Voxel-Based Multi-Class Classification of AD, MCI, and Elderly Controls

Blind Evaluation on an Independent Test Set

Ahmed Abdulkadir[1,2,3], Jessica Peter[2], Thomas Brox[3], Olaf Ronneberger[3,4], and Stefan Klöppel[1,2]

[1] Department of Psychiatry and Psychotherapy, University Medical Centre Freiburg, Freiburg, Germany
[2] Department of Neurology, University Medical Centre Freiburg, Freiburg, Germany
[3] Department of Computer Science, University of Freiburg, Freiburg, Germany
[4] BIOSS Centre for Biological Signalling Studies, University of Freiburg, Freiburg, Germany

Abstract. We trained a multi-class support vector machine (SVM) with probabilistic outputs on a large, publicly available sample ($n = 1429$) of healthy controls, individuals with mild cognitive impairment (MCI), and patients with probable Alzheimer's disease (AD). The test performance on a small validation set ($n = 30$) was similar to the cross-validation performance of the training set. Average area under the curve was 0.84 for the validation and 0.79 for the training set. The model was then applied to the test set ($n = 354$) of which no labels were known and the predictions were submitted to the CADDementia Challange.
The method required one hour computation time on a single CPU per subject, and almost no manual intervention.

1 Introduction

Diagnosis of dementia is an important task in clinical routine. In vivo brain imaging supplements clinical assessments by providing information about structure and function and can be used for assisting the diagnosis using automated machine learning methods [11, 15]. In a direct comparison, an automated method for diagnosing Alzheimer's disease (AD) performed as well as or better than clinicians [12]. In a previous study, that we conducted with four different data sets using functional and structural MRI markers, the structural markers were as sensitive as functional imaging markers in diagnosing pre-symptomatic Huntington's disease [2], despite the fact that functional dysfunction precedes structural degeneration in the central nervous system [10]. A direct comparison of different automated methods for diagnosing AD based on structural MRI was conducted by Cuingnet et *al.* [6]. The data used for the study was acquired on multiple centers for the Alzheimer's Disease Neuroimaging Initiative (ADNI). One of the best performing methods [13] used features similar to those in voxel-based morphometry (VBM) [4]. We showed in multiple studies that the typical pre-processing

for VBM studies leads to systematic differences between scanners, but at the same time, the extracted data was sufficiently robust to classify the presence of a disease across sites, and acquisition protocols [13, 1, 14, 20, 3]. Due to the high performance in previous study and robustness in different scenarios, VBM features were selected as the means of classification. In order to increase sensitivity and specificity, we applied data driven feature selection. Further, we aimed to reduce confounding effects of age, head size, and sex.

2 Materials

2.1 Test Data

The test data was provided through the web site on the challenge on Computer-Aided Diagnosis of Dementia (CADDementia) based on structural MRI data[1]. CADDementia provided T1 weighted MRI along with age and sex as basic demographic covariates. Data was acquired on three different scanners with five different scanning sequences. The 354 subjects included in the study were classified in three groups; Healthy controls (HC), individuals with mild cognitive impairment (MCI), and patients with AD. Patients labeled AD met the clinical criteria for probable AD according to [16, 17]. Patients labeled MCI met the criteria stipulated by [19]. One site used three different protocols; the other two sites acquired the images using a single protocol. No diagnostic labels were provided for the test data. Demographic data are shown on Table 1.

Table 1. Demographic data of the training, validation, and test set. HC: healthy controls, MCI: mild cognitive impairment, AD: Alzheimer's disease, F: female, M: male, N.A.: information was not available

	HC/MCI/AD	F/M	age [years]
ADNI	371/631/287	582/707	73.7±7.3
AIBL	79/31/30	80/60	74.5±7.4
CADDementia (validation)	12/9/9	13/17	65.2±7.0
CADDementia (test)	N.A.	141/213	65.1±7.8

2.2 Validation Data

The validation dataset - also provided by the CADDementia Challenge - consisted of thirty examples that also included class labels. Acquisition sites, parameters, and inclusion criteria were identical to the test data set.

[1] `http://caddementia.grand-challenge.org`

2.3 Training Data

Structural MRI from baseline scans of the Alzheimer's Disease Neuroimaging Initiative (ADNI) database[2] [18] and from the Australian Imaging, Biomarker & Lifestyle Flagship Study of Ageing (AIBL) database[3] [8] were used. A goal of ADNI has been to test whether serial magnetic resonance imaging, positron emission tomography, other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of MCI and early AD. The AIBL database consists of several hundred structural scans that were acquired on a single scanner. The study methodology has been reported previously [8]. We used baseline images from about 1429 subjects from AIBL and ADNI. Images from the ADNI were removed if either the subject converted or reverted during the course of the study. A demographic summary of the training data can be found on Table 1.

Four our study, individuals in the training data set were classified in AD, mild cognitive impairment (MCI), and healthy controls (HC).

3 Methods

3.1 Image pre-processing

The goal of image pre-processing was to obtain the input data for the automated classification process. We extracted very high-dimensional GM intensity maps for voxel-wise classification. Pre-processing of images was identical for training, validation, and test sets. Initially, the raw images were coregistered to the canonical T1 template in SPM8 using a rigid registration implemented in the SPM8 toolbox[4]. Then, using VBM8 toolbox[5], we computed voxel-wise densities of gray matter (GM) that were normalized to a reference space. Maps were subsequently modulated by the determinant of the Jacobian of the local deformation field. The modulation thus accounted for non-linear volume changes, but ignored global (affine) volume changes. If multiple baseline images were available for a subject, the mean of all available GM maps was taken. The initial registration failed in some cases, which required manually registering the images to the template. Thus, the employed method was semi automatic, although the manual intervention was minor and did not require expert knowledge. The image preprocessing per image took about one hour on a single core. Computation of one column of the kernel matrix, which included computing pair-wise dot-products between GM maps, took a couple of seconds. Manual registration (required in approximately 10% of the test cases) took a few minutes per subject.

[2] `http://adni.loni.usc.edu`
[3] `http://www.aibl.csiro.au`
[4] `http://www.fil.ion.ucl.ac.uk/spm/software/spm8/`
[5] `http://dbm.neuro.uni-jena.de/vbm`

3.2 Feature Selection

In previous studies [12, 3, 14, 2, 1] we used a linear SVM in combination with high-dimensional GM density maps including either all voxels or selecting voxels *a priori*. Here, we used the linear binary classifier with a hard margin in order to make the feature selection. Using 20% of the training data, one model for every binary classification was trained. For each classifier, using the method proposed by Gaonkar and Davatzikos [9] the p-values of the weights were computed and features were included only if the p-value was lower than a certain threshold. The threshold was 0.0001, 0.001, and 0.01 for ADvsHC, MCIvsNC, and ADvsMCI, respectively. The kernel computed from the subset of significant features was then used for the multi-class classification as explained in the next subsection. Computation for training a model and performing feature selection required about one minute of computation time, provided that the kernel matrix was computed and all required data was in the memory.

3.3 Nuisance Correction

We used kernel regression to correct for confounding effects such as age, sex and total intracranial volume to remove confounds from the linear dot-product matrix of the gray matter values. Computation time for this step was smaller than one second and thus was negligible compared to the time that was required for the pre-processing. Given the kernel matrix $\mathbf{K} \in \mathbb{R}^{N \times N}$, the detrended kernel $\tilde{\mathbf{K}}$ was computed as

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

where $\mathbf{I}$ was the identity matrix and $\mathbf{X} \in \mathbb{R}^{N \times 3}$ the design matrix of N subjects coding sex, age, and total intra-cranial volume. This method was the same as previously proposed by Dukart et *al.* [7], but uses sex and TIV as additional covariate and performs the detrending in kernel space.

Unlike in previous work [14], we did not correct for scanner/sequence for this study, since not enough training data was available. Specifically, only about four images of healthy controls per scanner/sequence were available in the validation set. Of note, the age distribution in the training and test sets differed significantly ($p < 0.05$, Student's t-test). Since age, and AD both are associated with neuronal degeneration in partially overlapping regions, we expected a bias due to age. Specifically, we expected a lower sensitivity, because AD progression is positively correlated with age [7]. Thus younger subjects are less likely to be classified as AD.

3.4 Multi-class Classification

Multi-class linear classification in the one-versus-one setting is not well suited for classification of controls, MCI, and AD, because MCI is in between controls and MCI. We therefore employed a non-linear SVM with radial basis function

$k(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma||\mathbf{x}_i - \mathbf{x}_j||)$. For classification, we used one versus all support vector machine (SVM) as implemented in libsvm [5] with the option for probabilistic multi-class outputs [21]. The SVM parameters were set manually to $C = 2$ and $\gamma = 0.0002$. These combination achieved similar performance on the training and validation set (Figure 1). The performance on the validation set was obtained by using the predictions by the model trained using all training examples that were not used to estimate p-values for feature selection. The performance on the training set was computed in a ten fold cross-validation.

4 Results

Classification results were obtained for the training set ($n = 1429$) by cross-validation, and on the validation set ($n = 30$) by applying the model trained on the entire training set. Test accuracy of binary classification of the validation (training) set of HC, MCI, and AD versus rest was 41.7 % (62.1 %), 66.7 % (64.8 %), and 77.8 % (70.2 %), respectively. Discriminability of the validation set in terms of AUC was highest for AD (96.8%), intermediate for HC (84.7%), and lowest for MCI (67.7%), as shown in Figure 1. The same order was observed in training set as well. Predictions on the test set were submitted to the

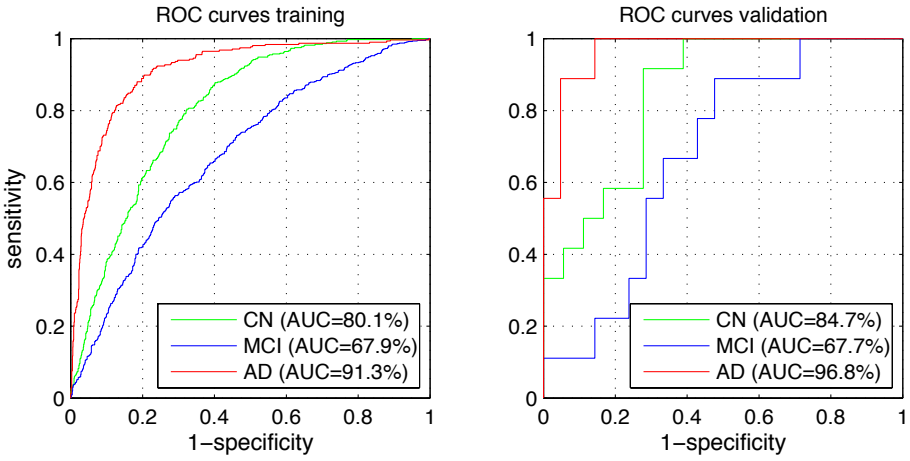

Fig. 1. Performance curves of training and validation set.

CADDementia committee.

5 Discussion

Training and test/validation sets differed in scanner hardware, acquisition parameters and inclusion criteria. Furthermore, the populations of controls and

patients were possibly more distinct in the training set, as conversion and reversion lead to exclusion. In addition, the test population was significantly younger than the test population. The correction for age effects and multi-centric studies conducted previously [3, 14], suggest that the most relevant factor that could lead to a discrepancy in cross-validated training performance versus test performance are the class-wise difference in population.

As expected, classification performance of the three classes HC, and AD was well above chance on the train and validation set. Discriminating MCI from the rest was more certain. Binary classification of HC and AD subjects reached up to 90 % accuracy. These results were obtained with an (almost fully) automated processing pipeline, which required no expert knowledge in the classification process.

One drawback of the presented methods is, that the classification process used the same features for all classification tasks. Although MCI can be seen as pre-state of AD, the optimally discriminative features between HC and MCI are not necessarily the same as the optimally discriminative features between MCI and AD or between HC and AD.

The SVM, as discriminative method, performed well in many similar classification tasks that were evaluated by cross-validation. In the present setting, the validation set remained entirely untouched. This reduced the risk of overfitting the model. However, since the parameters were hand-tuned and picked in such a way that the cross-validated performance on the training set was similar to the validation performance, there was a risk of overfitting to the validation set. We therefore expect a slightly lower performance on the test set.

6 Acknowledgments

This work was funded by a grant from the Deutsche Forschungsgemeinschaft (KL2415/2-1) to SK and OR and supported by the by the Excellence Initiative of the German Federal and State Governments (EXC 294) to OR. Data collection and sharing of the ADNI data set was funded by the Alzheimer's Disease Neuroimaging Initiative (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012).

References

1. Abdulkadir, A., Mortamet, B., Vemuri, P., Jack Jr., C.R., Krüger, G., Klöppel, S.: Effects of hardware heterogeneity on the performance of SVM Alzheimer's disease classifier. NeuroImage 58(3), 785–792 (2011)
2. Abdulkadir, A., Ronneberger, O., Christian Wolf, R., Pfleiderer, B., Saft, C., Klöppel, S.: Functional and Structural MRI Biomarkers to Detect Pre-Clinical Neurodegeneration. Current Alzheimer Research 10(2), 125–134 (2013)
3. Abdulkadir, A., Ronneberger, O., Tabrizi, S.J., Klöppel, S.: Reduction of confounding effects with voxel-wise gaussian process regression in structural mri. In: Pattern Recognition in NeuroImaging (PRNI), 2014 International Workshop on (2014)

4. Ashburner, J., Friston, K.J.: Voxel-based morphometry - The methods. NeuroImage 11(6), 805–821 (2000)
5. Chang, C.C., Lin, C.J.: LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology (TIST) 2(3), 27–27 (2011)
6. Cuingnet, R., Gerardin, E., Tessieras, J., Auzias, G., Lehéricy, S., Habert, M.O., Chupin, M., Benali, H., Colliot, O., Alzheimer's Disease Neuroimaging Initiative, Alzheimer's Disease Neuroimaging Initiative: Automatic classification of patients with Alzheimer's disease from structural MRI: A comparison of ten methods using the ADNI database. NeuroImage 56(2), 766–781 (2011)
7. Dukart, J., Schroeter, M.L., Mueller, K., Alzheimer's Disease Neuroimaging Initiative: Age Correction in Dementia – Matching to a Healthy Brain. PloS one 6(7), e22193 (2011)
8. Ellis, K.A., Bush, A.I., Darby, D., De Fazio, D., Foster, J., Hudson, P., Lautenschlager, N.T., Lenzo, N., Martins, R.N., Maruff, P., Masters, C., Milner, A., Pike, K., Rowe, C., Savage, G., Szoeke, C., Taddei, K., Villemagne, V., Woodward, M., Ames, D.: The Australian Imaging, Biomarkers and Lifestyle (AIBL) study of aging: methodology and baseline characteristics of 1112 individuals recruited for a longitudinal study of Alzheimer's disease. International Psychogeriatrics 21(04), 672–687 (2009)
9. Gaonkar, B., Davatzikos, C.: Analytic estimation of statistical significance maps for support vector machine based multi-variate image analysis and classification. NeuroImage 78, 270–283 (2013)
10. Gili, T., Cercignani, M., Serra, L., Perri, R., Giove, F., Maraviglia, B., Caltagirone, C., Bozzali, M.: Regional brain atrophy and functional disconnection across Alzheimer's disease evolution. Journal of Neurology, Neurosurgery & Psychiatry 82(1), 58–66 (2011)
11. Klöppel, S., Abdulkadir, A., Jack Jr, C.R., Koutsouleris, N.: Diagnostic neuroimaging across diseases. Neuroimage (2012)
12. Klöppel, S., Stonnington, C.M., Barnes, J., Chen, F., Chu, C., Good, C.D., Mader, I., Mitchell, L.A., Patel, A.C., Roberts, C.C., Fox, N.C., Jack, C.R., Ashburner, J., Frackowiak, R.S.J.: Accuracy of dementia diagnosis: a direct comparison between radiologists and a computerized method. Brain 131(Pt 11), 2969–2974 (2008)
13. Klöppel, S., Stonnington, C.M., Chu, C., Draganski, B., Scahill, R.I., Rohrer, J.D., Fox, N.C., Jack, C.R., Ashburner, J., Frackowiak, R.S.J.: Automatic classification of MR scans in Alzheimer's disease. Brain 131(Pt 3), 681–689 (2008)
14. Kostro, D., Abdulkadir, A., Durr, A., Roos, R., Leavitt, B.R.: Correction of inter-scanner and within-subject variance in structural MRI based automated diagnosing. NeuroImage (2014)
15. Lemm, S., Lemm, S., Blankertz, B., Blankertz, B., Dickhaus, T., Dickhaus, T., Müller, K.R., Müller, K.R.: Introduction to machine learning for brain imaging. NeuroImage 56(2), 387–399 (2011)
16. McKhann, G., Drachman, D., Folstein, M., Katzman, R., Price, D., Stadlan, E.M.: Clinical-Diagnosis of Alzheimers-Disease - Report of the Nincds-Adrda Work Group Under the Auspices of Department-of-Health-and-Human-Services Task-Force on Alzheimers-Disease. Neurology 34(7), 939–944 (1984)
17. McKhann, G.M., Knopman, D.S., Chertkow, H.: The diagnosis of dementia due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimer's & Dementia (2011)

18. Mueller, S.G., Weiner, M.W., Thal, L.J., Petersen, R.C., Jack, C., Jagust, W., Trojanowski, J.Q., Toga, A.W., Beckett, L.: The Alzheimer's Disease Neuroimaging Initiative. Neuroimaging Clinics of North America 15(4), 869–877 (2005)
19. Petersen, R.C.: Mild cognitive impairment as a diagnostic entity. Journal of Internal Medicine 256(3), 183–194 (2004)
20. Stonnington, C.M., Tan, G., Klöppel, S., Chu, C., Draganski, B., Jack, C.R., Chen, K., Ashburner, J., Frackowiak, R.S.J.: Interpreting scan data acquired from multiple scanners: a study with Alzheimer's disease. NeuroImage 39(3), 1180–1185 (2008)
21. Wu, T.F., Lin, C.J., Weng, R.C.: Probability estimates for multi-class classification by pairwise coupling. Journal of Machine Learning Research (2004)