Voxel-based dementia classification of AD, MCI and controls for the CADDementia data set

Esther E. Bron[1], Marion Smits[2], John C. van Swieten[3], Wiro J. Niessen[1,4], and Stefan Klein[1]

[1] Biomedical Imaging Group Rotterdam, Departments of Medical Informatics and Radiology, Erasmus MC, Rotterdam, the Netherlands, *e.bron@erasmusmc.nl*
[2] Department of Radiology, Erasmus MC, Rotterdam, the Netherlands
[3] Department of Neurology, Erasmus MC, Rotterdam, the Netherlands
[4] Imaging Physics, Applied Sciences, Delft University of Technology, the Netherlands

Abstract. In the CADDementia challenge, algorithms for computer-aided diagnosis of dementia based on structural MRI were objectively compared using a previously unseen multicenter data set. In this work, we present our submission to this challenge. Our algorithm computed features based on voxel-based morphometry. We used a linear support vector machine classifier for classification. For training, we used 509 scans of Alzheimer's disease patients, patients with mild cognitive impairment and healthy controls from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database. The tuning and validation of the algorithms was performed on the 30 subjects from the CADDementia training set. We evaluated eight variations to our algorithm that included confounder correction and prior optimization. On the training set, the best performing algorithm yielded an accuracy of 66.7% and and area under the receiver-operating-characteristic curve of 72.0% percent. Finally, five of those eight algorithms were selected to be submitted to the challenge.

1 Introduction

Computer-aided diagnosis of neurodegenerative diseases is an emerging research field in which machine learning approaches are used to distinguish for example Alzheimer's disease (AD) patients from normal (CN) controls [1]. Although in the literature very promising results of algorithms for computer-aided diagnosis of AD and mild cognitive impairment (MCI) have been reported, they are difficult to compare as different data sets and methodology were used for evaluation. In addition, it is unclear how such algorithms would generalize to new data, and thus, how they would perform in clinical practice when the algorithm cannot be adapted completely to a specific data set. For these reasons, the challenge on Computer-Aided Diagnosis of Dementia based on Structural MRI (CADDementia)[5] was initiated.

In this paper, we describe the algorithm that we submitted to the CADDementia challenge. The image processing and classification pipeline was based on

[5] http://caddementia.grand-challenge.org

that of Bron et al. [2]. We trained our algorithms on data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) and on the 30 training subjects provided by CADDementia. The method was tuned on the 30 training sets by training the algorithms on the ADNI data. Five variations to the algorithm were applied to the CADDementia test set. Some additions to the previously published methods were made. First, CADDementia addresses multi-class classification of AD, MCI and CN, so in this work we combine the pairwise predictions of these classes to make a final prediction. Second, as the subjects in the CADDementia data set are younger than those from ADNI, we incorporated confounder correction. Third, priors were optimized on the training set.

2 Methods

2.1 Data

The data used in this work is made public by the CADDementia challenge. This multi-center data set was composed consisting of imaging data of 384 subjects from three medical centers: VU University Medical Center (VUMC), Amsterdam, the Netherlands; Erasmus MC (EMC), Rotterdam, the Netherlands; University of Porto / Hospital de São João (UP), Porto, Portugal. The data set contained structural T1-weighted MRI (T1w) scans of patients with the diagnosis of probable AD, patients with the diagnosis of MCI, and CN without a dementia syndrome. In addition to the MR scans, the data set included demographic information (age, gender) and information on which institute the data came from. Within the three centers, the data sets of the three classes had a similar age and gender distribution. Of these 384 subjects, 30 were used for training and tuning the methods. This training set consisted of 9 AD patients (3 male, 66.1 ± 5.2 yrs), 9 MCI patients (5 male, 68.0 ± 8.5 yrs), and 12 controls (9 male, 62.5 ± 6.1 yrs). For the other 354 subjects the labels were kept secret, these were used as test set (213 male, 65.1 ± 7.8 yrs).

For training, data from the ADNI was used in addition. The cohort is adopted from [3] and consists of AD patients, MCI patients that converted to AD within 18 months (MCIc), MCI patients that did not convert to AD within 18 months (MCInc), and CN. The participants were 137 AD patients (67 male, age: 76.0 ± 7.3 yrs, MMSE: 23.2 ± 2.0), 210 MCI (127 male, 74.5 ± 7.3 yrs, MMSE: 26.9 ± 1.8), and 162 CN (76 male, 76.3 ± 5.4 yrs, MMSE: 29.2 ± 1.0). T1w imaging was acquired at 1.5T with a voxel size of $\sim 1\text{mm}^3$ [4].

2.2 Image Processing

Probabilistic tissue segmentations were obtained for white matter (WM), gray matter (GM) and cerebrospinal fluid (CSF) on the T1w image using the unified tissue segmentation method [5] of SPM8 (Statistical Parametric Mapping, London, UK).

For construction of a template space, the coordinate transformations from the template space to the subject's space were derived from pairwise image registration [6] of a subset of 150 T1w images (81 CN, 69 AD [3]). We performed pairwise registrations with consecutively a rigid (including isotropic scaling), affine, and non-rigid B-spline transformation model. The non-rigid B-spline registration used a three-level multi-resolution framework with isotropic control-point spacing of 24, 12, and 6 mm at the three resolutions respectively. A template image was created by averaging the deformed individual images. To transform the other subjects' images to template space, coordinate transformations were derived from pairwise registrations to the subset. The registrations to the template space were visually inspected to check if they were correct.

For computation of intracranial volume, a brain mask was constructed using a multi-atlas segmentation approach. We performed brain extraction [7] on the T1w images associated with a set of 30 atlases [8, 9], checked the brain extractions visually, and adjusted extraction parameters if needed. The T1w images of the atlases were registered to the subjects T1w image using a rigid, affine, and non-rigid B-spline transformation model consecutively. Registration was performed by maximization of mutual information [10] within dilated brain masks [7]. For initialization, the dilated brain masks were rigidly registered. For non-rigid registration, the same multi-resolution settings were used as in the template-space construction. The subjects T1w images were corrected for inhomogeneities to improve registrations [11]. To make a brain mask for each subject, the extracted brains of the atlases were transformed to each subjects T1w image and the labels were fused using a majority voting algorithm [12].

2.3 Classification

For classification, linear SVM classifiers [13] were applied using the LibSVM software package [14]. The features were based on voxel-based morphometry, which means that we use GM probabilistic segmentations in the template space that are modulated by the Jacobian determinant of the deformation field to take account of compression and expansion [5]. To correct for head size, the features were divided by intracranial volume (ICV). The features were normalized to zero mean and unit variance. Pairwise classifications were performed in the following settings: AD/CN, AD/MCI, and MCI/CN. For making the three-class classification of CN, MCI and AD, the output probabilities of the pairwise classification were multiplied and normalized [15].

2.4 Optimization for CADDementia

The original method [2], as described above, divides all features by the ICV to correct for head size. We also tried classification without ICV correction.

Because the ADNI subjects were on average 10 years older than those of the CADDementia dataset, confounder correction was introduced. For this, we used the method proposed by Abdulkadir et al. [16]. Using kernel regression, the confounding effect of age, sex, and optionally also ICV were removed from

the kernel matrix. This kernel matrix is the linear kernel which is used as input for the SVM classifier and is the dot-product matrix of the feature values. This kernel matrix $\boldsymbol{K} \in \mathbb{R}^{N \times N}$ is corrected for the confounding effects using the following equation:

$$\tilde{\boldsymbol{K}} = \boldsymbol{R}\boldsymbol{K}\boldsymbol{R}^T, \quad \boldsymbol{R} = \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T, \tag{1}$$

in which $\boldsymbol{I}$ is the identity matrix and $\boldsymbol{X} \in \mathbb{R}^{N \times a}$is the matrix with the confounders for N subjects [16]. As confounders, we used: 1) age and sex ($a = 2$), and 2) age, sex and ICV ($a = 3$).

For the CADDementia test set, the exact numbers of subject in each class were not given. The data users only knew that priors were roughly equal for all classes. For the training set, the priors were not exactly equal for the classes: $p_{CN} = 1.2$, $p_{MCI} = 0.9$, and $p_{AD} = 0.9$. To take this into account as an additional option, the output probabilities of all classifications were multiplied by these priors.

For evaluation on the training set, only the ADNI data was used for training the classifier. For the algorithms that were submitted to the challenge for classification of the test set, training was performed on both the ADNI data and the 30 training subjects.

2.5 Evaluation

Evaluation was performed according to the protocol specified by the CADDementia challenge. The Python scripts that we used for evaluation can be downloaded from the challenge web site[6]. The performance of the algorithms was quantified by the classification accuracy and the area under the receiver-operating characteristic (ROC) curve (AUC). Confidence intervals on the accuracy and AUC were determined with bootstrapping on the test set (1000 resamples). The performance was evaluated on the 30 training subjects.

2.6 Detailed explanation and scripts

A detailed step-by-step guide of the used methodology is available on the CADDementia wiki[7]. The used scripts are also available[8].

3 Results

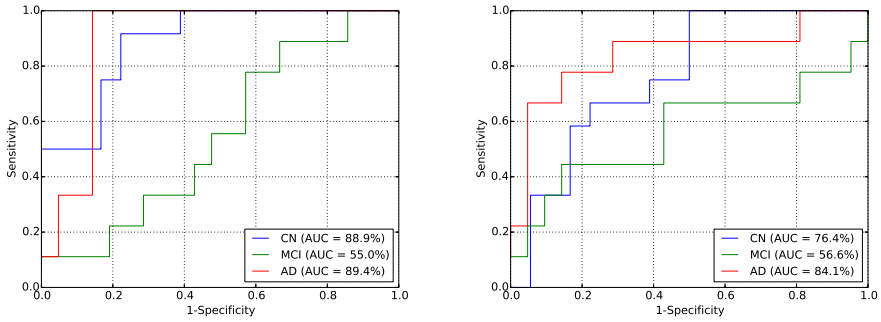
Table 1 reports the accuracy and AUC for the variations to the algorithm on the training set. The confusion matrix for all eight evaluated methods are shown in Table 2. The original method and the method without ICV correction yielded the same performance (accuracy=50%, AUC=77.8%). Correction for confounding effects improved the accuracy, but reduced AUC. The performance on the

[6] http://caddementia.grand-challenge.org
[7] http://wiki.caddementia.bigr.nl/index.php/Bron_et_al.
[8] https://bitbucket.org/bigr_erasmusmc/bron_iris_pipeline_adni

training set improved by taking the optimized priors ($p_{CN} = 1.2$, $p_{MCI} = 0.9$, and $p_{AD} = 0.9$) into account. The best accuracy was obtained by the algorithm with confounder correction for age, sex and ICV using the optimized priors. For this method (*age-sex-icv-op*), the accuracy is 66.7 % and the AUC 72.0%. The best AUC, 78.4 %, is obtained by the original method and by the method without ICV correction, both with the optimized priors. Figure 1 shows the ROC curves for the original method and for the method with the highest accuracy (*age-sex-icv-op*). The estimated computation time was about 4 hours per subject for image processing and a few minutes for classification.

Table 1. Performance on the train set, accuracy and area under the ROC-curve (AUC). OP denotes the optimized priors for the training set: $p_{CN} = 1.2$, $p_{MCI} = 0.9$, and $p_{AD} = 0.9$. CI = 95% confidence interval estimated with bootstrapping. The five methods in bold are submitted to the challenge to be applied to the test set.

Methods	Code	Accuracy (CI) [%]	AUC (CI) [%]
Original method [2]	***original***	50.0 (30.0 - 66.7)	77.8 (66.9 - 88.2)
Original method, OP	*original-op*	60.0 (36.7 - 73.3)	78.4 (67.4 - 88.8)
No ICV correction	*noicv*	50.0 (30.0 - 66.7)	77.8 (66.9 - 88.2)
No ICV correction, OP	*noicv-op*	60.0 (36.7 - 73.3)	78.4 (67.7 - 88.8)
Age and sex as confounders	***age-sex***	60.0 (40.0 - 73.3)	71.9 (55.3 - 86.1)
Age and sex as confounders, OP	***age-sex-op***	63.3 (43.3 - 76.7)	72.4 (57.3 - 85.0)
Age, sex, ICV as confounders	***age-sex-icv***	63.3 (43.3 - 76.7)	71.6 (52.8 - 86.1)
Age, sex, ICV as confounders, OP	***age-sex-icv-op***	66.7 (46.7 - 80.0)	72.0 (54.1 - 85.6)



(a) *Original* (b) *Age-sex-icv-op*

Fig. 1. The receiver-operating-characteristic (ROC) curves on the 30 training sets for two of the evaluated algorithms: (a) the original method, *Original*, and (b) the method with confounder correction for age, sex and ICV and optimized priors, *Age-sex-icv-op*. The area-under-the-ROC-curve (AUC) values are shown for every per-class ROC curve.

Table 2. The confusion matrices for the algorithms performance on the training set (n=30). The algorithms printed in bold were submitted for testing on the CADDementia test set.

original

		True class		
		CN	MCI	AD
Hypothesized class	CN	6	2	0
	MCI	6	4	4
	AD	0	3	5

original-op

		True class		
		CN	MCI	AD
Hypothesized class	CN	10	3	2
	MCI	2	3	2
	AD	0	3	5

noicv

		True class		
		CN	MCI	AD
Hypothesized class	CN	6	2	0
	MCI	6	4	4
	AD	0	3	5

noicv-op

		True class		
		CN	MCI	AD
Hypothesized class	CN	10	3	2
	MCI	2	3	2
	AD	0	3	5

age-sex

		True class		
		CN	MCI	AD
Hypothesized class	CN	8	3	1
	MCI	1	2	0
	AD	3	4	8

age-sex-op

		True class		
		CN	MCI	AD
Hypothesized class	CN	11	6	3
	MCI	0	2	0
	AD	1	1	6

age-sex-icv

		True class		
		CN	MCI	AD
Hypothesized class	CN	8	4	0
	MCI	2	3	1
	AD	2	2	8

age-sex-icv-op

		True class		
		CN	MCI	AD
Hypothesized class	CN	12	7	3
	MCI	0	2	0
	AD	0	0	6

4 Conclusion and Discussion

The algorithms in this work are based on [2] but adapted to multi-class classification of AD, MCI and CN. The original algorithm for voxel-based classification yielded an accuracy of 50.0% and an AUC of 77.8% on the training set. The methodological changes introduced in this work, confounder correction and prior optimization on the training set, improved accuracies on the training set, but reduced AUC.

Based on these performances, five algorithms were chosen to be evaluated on the CADDementia test, marked with bold font in Table 1. First, we chose the original method (*original*). In addition, four other methods with the highest accuracies on the training set were chosen. These were the algorithms that corrected for confounding effects: *age-sex* (accuracy=60.0%), *age-sex-op* (accuracy=63.3%), *age-sex-icv* (accuracy=63.3%), and *age-sex-icv-op* (accuracy=66.7%).

It should be noted that the authors of this work are the organizers of the CADDementia challenge. Therefore, the performance of our algorithms on the test set is listed on the web site, but will not be included in the official ranking of the CADDementia challenge.

5 Acknowledgments

This work was funded by an Erasmus MC grant on Advanced MR neuroimaging in presenile dementia.

References

1. Falahati, F., Westman, E., Simmons, A.: Multivariate Data Analysis and Machine Learning in Alzheimer's Disease with a Focus on Structural Magnetic Resonance Imaging. J Alzheimer Disease **41**(3) (2014) 685–708
2. Bron, E.E., Steketee, R.M.E., Houston, G.C., Oliver, R.a., Achterberg, H.C., Loog, M., van Swieten, J.C., Hammers, A., Niessen, W.J., Smits, M., Klein, S.: Diagnostic classification of arterial spin labeling and structural MRI in presenile early stage dementia. Hum Brain Mapp **35**(9) (2014) 4916–4931
3. Cuingnet, R., Gerardin, E., Tessieras, J., Auzias, G., Lehéricy, S., Habert, M.O.O., Chupin, M., Benali, H., Colliot, O.: Automatic classification of patients with Alzheimer's disease from structural MRI: A comparison of ten methods using the ADNI database. Neuroimage **56**(2) (2011) 766–781
4. Jack, C.R., Bernstein, M., Fox, N.C., Thompson, P., Alexander, G., Harvey, D., Borowski, B., Britson, P.J., L Whitwell, J., Ward, C., Dale, A.M., Felmlee, J.P., Gunter, J.L., Hill, D.L.G., Killiany, R., Schuff, N., Fox-Bosetti, S., Lin, C., Studholme, C., DeCarli, C.S., Krueger, G., Ward, H., Metzger, G.J., Scott, K.T., Mallozzi, R., Blezek, D., Levy, J., Debbins, J.P., Fleisher, A.S., Albert, M., Green, R., Bartzokis, G., Glover, G., Mugler, J., Weiner, M.W.: The Alzheimer's Disease Neuroimaging Initiative (ADNI): MRI methods. J Magn Reson Imaging **27**(4) (2008) 685–91
5. Ashburner, J., Friston, K.J.: Voxel-based morphometry - the methods. Neuroimage **11** (2000) 805–821
6. Seghers, D., D'Agostino, E., Maes, F., Vandermeulen, D., Suetens, P.: Construction of a brain template from MR images using state-of-the-art registration and segmentation techniques. In: Proc Intl Conf Med Image Comput Comp Ass Intervent, Springer (2004) 696–703
7. Smith, S.M.: Fast robust automated brain extraction. Hum Brain Map **17**(3) (2002) 143–155
8. Hammers, A., Allom, R., Koepp, M.J., Free, S.L., Myers, R., Lemieux, L., Mitchell, T.N., Brooks, D.J., Duncan, J.S.: Three-dimensional maximum probability atlas of the human brain, with particular reference to the temporal lobe. Hum Brain Mapp **19** (2003) 224–247
9. Gousias, I.S., Rueckert, D., Heckemann, R.A., Dyet, L.E., Boardman, J.P., Edwards, A.D., Hammers, A.: Automatic segmentation of brain MRIs of 2-year-olds into 83 regions of interest. Neuroimage **40** (2008) 672–684
10. Thévenaz, P., Unser, M.: Optimization of mutual information for multiresolution image registration. IEEE Trans Image Proc **9**(12) (2000) 2083–2099
11. Tustison, N.J., Avants, B.B., Cook, P.A., Zheng, Y., Egan, A., Yushkevich, P.A., Gee, J.C.: N4ITK: improved N3 bias correction. IEEE Trans Med Imag **29**(6) (2010) 1310–1320
12. Heckemann, R.A., Hajnal, J.V., Aljabar, P., Rueckert, D., Hammers, A.: Automatic anatomical brain MRI segmentation combining label propagation and decision fusion. Neuroimage **33** (2006) 115–126

13. Vapnik, V.N.: The nature of statistical learning theory. Springer-Verlag New York, Inc. (1995)
14. Chang, C.C., Lin, C.J.: LIBSVM: A library for support vector machines. ACM TIST **2**(3) (2011) 27–27
15. Tax, D., Breukelen, M.V., Duin, R., Kittler, J.: Combining multiple classifiers by averaging or by multiplying? Pattern recognition **33** (2000) 1475–1485
16. Abdulkadir, A., Peter, J., Brox, T., Ronneberger, O., Klöppel, S.: Voxel-based multi-class classification of AD, MCI, and elderly controls: Blind evaluation on an independent test set. In: Proc MICCAI workshop Challenge on Computer-Aided Diagnosis of Dementia Based on Structural MRI Data. (2014) 8–15