	874	22234	40420
53	389	13550	65677	141	1052	22923	47723
57	839	19188	38671				

367 *: images for which the cortical thickness extraction pipeline failed.

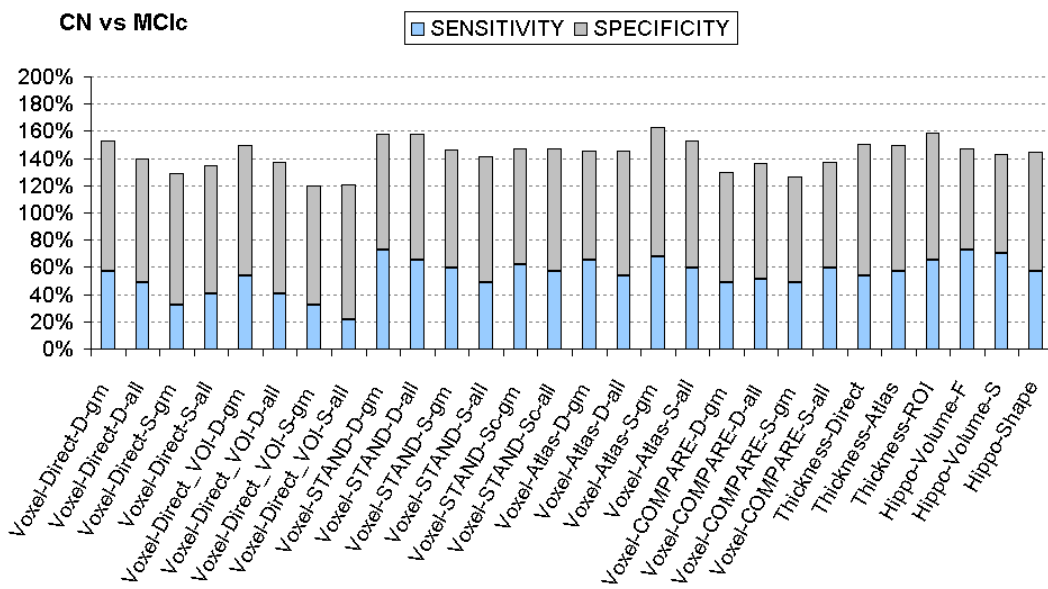
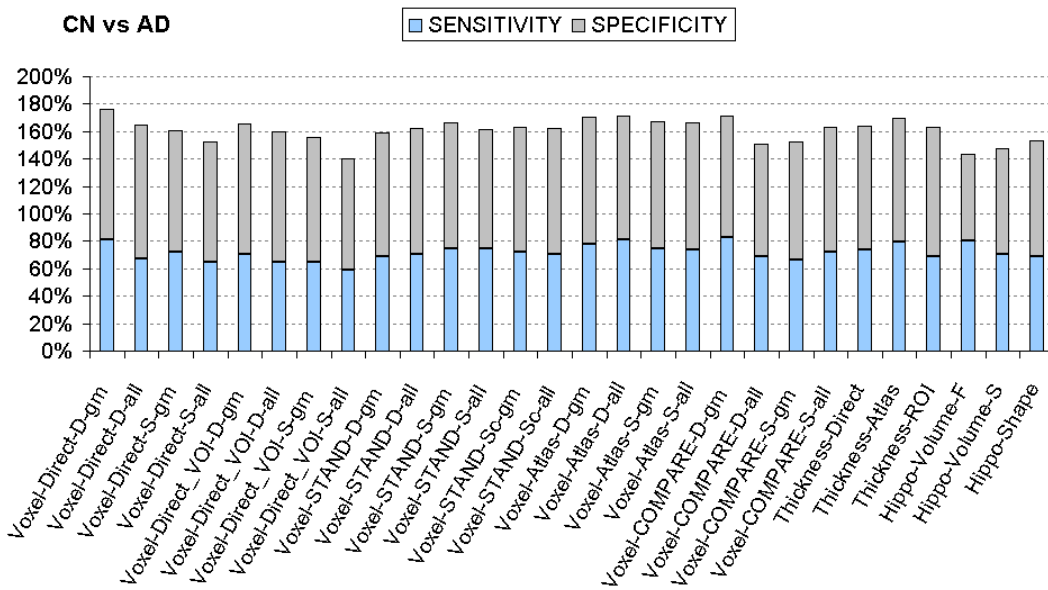
368 †: images for which the SPHARM computation pipeline failed.

369

370

1 **List of figures**

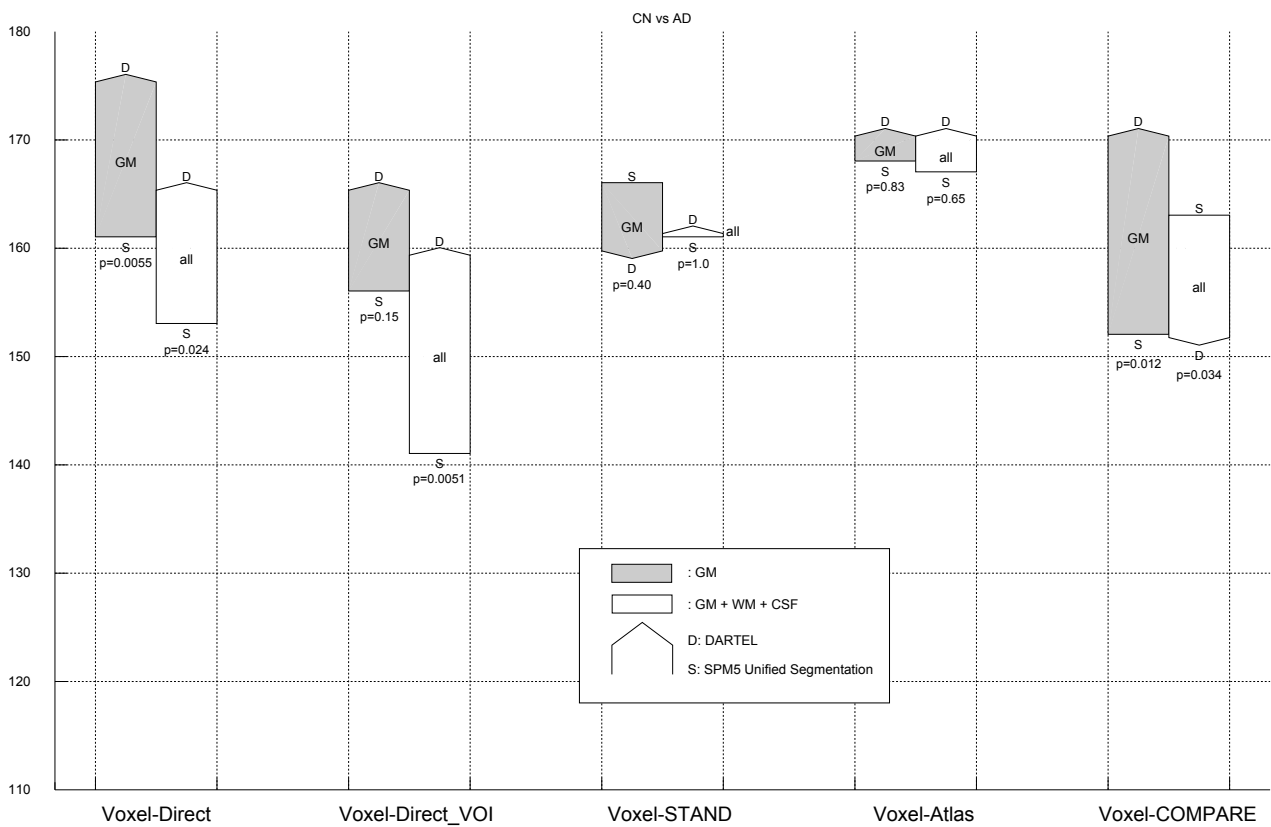
2 **Figure 1.** Classification results for the different methods.



3

4

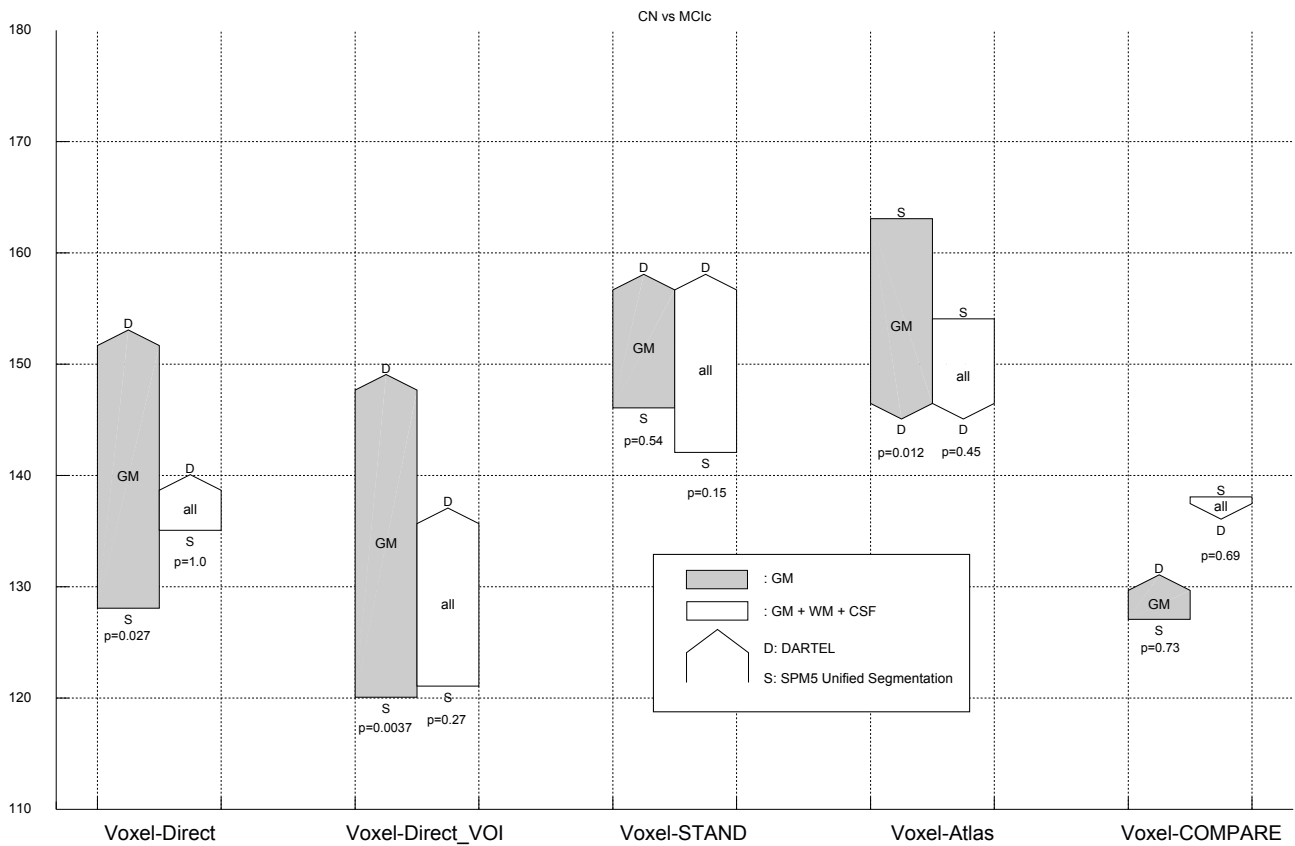
5 **Figure 2.** Impact of the preprocessing on the accuracy for CN vs AD. The sum of the sensitivity
6 and specificity is considered. The front tip of an arrow indicates the results obtained with DARTEL
7 whereas the back tip indicates the results obtained with SPM5 unified segmentation. The color of
8 the arrow indicates the features used. Grey arrows correspond to the use of GM probability maps
9 only whereas white arrows correspond to the use of GM, WM and CSF probability maps. The p-
10 values obtained with the McNemar's chi square test assessed the difference between the results
11 obtained with DARTEL and SPM5.



12

13

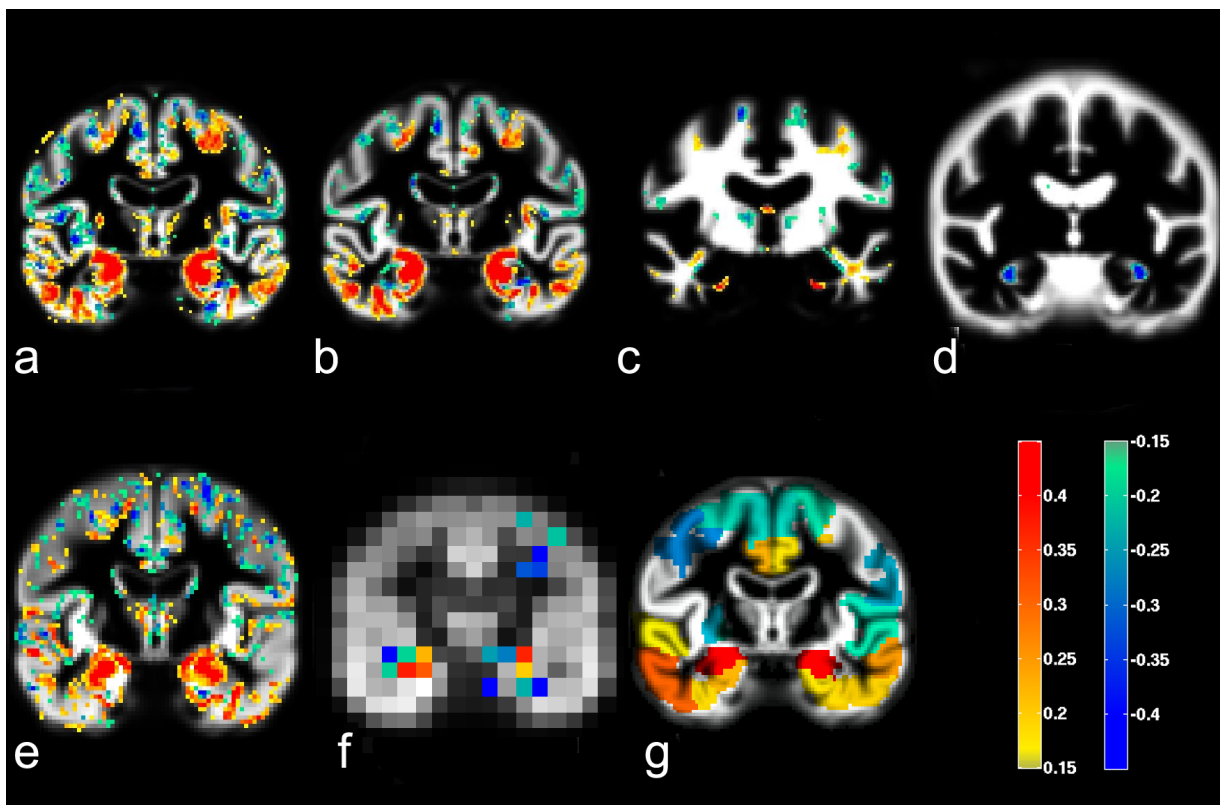
14 **Figure 3.** Impact of the preprocessing on the accuracy for CN vs MCIc. The sum of the sensitivity
 15 and specificity is considered. The front tip of an arrow indicates the results obtained with DARTEL
 16 whereas the back tip indicates the results obtained with SPM5 unified segmentation. The color of
 17 the arrow indicates the features used. Grey arrows correspond to the used of GM probability maps
 18 only whereas white arrows correspond to the use of GM, WM and CSF probability maps. The p-
 19 values obtained with the McNemar's chi square test assessed the difference between the results
 20 obtained with DARTEL and SPM5.



21

22

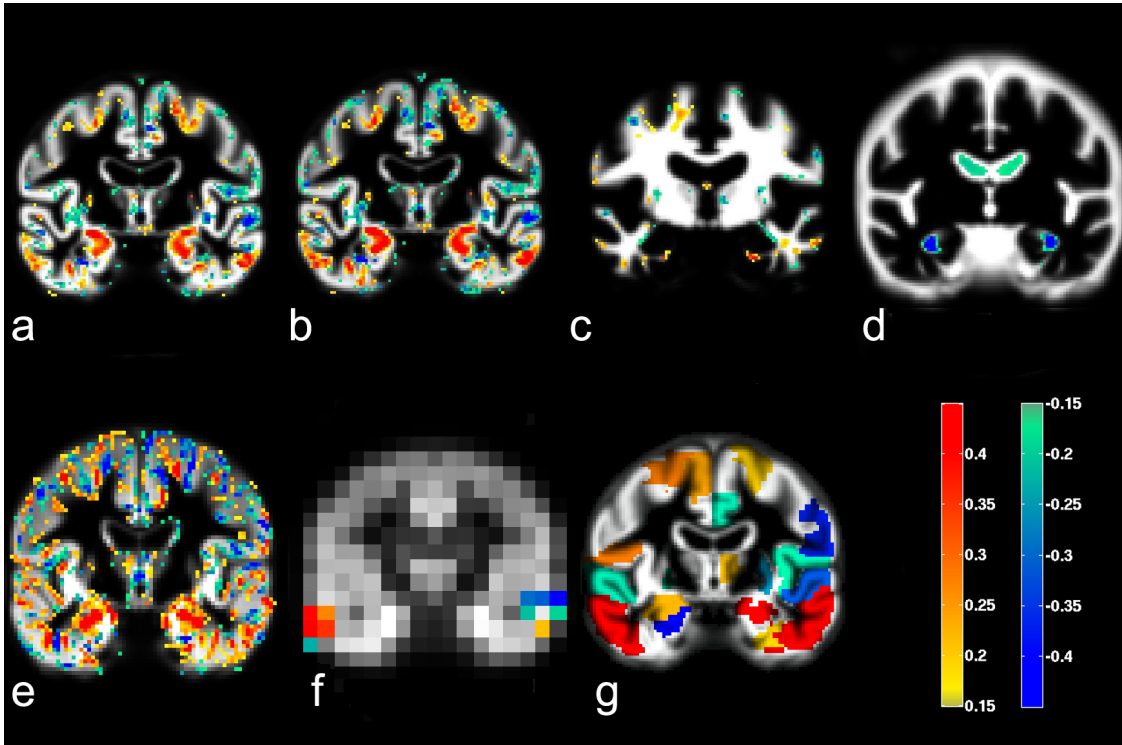
23 **Figure 4.** Optimal margin hyperplane in the CN vs AD experiments for *Voxel-Direct-D-gm* (a),
24 *Voxel-Direct-D-all* (b-d), *Voxel-Direct-S-gm* (e), *Voxel-STAND-D-gm* (f) and *Voxel-Atlas-D-gm* (g).
25 The figure displays the normalized vector orthogonal to the hyperplane superimposed on the tissue
26 average probability maps. The coronal slices are equivalent to $y = 9mm$ in the MNI-space. For
27 visualization purposes, only coefficients w_i greater than 0.15 in absolute value are displayed. For
28 regions in warm colors, tissue atrophy increases the likelihood of classification into AD or MCIc.
29 For regions in cool colors, it is the opposite.



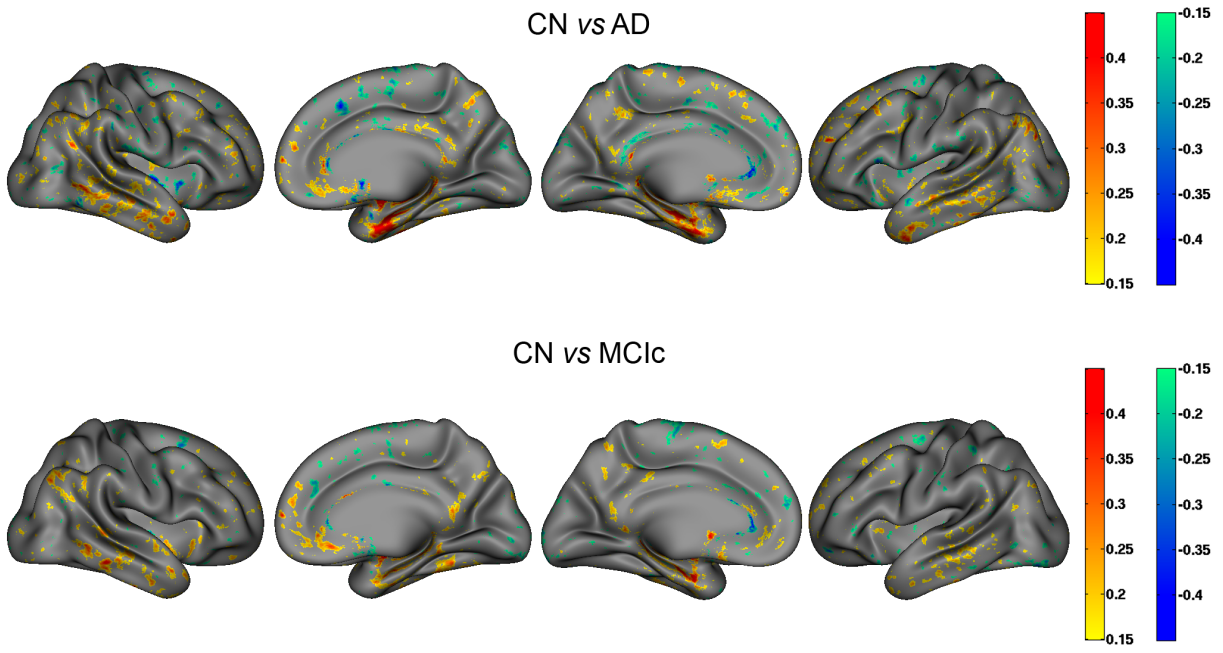
30

31

32 **Figure 5.** Optimal margin hyperplane in the CN vs MCIc experiments for *Voxel-Direct-D-gm* (a),
33 *Voxel-Direct-D-all* (b-d), *Voxel-Direct-S-gm* (e), *Voxel-STAND-D-gm* (f) and *Voxel-Atlas-D-gm* (g)
34 (please refer to Figure 4 for a complete description of the figure).

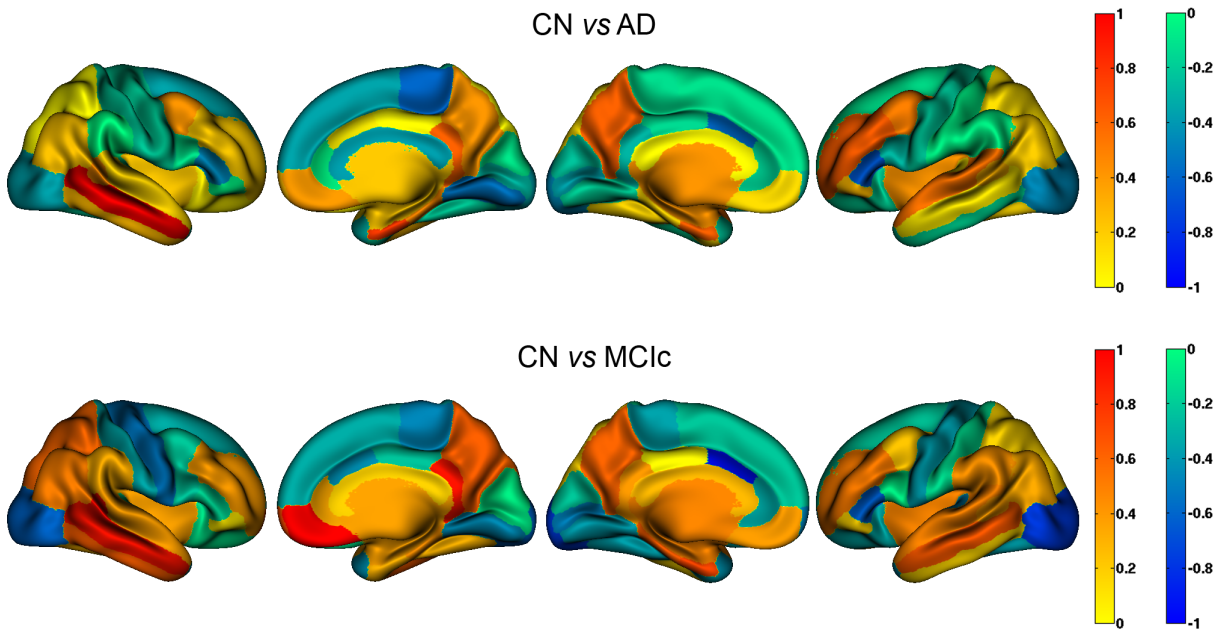


37 **Figure 6.** Optimal margin hyperplane for *Thickness-Atlas*. Upper rows: CN vs AD experiment.
38 Lower rows: CN vs MCIc experiment.



39
40

41 **Figure 7.** Optimal margin hyperplane for *Thickness-Direct*. Upper rows: CN vs AD experiment.
42 Lower rows: CN vs MCIc experiment.



43

44

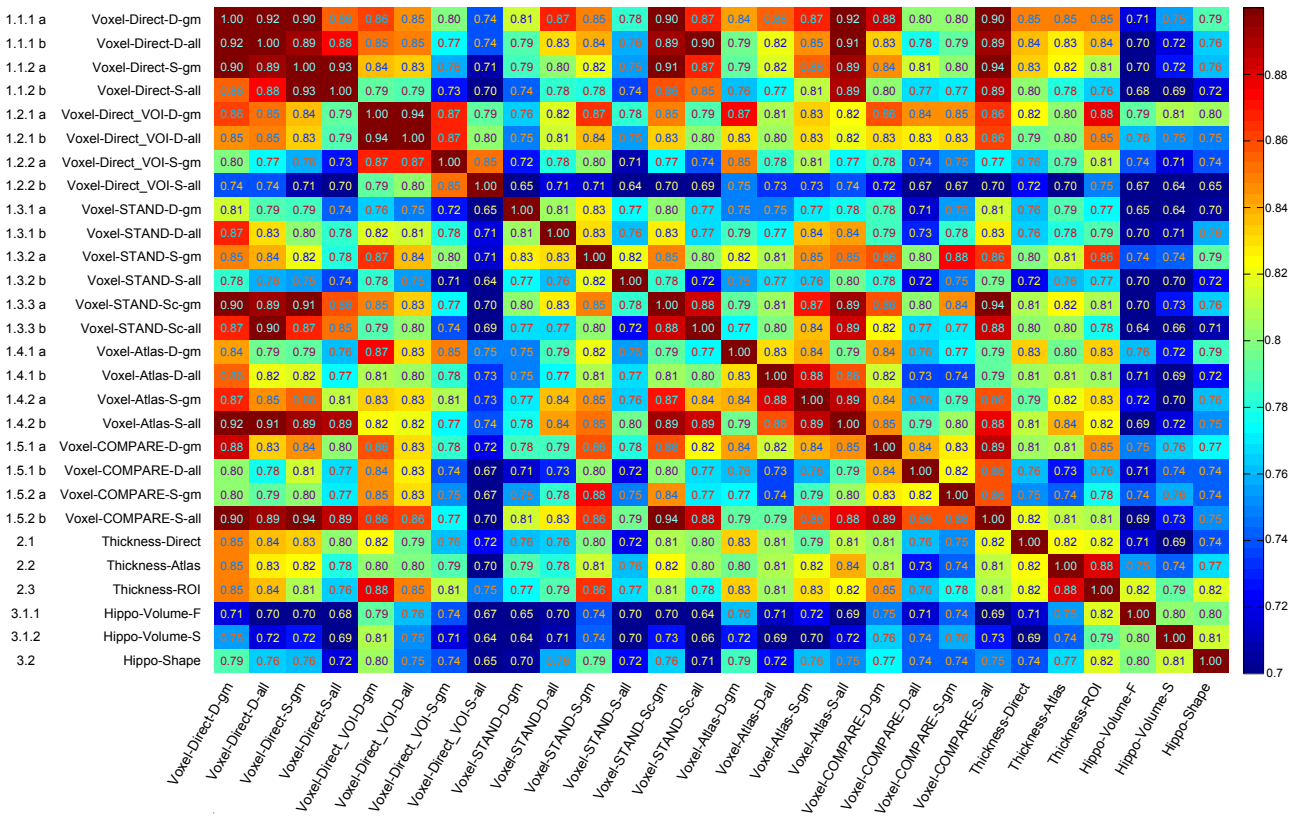
45 **List of figures for Supplementary Materials**

46 **Figure S1.** Jaccard index of similarity between the different classification methods for CN vs AD.

47 The Jaccard index of two methods is the number of subjects correctly classified by both methods

48 divided by the number of subjects correctly classified by at least one of the two methods.

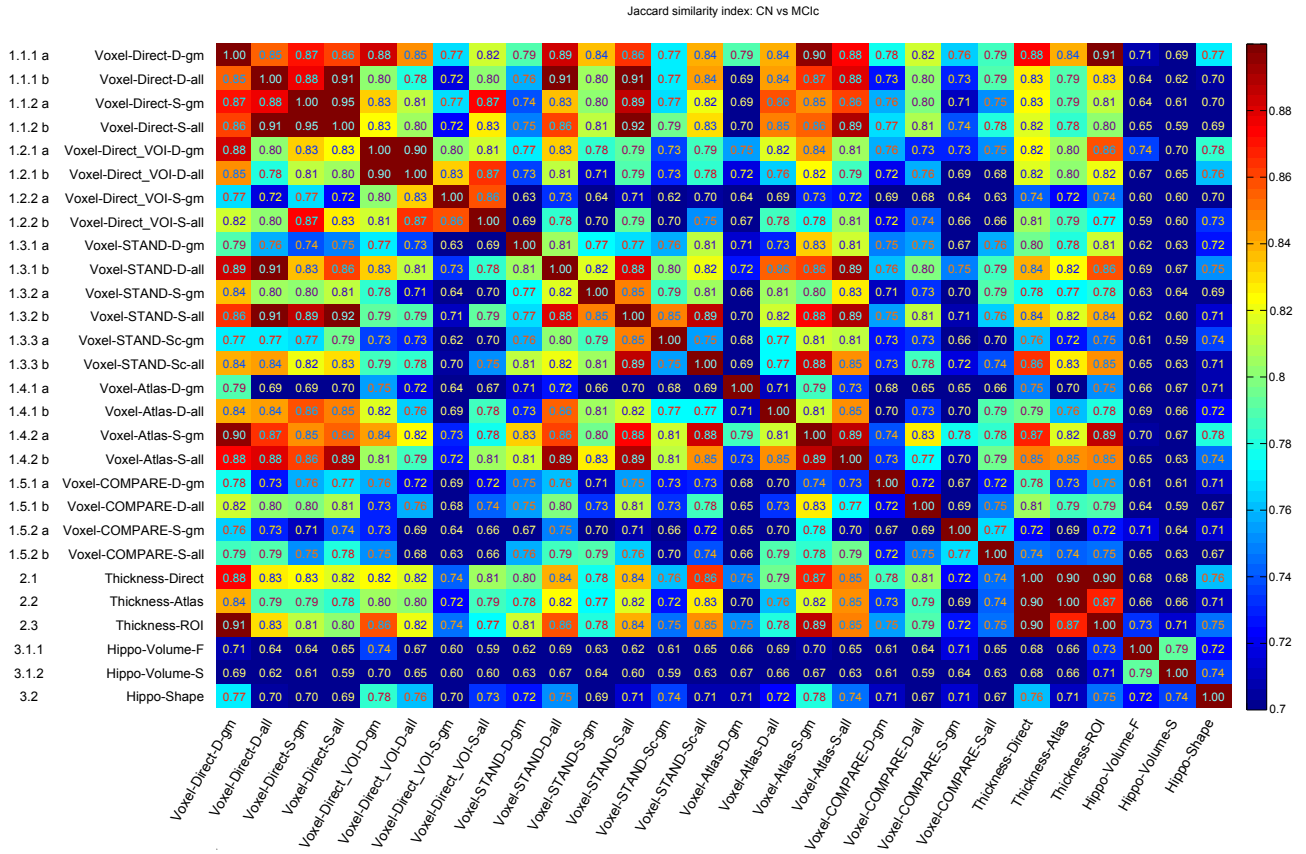
Jaccard similarity index: CN vs AD



49

50

51 **Figure S2.** Jaccard index of similarity between the different classification methods for
 52 CN vs MCIc. The Jaccard index of two methods is the number of subjects correctly classified by
 53 both methods divided by the number of subjects correctly classified by at least one of the two
 54 methods.



55