ARTICLE IN PRESS

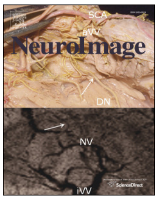
YNIMG-11955; No. of pages: 18; 4C: 10, 11, 12, 13, 14

NeuroImage xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

NeuroImage

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



Standardized evaluation of algorithms for computer-aided diagnosis of dementia based on structural MRI: The CADDementia challenge

Esther E. Bron [a,b,*], Marion Smits [c], Wiesje M. van der Flier [d,e], Hugo Vrenken [f], Frederik Barkhof [f], Philip Scheltens [d], Janne M. Papma [g,c], Rebecca M.E. Steketee [c], Carolina Méndez Orellana [c,g], Rozanna Meijboom [c], Madalena Pinto [h], Joana R. Meireles [h], Carolina Garrett [h,i], António J. Bastos-Leite [j], Ahmed Abdulkadir [k,l,m], Olaf Ronneberger [n,m], Nicola Amoroso [o,p], Roberto Bellotti [o,p], David Cárdenas-Peña [q], Andrés M. Álvarez-Meza [q], Chester V. Dolph [r], Khan M. Iftekharuddin [r], Simon F. Eskildsen [s], Pierrick Coupé [t], Vladimir S. Fonov [u], Katja Franke [v,w], Christian Gaser [v,w], Christian Ledig [x], Ricardo Guerrero [x], Tong Tong [x], Katherine R. Gray [x], Elaheh Moradi [y], Jussi Tohka [y], Alexandre Routier [z,aa], Stanley Durrleman [z,aa], Alessia Sarica [ab], Giuseppe Di Fatta [ac], Francesco Sensi [ad], Andrea Chincarini [ad], Garry M. Smith [ae,ac], Zhivko V. Stoyanov [ae,ac], Lauge Sørensen [af], Mads Nielsen [af], Sabina Tangaro [o], Paolo Inglese [o], Christian Wachinger [ag,ah], Martin Reuter [ag,ah], John C. van Swieten [g], Wiro J. Niessen [a,b,ai], Stefan Klein [a,b], for the Alzheimer's Disease Neuroimaging Initiative [1]

[a] *Biomedical Imaging Group Rotterdam, Department of Medical Informatics, Erasmus MC, Rotterdam, The Netherlands*
[b] *Biomedical Imaging Group Rotterdam, Department of Radiology, Erasmus MC, Rotterdam, The Netherlands*
[c] *Department of Radiology, Erasmus MC, Rotterdam, The Netherlands*
[d] *Alzheimer Center, Department of Neurology, VU University Medical Center, Neuroscience Campus Amsterdam, The Netherlands*
[e] *Department of Epidemiology & biostatistics, VU University Medical Center, Neuroscience Campus Amsterdam, The Netherlands*
[f] *Department of Radiology & Nuclear Medicine, VU University Medical Center, Neuroscience Campus Amsterdam, The Netherlands*
[g] *Department of Neurology, Erasmus MC, Rotterdam, The Netherlands*
[h] *Department of Neurology, Hospital de São João, Porto, Portugal*
[i] *Department of Clinical Neurosciences and Mental Health, Faculty of Medicine, University of Porto, Porto, Portugal*
[j] *Department of Medical Imaging, Faculty of Medicine, University of Porto, Porto, Portugal*
[k] *Department of Psychiatry & Psychotherapy, University Medical Centre Freiburg, Germany*
[l] *Department of Neurology, University Medical Centre Freiburg, Germany*
[m] *Department of Computer Science, University of Freiburg, Germany*
[n] *BIOSS Centre for Biological Signaling Studies, University of Freiburg, Germany*
[o] *National Institute of Nuclear Physics, Branch of Bari, Italy*
[p] *Department of Physics, University of Bari, Italy*
[q] *Signal Processing and Recognition Group, Universidad Nacional de Colombia, Colombia*
[r] *Vision Lab, Old Dominion University, Norfolk, VA 23529, USA*
[s] *Center of Functionally Integrative Neuroscience and MINDLab, Aarhus University, Aarhus, Denmark*
[t] *Laboratoire Bordelais de Recherche en Informatique, Unit Mixte de Recherche CNRS (UMR 5800), PICTURA Research Group, Bordeaux, France*
[u] *McConnell Brain Imaging Centre, Montreal Neurological Institute, McGill University, Montreal, Canada*
[v] *Structural Brain Mapping Group, Department of Neurology, Jena University Hospital, Germany*
[w] *Structural Brain Mapping Group, Department of Psychiatry, Jena University Hospital, Germany*
[x] *Biomedical Image Analysis (BioMedIA) Group, Department of Computing, Imperial College London, UK*
[y] *Department of Signal Processing, Tampere University of Technology, Finland*
[z] *Inserm U1127, CNRS UMR 7225, Sorbonne Universités, UPMC Univ Paris 06 UMR S 1127, Institut du Cerveau et de la Moelle épinière, ICM, Inria Paris-Rocquencourt, F-75013 Paris, France*
[aa] *Centre d'Acquisition et de Traitement des Images (CATI), Paris, France*
[ab] *Bioinformatics Laboratory, Department of Medical and Surgical Sciences, Magna Graecia University, Catanzaro, Italy*
[ac] *School of Systems Engineering, University of Reading, Reading RG6 6AY, UK*
[ad] *National Institute of Nuclear Physics, Branch of Genoa, Italy*
[ae] *Centre for Integrative Neuroscience and Neurodynamics, University of Reading, RG6 6AH, UK*
[af] *Department of Computer Science, University of Copenhagen, Denmark*
[ag] *Computer Science and Artificial Intelligence Lab, MA Institute of Technology, Cambridge, USA*
[ah] *Massachusetts General Hospital, Harvard Medical School, Cambridge, USA*
[ai] *Imaging Physics, Applied Sciences, Delft University of Technology, The Netherlands*

* Corresponding author at: Biomedical Imaging Group Rotterdam, Departments of Medical Informatics and Radiology, Erasmus MC, Rotterdam, The Netherlands.
E-mail addresses: caddementia@bigr.nl, e.bron@erasmusmc.nl (E.E. Bron).

[1] Some data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report.

http://dx.doi.org/10.1016/j.neuroimage.2015.01.048
1053-8119/© 2015 Published by Elsevier Inc.

A R T I C L E I N F O

Article history:
Accepted 24 January 2015
Available online xxxx

Keywords:
Alzheimer's disease
Challenge
Classification
Computer-aided diagnosis
Mild cognitive impairment
Structural MRI

A B S T R A C T

Algorithms for computer-aided diagnosis of dementia based on structural MRI have demonstrated high performance in the literature, but are difficult to compare as different data sets and methodology were used for evaluation. In addition, it is unclear how the algorithms would perform on previously unseen data, and thus, how they would perform in clinical practice when there is no real opportunity to adapt the algorithm to the data at hand. To address these comparability, generalizability and clinical applicability issues, we organized a *grand challenge* that aimed to objectively compare algorithms based on a clinically representative multi-center data set. Using clinical practice as the starting point, the goal was to reproduce the clinical diagnosis. Therefore, we evaluated algorithms for multi-class classification of three diagnostic groups: patients with probable Alzheimer's disease, patients with mild cognitive impairment and healthy controls. The diagnosis based on clinical criteria was used as reference standard, as it was the best available reference despite its known limitations. For evaluation, a previously unseen test set was used consisting of 354 T1-weighted MRI scans with the diagnoses blinded. Fifteen research teams participated with a total of 29 algorithms. The algorithms were trained on a small training set ($n = 30$) and optionally on data from other sources (e.g., the Alzheimer's Disease Neuroimaging Initiative, the Australian Imaging Biomarkers and Lifestyle flagship study of aging). The best performing algorithm yielded an accuracy of 63.0% and an area under the receiver-operating-characteristic curve (AUC) of 78.8%. In general, the best performances were achieved using feature extraction based on voxel-based morphometry or a combination of features that included volume, cortical thickness, shape and intensity. The challenge is open for new submissions via the web-based framework: http://caddementia.grand-challenge.org.

Introduction

In 2010, the number of people over 60 years of age living with dementia was estimated at 35.6 million worldwide. This number is expected to almost double every twenty years (Prince et al., 2013). Accordingly, the cost of care for patients with Alzheimer's disease (AD) and other dementias is expected to increase dramatically, making AD one of the costliest chronic diseases to society (Alzheimer's Association, 2014). Early and accurate diagnosis has great potential to reduce the costs related to care and living arrangements as it gives patients access to supportive therapies that can help them maintain their independence for longer and delay institutionalization (Paquerault, 2012; Prince et al., 2011). In addition, early diagnosis supports new research into understanding the disease process and developing new treatments (Paquerault, 2012; Prince et al., 2011).

While early and accurate diagnosis of dementia is challenging, it can be aided by assessment of quantitative biomarkers. The five most commonly investigated biomarkers were recently included in the revised diagnostic criteria for AD (McKhann et al., 2011; Jack et al., 2011) and in the revised diagnostic criteria for mild cognitive impairment (MCI) due to AD (Albert et al., 2011). These five biomarkers can be divided into two categories: 1) measures of brain amyloid, which include cerebrospinal fluid (CSF) measures of Aβ42 and amyloid positron emission tomography (PET) imaging, and 2) measures of neuronal injury and degeneration, which include CSF tau measurement, fluoro deoxyglucose (FDG) PET and structural MRI (Jack et al., 2012). Of these biomarkers, structural MRI is very important as it is widely available and non-invasive. Also, it is a good indicator of progression to AD in an individual subject, because it becomes abnormal in close temporal proximity to the onset of the cognitive impairment (Jack et al., 2010, 2013).

Structural MRI data can be used to train computer-aided diagnosis methods. These methods make use of machine-learning and other multivariate data-analysis techniques that train a model (classifier) to categorize groups (e.g., patients and controls). Computer-aided diagnosis techniques use features derived from neuroimaging or related data, and may therefore benefit from the large amounts of neuroimaging data that have become available over the last years. The techniques may improve diagnosis as they can potentially make use of group differences that are not noted during qualitative visual inspection of brain imaging data, potentially leading towards an earlier and more objective diagnosis than when using clinical criteria (Klöppel et al., 2012). In addition, computer-aided diagnosis algorithms can be used to 1) improve diagnosis in hospitals with limited neurological and neuroradiological expertise, 2) increase the speed of diagnosis, and 3) aid the recruitment of specific, homogeneous patient populations for clinical trials in pharmacological research (Klöppel et al., 2012).

Structural-MRI-based computer-aided diagnosis methods for dementia, mainly for AD and MCI, have previously shown promising results in the literature. A few years ago, Cuingnet et al. (2011) compared the performance of various feature extraction methods (e.g., voxel-based features, cortical thickness, hippocampal shape and volume) for dementia classification using a support vector machine (SVM) based on structural MRI. Using data from 509 subjects from the Alzheimer's Disease Neuroimaging Initiative (ADNI) cohort, three classification experiments were performed: 1) AD versus healthy controls (CN), 2) patients with MCI versus CN, and 3) MCI who had converted to AD within 18 months (MCI converters, MCIc) versus MCI who had not converted to AD within 18 months (MCI non-converters, MCInc). For the AD/CN classification, the best results were obtained with whole-brain methods (voxel-based and cortical thickness) achieving 81% sensitivity and 95% specificity for the best method. The performances of the MCI/CN classifications were much lower than those of AD/CN, and the MCIc/MCInc classifications yielded no performances better than chance. A recent review paper by Falahati et al. (2014) discussed the literature on AD classification and MCI prediction. The research field of computer-aided diagnosis of dementia based on structural MRI is rather extensive, as evidenced by this paper reviewing 50 papers with at least 50 subjects per diagnostic group. The reviewed papers mainly trained a classification model on the AD/CN groups and subsequently tested this model on both AD/CN and MCIc/MCInc classifications. The paper concluded that classification methods are difficult to compare, because the outcome is influenced by many factors, such as feature extraction, feature selection, robustness of the validation approach, image quality, number of training subjects, demographics, and clinical diagnosis criteria. In general, the accuracy obtained for AD/CN classification was 80–90%, and the accuracy for prediction of MCI conversion is somewhat lower. To promote comparison of algorithms, Sabuncu and Konukoglu (2014) published results based on six large publicly available data sets for AD and other diseases (e.g., schizophrenia, autism). A comparison was performed using four feature extraction strategies, including volumetric and cortical thickness features computed with FreeSurfer (Fischl, 2012), and three types of machine learning techniques (SVM, neighborhood approximation forest (Konukoglu et al., 2013), and relevance voxel machine (Sabuncu and Van Leemput, 2012)). Using the ADNI database, the accuracies ranged from 80–87% for AD/CN classification and 58–66% for MCI/CN classification. The authors made all processed data and computational tools available to promote extension of their benchmark results.

Taken together, these publications show very promising results of algorithms for computer-aided diagnosis of AD and MCI. However, they are difficult to compare as different data sets and methodology were used for evaluation. In addition, it is unclear how the algorithms would perform on previously unseen data, and thus, how they would perform in clinical practice when there is no opportunity to adapt the algorithm to the data at hand. Adaptation of an algorithm would be necessary if the algorithm had been trained or optimized on data that are not representative for the data used in a clinical setting. This seriously hampers clinical implementation of algorithms for computer-aided diagnosis. In medical image analysis research, issues related to comparability and clinical applicability have been addressed in grand challenges[2]. Such grand challenges have the goal of comparing algorithms for a specific task on the same clinically representative data using the same evaluation protocol. In such challenges, the organizers supply reference data and evaluation measures on which researchers can evaluate their algorithms. For this work, we initiated a grand challenge on computer-aided diagnosis of dementia (CADDementia). The CADDementia challenge aims to objectively compare algorithms for classification of AD and MCI based on a clinically representative multi-center data set. We recently organized a workshop at the 17th International Conference on Medical Image Computing and Computer-Assisted Interventions (MICCAI). At this workshop, the methods and results of the algorithms were presented by the 15 teams that originally participated in the challenge.

In the CADDementia challenge, we evaluated algorithms that made a multi-class classification of three diagnostic groups: patients with AD, patients with MCI and CN. The algorithms covered the complete image-processing and classification pipeline starting from structural MRI images. The current clinical diagnosis criteria for AD and MCI (McKhann et al., 2011; Petersen, 2004) were used as the reference standard. Although MCI is known to be heterogeneous, as some of the patients will convert to AD and others will not, it is considered to be one diagnostic entity according to these clinical diagnosis criteria. Hence, in this challenge we did not address prediction of MCI progression, but focused on diagnosis as a crucial first step. Regarding diagnostic classification, binary AD/CN classification overestimates true clinical performance as the most difficult to diagnose patients are left out. Therefore we chose to stay close to the clinical problem and address the three-class classification problem.

An evaluation framework was developed consisting of evaluation measures and a reference data set. All methodological choices for the evaluation framework are based on considerations related to our aim to take a step towards clinical implementation of algorithms for computer-aided diagnosis of dementia. This can be summarized in three key points: comparability, generalizability, and clinical applicability. First, by evaluating all algorithms using the same data set and evaluation methods, the results of the algorithms were better comparable. Second, by providing a previously unseen multi-center data set with blinded ground truth diagnoses, overtraining was avoided and generalizability of the algorithms is promoted. Third, according to the current clinical standards, a multi-class diagnosis of AD, MCI and CN was evaluated. The data for the evaluation framework consisted of clinically-representative T1-weighted MRI scans acquired at three centers. For testing the algorithms, we used scans of 354 subjects with the diagnoses blinded to the participants. Because the aim of this challenge was to evaluate the performance in a clinical situation, when not much data are available, we decided to make only a small training set available. This training set consisted of 30 scans equally representing the three data-supplying centers and the diagnostic groups. The diagnostic labels for the training set were made available. For both training and test data, age and sex were provided. In addition to the provided training data, teams were encouraged to use training data from other sources. For this purpose, most algorithms used data from the Alzheimer's Disease Neuroimaging Initiative (ADNI)[3] or from the Australian Imaging Biomarker and Lifestyle flagship study of aging (AIBL)[4].

In this article, we present the CADDementia challenge for objective comparison of computer-aided diagnosis algorithms for AD and MCI based on structural MRI. The article describes the standardized evaluation framework consisting of evaluation measures and a multi-center structural MRI data set with clinical diagnoses as reference standard. In addition, this paper presents the results of 29 algorithms for the classification of dementia developed by 15 international research teams that participated in the challenge.

Evaluation framework

In this section, we describe our evaluation framework including the data set, the reference standard, the evaluation measures and the algorithm ranking methods.

Web-based evaluation framework

The evaluation framework as proposed in this work is made publicly available through a web-based interface[5]. From this protected web site, the data and the evaluation software are available for download. The data available for download are, for the training set: a total of 30 structural MRI scans from the probable AD, MCI and CN groups including diagnostic label, age, sex and scanner information; and for the test set: 354 structural MRI scans from the probable AD, MCI and CN groups including age, sex and scanner information. The data set and the evaluation measures are detailed in the following sections. Everyone who wishes to validate their algorithm for classification of AD, MCI and CN can use the data set for validation. To be allowed to download the data, participants are required to sign a data usage agreement and to send a brief description of their proposed algorithm. The predictions and a short article describing the algorithm are submitted via the web site[4]. The algorithms are validated with the software described in the following sections. The web site remains open for new submissions to be included in the ranking.

Data

A 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, sex) 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 sex distribution.

The data characteristics are listed in Table 1 and the sizes of the complete data set, training set and test set are listed in Table 2. Most of the data were used for evaluation of performance: the test set. Only after the workshop, we released the class sizes of the test set, marked with an * in Table 2. Therefore only the prior for each class (~1/3) was known to the authors of the algorithms in this paper. A small training data set with diagnostic labels was made available, which consisted of 30 randomly chosen scans distributed over the diagnostic groups. Suitable data from other sources could be used for training (see Training data from other sources section). Q3

[2] http://www.grand-challenge.org.

[3] http://adni.loni.usc.edu.

[4] http://aibl.csiro.au.

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

Table 1
Data characteristics. ASSET: array spatial sensitivity encoding technique, FSPGR: fast spoiled gradient-recalled echo, IR: inversion recovery, MPRAGE: magnetization prepared rapid acquisition gradient echo, TE: echo time, TI: inversion time, TR: repetition time.

	VUMC	EMC	UP
Scanner	3T, GE Healthcare Signa HDxt	3T, GE Healthcare Protocol 1: Discovery MR750 Protocol 2: Discovery MR750 Protocol 3: HD platform	3T, Siemens Trio A Tim
Sequence	3D IR FSPGR	3D IR FSPGR	3D MPRAGE
Scan parameters (TI/TR/TE)	450 ms/7.8 ms/3.0 ms	Protocol 1: 450 ms/7.9 ms/3.1 ms Protocol 2: 450 ms/6.1 ms/2.1 ms Protocol 3: 300 ms/10.4 ms/2.1 ms	900 ms/2300 ms/3.0 ms
Parallel imaging	Yes (ASSET factor = 2)	PROTOCOL 1: YES (ASSET FACTOR = 2) PROTOCOL 2: PARALLEL IMAGING: NO PROTOCOL 3: PARALLEL IMAGING: NO	No
Resolution	0.9 × 0.9 × 1 mm (sagittal)	Protocol 1: 0.9 × 0.9 × 1.0 mm (sagittal) Protocol 2: 0.9 × 0.9 × 0.8 mm (axial) Protocol 3: 0.5 × 0.5 × 0.8 mm (axial)	1 × 1 × 1.2 mm (sagittal)
Number of scans	180	174	30
Age mean (Std)			
Overall	62.2 (5.9) years	68.6 (7.8) years	67.8 (9.1) years
CN	62.1 (6.0) years	65.5 (7.3) years	64.1 (8.8) years
MCI	62.5 (5.5) years	73.1 (5.5) years	70.0 (8.5) years
AD	62.0 (6.0) years	67.2 (8.4) years	64.6 (7.8) years
Percentage of males			
Overall	59%	63%	50%
CN	62%	61%	40%
MCI	68%	69%	60%
AD	47%	57%	50%

Reference standard

The clinical diagnosis was used as the reference standard in this evaluation framework. The data were acquired either as part of clinical routine or as part of a research study at the three centers. All patients underwent neurological and neuropsychological examination as part of their routine diagnostic work up. The clinical diagnosis was established by consensus of a multidisciplinary team. Patients with AD met the clinical criteria for probable AD (McKhann et al., 1984, 2011). MCI patients fulfilled the criteria specified by Petersen (2004): i.e., memory complaints, cognitive impairment in one or multiple domains confirmed by neuropsychological testing, not demented, intact global cognitive function, clinical dementia rating score = 0.5. No hard threshold values were used, but all mentioned criteria were considered. Subjects with psychiatric disorder or other underlying neurological diseases were excluded. Center-specific procedures are specified in the following sections.

VU University Medical Center (VUMC), Amsterdam, The Netherlands

Patients with AD, patients with MCI and controls with subjective complaints were included from the memory-clinic based Amsterdam Dementia Cohort (van der Flier et al., 2014). The protocol for selection of patients and controls was the same as used by Binnewijzend et al. (2013). Controls were selected based on subjective complaints and had at least 1 year of follow-up with stable diagnosis. For the controls, the findings from all investigations were normal; they did not meet the criteria for MCI. The patients' T1w-scans showed no stroke or other abnormalities. All patients gave permission for the use of the data for research.

Table 2
Sizes of the complete data set, training set and test set, distributed over the three data-supplying centers and the three classes. The numbers in the columns marked by a * were unknown to the authors of the algorithms discussed in this paper.

	Complete data set				Training data				Test data			
	n_{AD}*	n_{MCI}*	n_{CN}*	n	n_{AD}	n_{MCI}	n_{CN}	n	n_{AD}*	n_{MCI}*	n_{CN}*	n
VUMC	60	60	60	180	5	4	5	14	55	56	55	166
EMC	42	61	71	174	3	4	6	13	39	57	65	161
UP	10	10	10	30	1	1	1	3	9	9	9	27
Total	112	131	141	384	9	9	12	30	103	122	129	354

Erasmus MC (EMC), Rotterdam, The Netherlands

From the Erasmus MC, the data were acquired either as part of clinical routine or as part of a research study. All patients were included from the outpatient memory clinic. Diagnostic criteria for AD and MCI (Papma et al., 2014) were as mentioned above. Healthy control subjects were volunteers recruited in research studies and did not have any memory complaints. All subjects signed informed consent and the study was approved by the local medical ethical committee.

University of Porto/Hospital de São João (UP), Porto, Portugal

The majority of the included patients were included from the outpatient dementia clinic of Hospital de São João (Porto, Portugal). Two patients with AD were referred from external institutions for a second opinion. In addition, healthy control subjects were volunteers recruited in research studies. All subjects provided consent to be included in this study.

Data preprocessing

The T1w MRI data was anonymized and facial features were masked (Leung et al., 2014). All anonymized scans were visually inspected to check if no brain tissue was accidentally removed by the facial masking. Skull stripping was performed by the participants themselves, if needed for their algorithm. Next to the original anonymized T1w scans, we provided these scans after non-uniformity correction with N4ITK (Tustison et al., 2010) using the following settings: shrink factor = 4, number of iterations = 150, convergence threshold = 0.00001, and initial b-spline mesh resolution = 50 mm. Images were stored in NIfTI-1 file format.[6]

[6] http://nifti.nimh.nih.gov.

Table 3
Confusion matrix for a three-class classification problem.

		True class		
		c_0	c_1	c_2
Hypothesized class	C_0	$n_{0,0}$	$n_{0,1}$	$n_{0,2}$
	C_1	$n_{1,0}$	$n_{1,1}$	$n_{1,2}$
	C_2	$n_{2,0}$	$n_{2,1}$	$n_{2,2}$
Column totals:		n_0	n_1	n_2

Evaluation measures

The performance of the algorithms was quantified by the classification accuracy, area under the receiver-operating-characteristic (ROC) curve (AUC) and the true positive fraction for the three classes. The performance was evaluated on all 354 test subjects (ALL) and in addition per data-providing center (VUMC, EMC, UP).

Accuracy for multi-class classification

Classification accuracy is in case of a binary design defined as the number of correctly classified samples divided by the total number of samples. For extending the accuracy measure to three-class classification, there are two main options (Hand and Till, 2001). The difference between these is whether or not the difference between the two other classes is taken into account when the performance for one class is assessed.

To determine a simple measure of accuracy, all diagonal elements of the confusion matrix (Table 3), the true positives (tp) and true negatives (tn), are divided by the total number of samples (n):

$$\text{accuracy} = \frac{tp + tn}{n} = \frac{n_{0,0} + n_{1,1} + n_{2,2}}{n_0 + n_1 + n_2}. \tag{1}$$

The alternative, the average accuracy,

$$\text{accuracy}_{\text{average}} = \frac{1}{L}\sum_{i=0}^{L-1}\frac{tp_i + tn_i}{n} = \frac{1}{L}\sum_{i=0}^{L-1}\frac{n_{i,i} + \Sigma_{j=0,j\neq i}^{L-1}\Sigma_{k=0,k\neq i}^{L-1} n_{j,k}}{n}, \tag{2}$$

asseses the accuracy separately for each class without distinguishing between the two other classes. For calculation of the accuracy for $i = 0$, the true positive samples (tp_i) are $n_{0,0}$. The true negative samples in this case (tn_i) are $n_{1,1}$, $n_{1,2}$, $n_{2,1}$ and $n_{2,2}$. The separate per-class accuracies are averaged to yield the final accuracy. L denotes the number of classes.

Eq. (2) is mainly applicable when the class sizes are very different. In this evaluation framework, we use the accuracy in Eq. (1) as it provides a better measure for the overall classification accuracy (Hand and Till, 2001).

AUC for multi-class classification

The performance of a binary classifier can be visualized as an ROC curve by applying a range of thresholds on the probabilistic output of the classifier and calculating the sensitivity and specificity. The AUC is a performance measure which is equivalent to the probability that a randomly chosen positive sample will have a higher probability of being positively classified than a randomly chosen negative sample (Fawcett, 2006). The advantage of ROC analysis – and accordingly the AUC measure – is that the performance of a classifier is measured independently of the chosen threshold. When more than two dimensions are used the ROC-curve becomes more complex. With L classes, the confusion matrix consists of L^2 elements: L diagonal elements denoting the correct classifications, and $L^2 - L$ off-diagonal elements denoting the incorrect classifications. For ROC analysis, the trade-off between these off-diagonal elements is varied. For three-class classification, there are $3^2 - 3 = 6$ off-diagonal elements, resulting in a 6-dimensional ROC-curve. Therefore, for simplicity, multi-class ROC analysis is often generalized to multiple per-class or pairwise ROC curves (Fawcett, 2006).

Similarly to accuracy in the previous section, the multi-class AUC measure can be defined in two ways. The difference between the two definitions is whether or not the third class is taken into account when the difference between a pair of classes is assessed.

First, Provost and Domingos (2001) calculate the multi-class AUC by generating an ROC curve for every class and measuring the AUCs. These per-class AUCs are averaged using the class priors $p(c_i)$ as weights:

$$\text{AUC}_1 = \sum_{i=0}^{L-1}\text{AUC}(c_i)\cdot p(c_i). \tag{3}$$

This method has the advantage that the separate ROC curve can be easily generated and visualized. The method calculates an AUC for every class separately, which is sensitive for the class distributions. Even though the class priors are used in averaging, the total AUC still depends on the class sizes.

Second, Hand and Till (2001) proposed a different method for multi-class AUC which is based on calculating an AUC for every pair of classes, without using information from the third class. The method is based on the principle that the AUC is equivalent to the probability that a randomly chosen member of class c_i will have a larger estimated probability of belonging to class C_i than a randomly chosen member of class c_j. Using this principle, the AUC can also be calculated directly from the ranks of test samples instead of first calculating the ROC curves. To achieve this, the class c_i and c_j test samples are ranked in increasing order of the output probability for class C_i. Let S_i be the sum of the ranks of the class c_i test samples. The AUC for a class c_i given another class, $\hat{A}(c_i|c_j)$, is then given by

$$\hat{A}\left(c_i|c_j\right) = \frac{S_i - n_i(n_i+1)/2}{n_i n_j}, \tag{4}$$

see Hand and Till (2001) for the complete derivation.

For situations with three or more classes, $\hat{A}(c_i|c_j) \neq \hat{A}(c_j|c_i)$. Therefore, the average of both is used:

$$\hat{A}\left(c_i, c_j\right) = \frac{\hat{A}\left(c_i|c_j\right) + \hat{A}\left(c_j|c_i\right)}{2}. \tag{5}$$

The overall AUC is obtained by averaging this over all pairs of classes:

$$\text{AUC}_2 = \frac{2}{L(L-1)}\sum_{i=0}^{L-1}\sum_{j=0}^{i}\hat{A}\left(c_i, c_j\right), \tag{6}$$

in which the number of pairs of classes is $\frac{L(L-1)}{2}$.

In contrast to the accuracy, AUC measurement does not require a threshold on the classifier's output probabilities and therefore the AUC generally does not rely on the class priors (Hand and Till, 2001). However, the first multi-class approach is dependent on the class priors as these are used for averaging the per-class AUCs.

Please cite this article as: Bron, E.E., et al., Standardized evaluation of algorithms for computer-aided diagnosis of dementia based on structural MRI: The CADDementia challenge, NeuroImage (2015), http://dx.doi.org/10.1016/j.neuroimage.2015.01.048

Therefore for this challenge, the second approach for AUC was adopted (Fawcett, 2006).

True positive fraction

For binary classifications in computer-aided diagnosis, often the sensitivity and the specificity are reported in addition to the accuracy. For this multi-class application, the true positive fractions (TPF) for the three classes provide the same information:

$$\mathrm{TPF}_i = \frac{n_{i,i}}{n_i}, \quad i \in 0,1,2. \tag{7}$$

The TPF for the diseased class ($\mathrm{TPF}_{\mathrm{AD}}$; $\mathrm{TPF}_{\mathrm{MCI}}$) can be interpreted as the two-class sensitivity, and the TPF for the control group equals the two-class specificity.

Submission guidelines

In this challenge, the participating teams were allowed to submit up to five algorithms. Submitting the diagnostic label for each sample of the test set was obligatory. Additionally, the output probabilities for each label were requested but this was optional to not rule out approaches that do not produce probabilistic outcomes. Every team had to write one full workshop paper describing their algorithms in the style of Lecture Notes in Computer Science.

Final results and ranking

For every algorithm, a confusion matrix was made based on the test data. Accuracy (Eq. (1)) and the TPF_i (Eq. (7)) for the three classes were calculated from the diagnostic labels. For every class, an ROC curve and per-class AUCs were calculated from the output probabilities reduced to a binary solution, e.g., AD versus non-AD, showing the ability of the classifier to separate that class from the other two classes. An overall AUC was calculated using Eqs. (4)–(6). Confidence intervals on the accuracy, AUC and TPF were determined with bootstrapping on the test set (1000 resamples). To assess whether the difference in performance between two algorithms was significant, the McNemar test (Dietterich, 1996) was used. Evaluation measures were implemented in Python scripting language (version 2.7.6) using the libraries Scikit-learn[7] (version 14.1) and Scipy[8] (version 14.0).

If an algorithm failed to produce an output for certain subjects, these subjects were considered misclassified as a fourth class. This fourth class was considered in the calculation of all performance measures. For calculation of the per-class ROC curves, sensitivity and specificity were determined on the subjects that were classified by the algorithm and subsequently scaled to the total data set to take missing samples into account.

The participating algorithms were ranked based on accuracy of diagnosing the cases in the test set. Algorithms for which output probabilities were available were also ranked based on the AUC of diagnosing the cases in the test set. The algorithm with the best accuracy (rank = 1) on the test set, was considered the winning algorithm. In case two or more algorithms had equal accuracies, the average rank was assigned to these algorithms.

MICCAI 2014 workshop

The evaluation framework was launched in March 2014 and the deadline for the first submissions was in June 2014. The evaluation framework and the results of the first participating algorithms were presented at the Challenge on Computer-Aided Diagnosis of Dementia Based on Structural MRI Data workshop that was organized on September 18th 2014 in conjunction with the 17th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI) conference in Boston (USA).

We invited around 100 groups from academia and industry by email to participate in the challenge. The challenges were advertised by the MICCAI organizers as well. Eighty-one teams made an account on the web site, of which 47 sent a data usage agreement and a brief description of the proposed algorithm, which was required for downloading the data. Finally, 16 teams submitted results, of which 15 were accepted for participation in the workshop. One team was excluded from participation because their workshop submission did not meet the requirements and because they only submitted results for AD/CN classification. The 15 participating teams submitted a total of 29 algorithms. These algorithms are described in the Algorithms section. More details can be found in the short articles that all authors submitted for the workshop (Bron et al., 2014).

Training data from other sources

In addition to the provided training data set of 30 scans, other sources of training data could be used by the participants. All algorithms except for two were trained on data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database[9]. The ADNI was launched in 2003 by the National Institute on Aging (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the Food and Drug Administration (FDA), private pharmaceutical companies and non-profit organizations, as a $60 million, 5-year public-private partnership. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessments can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). Determination of sensitive and specific markers of very early AD progression is intended to aid researchers and clinicians to develop new treatments and monitor their effectiveness, as well as lessen the time and cost of clinical trials. For up-to-date information, see www.adni-info.org. Acquisition of these data had been performed according to the ADNI acquisition protocol (Jack et al., 2008).

Two teams additionally trained on data from the Australian Imaging Biomarkers and Lifestyle (AIBL) flagship study of aging[10] funded by the Commonwealth Scientific and Industrial Research Organisation (CSIRO). These data were collected by the AIBL study group. AIBL study methodology has been reported previously (Ellis et al., 2009).

Algorithms

In this section, the 29 algorithms submitted by 15 teams are summarized. In Table 4, an overview of the algorithms is presented including a listing of the size of the used training set and the performance on the provided 30 training scans.

Abdulkadir et al.

Algorithm: *Abdulkadir* (Abdulkadir et al., 2014).

Features: Voxel-based morphometry (VBM) of gray matter (GM).

Classifier: Radial-basis kernel SVM.

Training data: 1289 ADNI subjects and 140 AIBL subjects. The 30 training subjects provided by the challenge were used for parameter selection.

Feature selection: SVM significance maps (Gaonkar and Davatzikos, 2013).

Confounder correction: Yes, for age, sex and intracranial volume (ICV) using kernel regression.

[7] http://scikit-learn.org.

[8] http://www.scipy.org.

[9] http://adni.loni.usc.edu.

[10] http://aibl.csiro.au.

Please cite this article as: Bron, E.E., et al., Standardized evaluation of algorithms for computer-aided diagnosis of dementia based on structural MRI: The CADDementia challenge, NeuroImage (2015), http://dx.doi.org/10.1016/j.neuroimage.2015.01.048

Table 4
Overview of the participating algorithms. The training accuracy was computed on the 30 training subjects by training on the data from different sources only. As indicated below, three algorithms instead trained on all data using 5-fold or 10-fold cross-validation.

	2*Algorithm	2*Features	2*Classifier	Size training data	Training accuracy [%]
1	*Abdulkadir*	VBM	SVM	1492	60
2	*Amoroso*	Volume and intensity relations	Neural network	288	$67^{5\text{-fold}}$
3	*Cárdenas-Peña*	Raw intensities	SVM	451	83
4	*Dolph*	Volumes	SVM	30	$80^{10\text{-fold}}$
5	*Eskildsen-ADNI1*	Volume and intensity relations	Regression	794	77
6	*Eskildsen-ADNI2*	Volume and intensity relations	Regression	304	70
7	*Eskildsen-Combined*	Volume, thickness and intensity relations	Regression	1098	73
8	*Eskildsen-FACEADNI1*	Volume, thickness and intensity relations	Regression	794	70
9	*Eskildsen-FACEADNI2*	Volume, thickness and intensity relations	Regression	304	67
10	*Franke*	VBM	Regression	591	90
11	*Ledig-ALL*	Volume, thickness and intensity relations	Random forest	734	68
12	*Ledig-CORT*	Cortical thickness	Random forest	734	58
13	*Ledig-GRAD*	Intensity relations	Random forest	734	67
14	*Ledig-MBL*	Intensity relations	Random forest	734	66
15	*Ledig-VOL*	Volumes	Random forest	734	56
16	*Moradi*	VBM	SVM	835	77
17	*Routier-adni*	Shapes	Regression	539	50
18	*Routier-train*	Shapes	Regression	539	73
19	*Sarica*	Volume and thickness	SVM	210	70
20	*Sensi*	Intensity relations	Random forest, SVM	581	73
21	*Smith*	Volume and raw intensities	Regression	189	80
22	*Sørensen-equal*	Volume, thickness, shape, intensity relations	LDA	679	73
23	*Sørensen-optimized*	Volume, thickness, shape, intensity relations	LDA	679	80
24	*Tangaro*	Volume and thickness	SVM	190	$73^{5\text{-fold}}$
25	*Wachinger-enetNorm*	Volume, thickness and shape	Regression	781	73
26	*Wachinger-man*	Volume, thickness and shape	Regression	781	67
27	*Wachinger-step1*	Volume, thickness and shape	Regression	781	77
28	*Wachinger-step1Norm*	Volume, thickness and shape	Regression	781	77
29	*Wachinger-step2*	Volume, thickness and shape	Regression	781	80

Automatic: Yes. Registration required manual intervention for some subjects.

Computation time: 1 h per subject.

Amoroso et al.

Algorithm: *Amoroso* (Amoroso et al., 2014).

Features: Volume features (FreeSurfer) and intensity features of the peri-hippocampal region (mean, standard deviation, kurtosis, and skewness).

Classifier: Back propagation neural network (1 hidden layer, 10 neurons). For every pairwise classification, 100 networks were trained on 50 randomly selected features. For final classification, the output scores were averaged.

Training data: 258 ADNI subjects + the 30 training subjects.

Feature selection: Unsupervised filter based on correlation and linear dependencies.

Confounder correction: –.

Automatic: Yes.

Computation time: 13 h per subject, of which 12 h were due to FreeSurfer processing time.

Cárdenas-Peña et al.

Algorithm: *Cárdenas-Peña* (Cárdenas-Peña et al., 2014)

Features: Features were based on similarities in MRI intensities between subjects. As a first step, similarities between slices of a subject's scan were calculated along each axis resulting in an interslice kernel (ISK) matrix. Second, pairwise similarities between the subjects' ISK matrices were computed using the Mahalanobis distance. Third, the dependence between the resulting matrix of the previous step and the class labels was optimized using a kernel centered alignment function. The eigenvalues of the resulting matrix were used as features.

Classifier: Radial-basis kernel SVM.

Training data: 451 ADNI subjects.

Feature selection: –.

Confounder correction: –.

Automatic: Yes.

Computation time: 22.3 s per subject.

Dolph et al.

Algorithm: *Dolph* (Dolph et al., 2014).

Features: Volume ratio of white matter (WM) and CSF for axial slices.

Classifier: Radial-basis kernel SVM.

Training data: The 30 training subjects.

Feature selection: SVM wrapper.

Confounder correction: –.

Automatic: Yes, but parameters for skull stripping and tissue segmentation were set manually.

Computation time: 30 min per subject.

Eskildsen et al.

Algorithm: *Eskildsen* (Eskildsen et al., 2014, 2015):

Features: Volume and intensity features of the hippocampus (HC) and entorhinal cortex (ERC) were calculated with Scoring by Non-local Image Patch Estimator (SNIPE). By comparing small image patches to a training library, this method segmented these brain regions and computed a grading value per voxel reflecting the proximity between a patch and the classes. As features, the volumes and average grading values for HC and ERC were used.

Cortical thickness was computed with Fast Accurate Cortex Extraction (FACE). As features, the mean cortical thickness was used in regions with large differences in cortical thickness between the classes.

These features were combined:

1. *Eskildsen-FACEADNI1*: Volume, intensity and cortical thickness features
2. *Eskildsen-ADNI1*: Volume and intensity features
3. *Eskildsen-FACEADNI2*: Volume, intensity and cortical thickness features
4. *Eskildsen-ADNI2*: Volume and intensity features

5. *Eskildsen-Combined*: A combination of the other four methods by averaging the posterior probabilities

Classifier: Sparse logistic regression. Ensemble learning was used to combine twenty-five models that were trained using different parameters and different samplings of the data.

Training data:

1. *Eskildsen-FACEADNI1*: 794 ADNI1 subjects
2. *Eskildsen-ADNI1*: 794 ADNI1 subjects
3. *Eskildsen-FACEADNI2*: 304 ADNI2 subjects
4. *Eskildsen-ADNI2*: 304 ADNI2 subjects
5. *Eskildsen-Combined*: 794 ADNI1 and 304 ADNI2

Regression parameters were optimized on the 30 training subjects.

Feature selection: –.

Confounder correction: Yes, for age, sex and differences in class priors.

Automatic: Yes.

Computation time: 55 min per subject.

Franke et al.

Algorithm: *Franke* (Franke and Gaser, 2014)

Features: VBM of GM and WM.

Classifier: Relevance vector regression. An age prediction model was trained on healthy controls. Classification of AD, MCI and CN was performed by thresholding the age difference between the predicted age and the real age.

Training data: 561 healthy subjects (IXI cohort[11]). The age difference threshold was optimized on the 30 training subjects.

Feature selection: Principal component analysis (PCA).

Confounder correction: Yes. Age was used in the modeling. Separate models were trained for males and females.

Automatic: Yes, except for the optimization of the age difference threshold.

Computation time: 10 min per subject.

Ledig et al.

Algorithm: *Ledig* (Ledig et al., 2014).

Features: Five feature sets were used:

1. *Ledig-VOL*: Volumes of regions-of-interest (ROIs) obtained with multi-atlas label propagation and expectation-maximization-based refinement (MALP-EM).
2. *Ledig-CORT*: Cortical thickness features (mean and standard deviation) and surface features (surface area, relative surface area, mean curvature, Gaussian curvature) for the whole cortex and cortex regions.
3. *Ledig-MBL*: Features describing the manifold-based learning (MBL) space. The manifold was trained on intensity texture descriptors for 1701 ADNI subjects.
4. *Ledig-GRAD*: Intensity patterns in patches. Grading features were learned using data of 629 ADNI and the 30 training subjects. The method was based on SNIPE (Eskildsen et al., 2014).
5. *Ledig-ALL*: A combination of all features above.

Classifier: Random forest classifier.

Training data: 734 ADNI subjects.

Feature selection: Only for *Ledig-MBL* and *Ledig-Grad*. *Ledig-MBL*: PCA and sparse regression using local binary intensity patterns and mini mental-state examination (MMSE) scores of 292 ADNI subjects. *Ledig-Grad*: elastic net sparse regression.

Confounder correction: –.

Automatic: Yes.

Computation time: 4 h per subject.

Moradi et al.

Algorithm: *Moradi* (Moradi et al., 2014).

Features: VBM of GM.

Classifier: Transductive SVM. Unsupervised domain adaptation was used to adapt the ADNI data to the 30 training sets. To increase both class separability and within-class clustering, low density separation was applied to both labeled and unlabeled data. The SVM used a graph-distance derived kernel. The classifications were repeated 101 times and combined with majority vote. Classification was performed in two stages: 1) AD/CN classification, 2) a further division of AD/MCI and CN/MCI.

Training data: 835 ADNI subjects.

Feature selection: Elastic net logistic regression.

Confounder correction: Yes. Age effects were removed with linear regression.

Automatic: Yes.

Computation time: 10 min per subject.

Routier et al.

Algorithm: *Routier* (Routier et al., 2014).

Features: Features derived from shape models of 12 brain structures: caudate nucleus, putamen, pallidum, thalamus, hippocampus and amygdala of each hemisphere. The segmentations were obtained with FreeSurfer. 3D triangular meshes of the shapes were obtained with a marching-cubes algorithm. Anatomical models of the shapes were built for AD, MCI and CN using Deformetrica[12] (Durrleman et al., 2014). The shape models were registered to the test subjects, thus computing the likelihood of the data for each model.

Classifier: Maximum-likelihood regression.

Training data: 509 ADNI subjects.

Thresholds were optimized on:

1. *Routier-adni*: the ADNI data
2. *Routier-train*: the 30 training sets

Feature selection: –.

Confounder correction: –.

Automatic: Yes.

Computation time: 4 days for training the anatomical models and additionally 11 h per subject.

Sarica et al.

Algorithm: *Sarica* (Sarica et al., 2014).

Features: Volume and cortical thickness features (FreeSurfer).

Classifier: Radial-basis kernel SVM. Pairwise classifications were combined with voting.

Training data: 210 ADNI subjects. The 30 training sets were used for model selection.

Feature selection: Three methods (correlation filter, random forest filter, and SVM wrapper) and their combination were evaluated. The models with best performance on the 30 training subjects were selected: the methods without ICV correction using the random forest filter (AD/CN, AD/MCI) and the correlation filter (CN/MCI).

Confounder correction: Yes. Age and sex were included as features. Experiments were performed with and without ICV correction.

Automatic: Yes, except for the model selection.

Computation time: 5 h per subject.

Note: Three test subjects were excluded as FreeSurfer failed.

Sensi et al.

Algorithm: *Sensi* (Sensi et al., 2014).

Features: Intensity and textural features of cuboid regions in the medial temporal lobe. The cuboid regions were placed around the entorhinal cortex, perirhinal cortex, hippocampus, and parahippocampal

[11] http://www.brain-development.org.

[12] http://www.deformetrica.org.

gyrus. In addition, two control regions that are relatively spared by AD (rolandic areas) were placed. In each region, voxel intensities were normalized for each tissue by the tissue mean calculated in an additional cuboid region positioned around the corpus callosum in a reference template. To obtain the features, the voxels in the cuboid volumes were processed with 18 filters (e.g., Gaussian mean, standard deviation, range, entropy, Mexican hat) with different voxel radii.

Classifier: Radial-basis kernel SVM and random forest classifier, combined by the weighted mean. Using probability density functions estimated on the 30 training subjects, the output probabilities were mapped to the classes.

Training data: 551 ADNI subjects + the 30 training subjects. For the ADNI data, MCIc patients were included in the AD group.

Feature selection: Random forest classifier.

Confounder correction: –.

Automatic: Yes.

Computation time: 45 min per subject.

Smith et al.

Algorithm: *Smith* (Smith et al., 2014).

Features: Surface area, volume and fragility of a thresholded ROI containing mainly the WM. The fragility originates from network theory and measures how close the structure is from breaking apart into smaller components.

Classifier: Multinomial logistic regression.

Training data: 189 ADNI subjects + the 30 training subjects.

Feature selection: –.

Confounder correction: Yes. Age was used as a feature. Separate thresholds for males and females were used for the WM ROI.

Automatic: Yes, except for the optimization of the threshold for the WM ROI.

Computation time: 7–24 min per subject.

Sørensen et al.

Algorithm: *Sørensen* (Sørensen et al., 2014).

Features: Five types of features were combined: 1) volumes of seven bilaterally joined regions (amygdala, caudate nucleus, hippocampus, pallidum, putamen, ventricles, whole brain; FreeSurfer), 2) cortical thickness of four lobes and the cingulate gyrus (FreeSurfer), 3) the volume of both hippocampi segmented with a multi-atlas, non-local patch-based segmentation technique (using 40 manual segmentations from the Harmonized Hippocampal Protocol as atlases (Frisoni and Jack, 2011)), 4) two hippocampal shape scores (left and right) computed by a Naive Bayes classifier on the principal components of surface landmarks trained on ADNI and AIBL AD/CN data, 5) a hippocampal texture score computed by a radial-basis kernel SVM on a Gaussian-filter-bank-based texture descriptor trained on ADNI and AIBL AD/CN data.

Classifier: Regularized linear discriminant analysis (LDA).

Different priors were used:

1. *Sørensen-equal*: equal class priors
2. *Sørensen-optimized*: class priors optimized on the 30 training subjects ($p_{\mathrm{CN}} = \frac{1}{8}, p_{\mathrm{MCI}} = \frac{3}{8}, p_{\mathrm{AD}} = \frac{1}{2}$).

Training data: 504 ADNI and 145 AIBL subjects.

Feature selection: –.

Confounder correction: Yes. Features were z-score transformed dependent on the age. Volume features were explicitly normalized by dividing by ICV.

Automatic: Yes.

Computation time: 19 h per subject, of which 18 h were due to FreeSurfer processing time.

Tangaro et al.

Algorithm: *Tangaro* (Tangaro et al., 2014).

Features: Volume and cortical thickness features (FreeSurfer). Hippocampus segmentations were obtained with random forest classification based on Haar-like features.

Classifier: Linear SVM. Pairwise classifications were combined by multiplication and normalization of the output probabilities.

Training data: 160 ADNI subjects + the 30 training subjects

Feature selection: –.

Confounder correction: –.

Automatic: Yes.

Computation time: 13 h per subject, of which 12 h were due to FreeSurfer processing time.

Wachinger et al.

Algorithm: *Wachinger* (Wachinger et al., 2014a).

Features: Volume, cortical thickness and shape features (FreeSurfer). For computation of shape features, a spectral shape descriptor ('ShapeDNA') was derived from volume (tetrahedral) and surface (triangular) meshes obtained from FreeSurfer labels with the marching cubes algorithm. This shape descriptor computes the intrinsic geometry with a method that does not require alignment between shapes (Reuter et al., 2006). Using 50 eigenvalues of the shape descriptor, two types of shape features were computed (Wachinger et al., 2014b): 1) the principal component for 44 brain structures ('BrainPrint'), and 2) the shape differences between left and right for white matter, gray matter, cerebellum white matter and gray matter, striatum, lateral ventricles, hippocampus and amygdala.

Classifier: Generalized linear model.

Training data: 751 ADNI subjects + the 30 training subjects.

Feature selection: Five methods were used:

1. *Wachinger-man*: manual selection of ROIs.
2. *Wachinger-step1*: stepwise selection using the Akaike information criterion on ADNI.
3. *Wachinger-step2*: stepwise selection using the Akaike information criterion on ADNI and the provided training data.
4. *Wachinger-step1Norm*: stepwise selection using the Akaike information criterion on ADNI with normalization by the Riemannian volume of the structure.
5. *Wachinger-enetNorm*: elastic net regularization with normalization by the Riemannian volume of the structure.

Confounder correction: Yes. Age was corrected for by linear regression, volume measures were normalized by the ICV.

Automatic: Yes.

Computation time: 17.4 h per subject, of which 16.8 h were due to FreeSurfer processing.

Results

The results presented in this section are based on the 29 algorithms presented at the CADDementia workshop (MICCAI 2014 workshop section). Q6

Classification performance

Table 5 and Fig. 1 show the accuracies and TPFs for the algorithms. The algorithms are ranked by accuracy. The accuracies ranged from 32.2% to 63.0%. As a three-class classification problem was analyzed, the accuracy for random guessing would be ~33.3%. If all subjects were estimated to be in the largest class (CN), the accuracy would be $n_{\mathrm{CN}}/n = {}^{129}/_{354} = 36.4\%$. It can thus be observed that 27 out of the 29 algorithms performed significantly better than guessing. The algorithm with the best accuracy was *Sørensen-equal*, with an accuracy of 63.0%. According to the McNemar test, *Sørensen-equal* was significantly better than most other algorithms ($p < 0.05$) except for *Sørensen-optimized* ($p = 0.23$), *Wachinger-enetNorm* ($p = 0.21$), *Moradi* ($p = 0.14$), *Ledig-*

Table 5
Accuracy and true positive fractions (TPFs) on the test data for the participating algorithms. CI = 95% confidence interval estimated with bootstrapping.

Rank	Algorithm	Accuracy [%] (CI)	TPF_{CN} [%] (CI)	TPF_{MCI} [%] (CI)	TPF_{AD} [%] (CI)
1	*Sørensen-equal*	63.0 (57.9–67.5)	96.9 (92.9–99.2)	28.7 (21.3–37.4)	61.2 (51.6–69.8)
2	*Sørensen-optimized*	59.9 (54.8–64.7)	70.5 (62.8–77.8)	41.0 (33.3–50.0)	68.9 (59.6–77.2)
3	*Wachinger-enetNorm*	59.0 (54.0–63.6)	72.1 (63.4–79.2)	51.6 (43.5–61.3)	51.5 (41.5–61.2)
4	*Ledig-ALL*	57.9 (52.5–62.7)	89.1 (83.7–93.8)	41.0 (32.4–49.6)	38.8 (30.7–50.0)
5	*Moradi*	57.6 (52.3–62.4)	57.4 (48.7–66.1)	59.8 (51.3–68.1)	55.3 (46.7–65.2)
6	*Franke*	56.2 (50.8–61.3)	58.9 (50.4–67.5)	43.4 (34.8–51.7)	68.0 (58.8–77.1)
7.5	*Sensi*	55.1 (50.0–60.2)	71.3 (63.6–78.8)	40.2 (31.2–49.6)	52.4 (42.7–62.0)
7.5	*Ledig-CORT*	55.1 (49.7–59.9)	68.2 (60.5–76.0)	45.1 (35.3–53.4)	50.5 (41.2–60.5)
9.5	*Ledig-GRAD*	54.0 (48.9–59.3)	87.6 (81.7–92.6)	37.7 (29.3–47.5)	31.1 (22.4–40.4)
9.5	*Wachinger-step1*	54.0 (48.9–59.0)	68.2 (60.2–75.4)	41.0 (31.9–50.9)	51.5 (42.2–61.1)
12.5	*Wachinger-step1Norm*	53.7 (48.6–58.8)	63.6 (54.9–71.9)	47.5 (38.4–56.6)	48.5 (39.6–59.1)
12.5	*Sarica*	53.7 (48.3–58.8)	65.9 (57.4–74.2)	39.3 (30.0–48.2)	55.3 (44.9–64.9)
12.5	*Wachinger-step2*	53.7 (47.5–58.8)	66.7 (58.1–74.1)	38.5 (30.1–48.1)	55.3 (45.5–65.0)
12.5	*Abdulkadir*	53.7 (48.3–58.2)	45.7 (37.0–53.6)	65.6 (56.1–73.0)	49.5 (39.4–58.8)
15	*Ledig-MBL*	53.4 (47.7–57.9)	82.9 (76.0–88.7)	43.4 (35.1–52.9)	28.2 (20.2–37.4)
16	*Wachinger-man*	53.1 (47.7–57.9)	61.2 (53.5–69.6)	60.7 (51.7–70.0)	34.0 (25.7–44.7)
17.5	*Eskildsen-ADNI1*	52.0 (46.6–56.8)	65.1 (56.9–73.2)	32.0 (24.1–40.9)	59.2 (49.5–68.3)
17.5	*Eskildsen-FACEADNI1*	52.0 (46.9–57.1)	65.1 (56.6–73.1)	36.1 (28.1–45.5)	54.4 (44.6–63.6)
19	*Eskildsen-Combined*	51.1 (45.5–56.2)	64.3 (56.2–72.3)	35.2 (27.1–44.3)	53.4 (43.0–62.9)
20	*Dolph*	49.7 (44.6–54.8)	84.5 (77.9–90.4)	23.0 (16.4–31.2)	37.9 (28.9–47.3)
21	*Routier-adni*	49.2 (43.5–54.2)	94.6 (89.8–97.7)	11.5 (6.2–17.7)	36.9 (27.4–46.5)
22.5	*Eskildsen-FACEADNI2*	48.3 (43.2–53.4)	48.8 (40.5–57.4)	42.6 (33.9–51.3)	54.4 (45.5–64.0)
22.5	*Routier-train*	48.3 (42.9–53.4)	48.1 (39.8–56.9)	21.3 (14.8–29.0)	80.6 (72.2–87.3)
24.5	*Ledig-VOL*	47.7 (42.1–52.8)	66.7 (57.1–74.1)	36.9 (28.9–45.9)	36.9 (28.6–47.2)
24.5	*Eskildsen-ADNI2*	47.7 (42.1–52.8)	59.7 (51.2–68.4)	38.5 (29.9–47.3)	43.7 (33.7–53.8)
26	*Amoroso*	46.9 (41.5–52.3)	67.4 (58.5–75.2)	42.6 (33.6–51.1)	26.2 (18.3–35.4)
27	*Tangaro*	46.6 (41.0–51.4)	68.2 (60.2–76.5)	37.7 (29.2–46.3)	30.1 (21.7–39.0)
28	*Cárdenas-Peña*	39.0 (33.9–43.8)	50.4 (41.5–59.1)	28.7 (21.6–38.5)	36.9 (27.4–46.8)
29	*Smith*	32.2 (27.4–36.7)	48.1 (39.6–57.1)	20.5 (13.9–28.3)	26.2 (18.3–35.0)

ALL ($p = 0.09$), and *Franke* ($p = 0.06$). The TPFs had a large variability between the algorithms, showing that the different algorithms chose different priors for the classification. Appendix A lists the confusion matrices for all algorithms.

For 19 of the methods, output probabilities were submitted, enabling ROC-analysis. Table 6 and Fig. 2 show the overall AUC and the per-class AUCs ($AUC(c_i)$) for the algorithms ranked by AUC. The AUC ranged from 50.4% to 78.8%. This was better than random guessing for all algorithms except for one having an AUC of 50.4% (46.7%–54.6%). The two algorithms by Sørensen et al. (*Sørensen-equal*, *Sørensen-optimized*) had the highest AUC (78.8%), followed by the algorithm of *Abdulkadir* (AUC = 77.7%). Fig. 3 shows the per-class ROC curves for *Sørensen-equal*. For most algorithms, the per-class AUCs for CN (range: 54.1%–86.6%) and AD (range: 46.6%–89.2%) were higher than the overall AUC. Except for *Smith*, AUC_{MCI} (range: 50.0%–63.1%) was always smaller than the overall AUC.

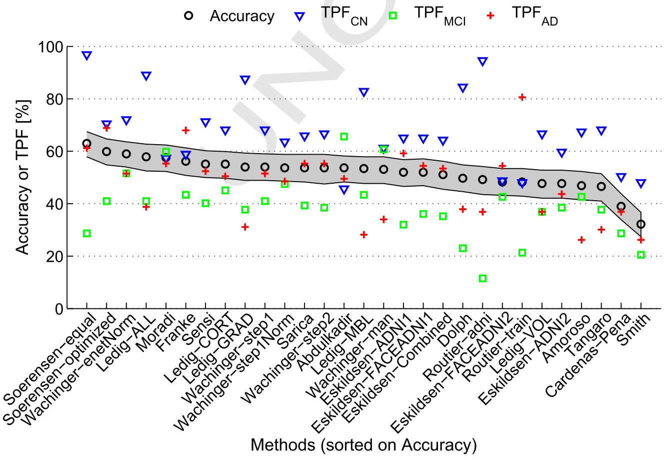


Fig. 1. Accuracy and TPFs on the test data for the participating algorithms. For the accuracy, the 95% confidence interval is shown in gray.

For the AD and CN classes, the evaluated algorithms obtained relatively high values for TPF and AUC. However, TPF and AUC for the MCI class were lower than those for the other classes, indicating that classification of MCI based on MRI is a difficult problem. This might be due to several factors including the heterogeneity of the MCI class and the use of the clinical diagnosis as reference standard (see Clinical applicability section).

The test data consisted of three subsets of data from three centers (Table 2). Fig. 4 shows how the performances of the algorithms varied between the subsets provided by different centers. The performances on the UP data set were mostly higher than those using all data, but the variation in performance across algorithms was rather high. Performances on the VUMC data were slightly better than those for all data, and performances on the EMC data were slightly worse than those for all data.

Feature extraction and classifiers

As shown in Table 4, the algorithms used a wide range of approaches. Out of the 29 methods, most methods included features based on volume (N = 19), 14 algorithms included features based on cortical thickness, 14 algorithms included features based on intensity (of which two algorithms used raw intensities and the rest more complex intensity relations), 9 algorithms included features based on shape, and 3 algorithms used voxel-based morphometry (VBM). Volume, cortical thickness, intensity and shape features were often combined. The combination of volume, cortical thickness and intensity was most often used (N = 8). We noted from Fig. 5 that the performance differences between the different feature extraction strategies were small, but in general we observed that the best performances were achieved with VBM and the combination of volume and cortical thickness with either shape, intensity or both. Also the classifiers

Please cite this article as: Bron, E.E., et al., Standardized evaluation of algorithms for computer-aided diagnosis of dementia based on structural MRI: The CADDementia challenge, NeuroImage (2015), http://dx.doi.org/10.1016/j.neuroimage.2015.01.048

Table 6
Area under the ROC-curve (AUC) on the test data for the participating algorithms that computed probabilistic outputs. CI = 95% confidence interval estimated with bootstrapping.

Rank	Algorithm	AUC [%] (CI)	AUC_{CN} [%] (CI)	AUC_{MCI} [%] (CI)	AUC_{AD} [%] (CI)
1.5	*Sørensen-equal*	78.8 (75.6–82.0)	86.3 (81.8–89.3)	63.1 (56.6–68.3)	87.5 (83.4–91.1)
1.5	*Sørensen-optimized*	78.8 (75.5–82.1)	86.3 (81.9–89.3)	62.7 (56.8–68.4)	86.7 (82.3–90.4)
3	*Abdulkadir*	77.7 (74.2–81.0)	85.6 (81.4–89.0)	59.9 (54.1–66.4)	86.7 (82.3–90.3)
4	*Wachinger-enetNorm*	77.0 (73.6–80.3)	83.3 (78.5–87.0)	59.4 (52.9–65.5)	88.2 (83.8–91.4)
5	*Ledig-ALL*	76.7 (73.6–79.8)	86.6 (82.7–89.8)	59.7 (53.3–65.1)	84.9 (79.7–88.7)
6	*Ledig-GRAD*	75.4 (72.4–78.6)	85.6 (81.5–88.9)	60.3 (53.9–66.5)	81.7 (76.3–86.1)
7	*Ledig-MBL*	75.2 (72.0–78.1)	82.5 (77.8–86.0)	57.3 (50.9–63.6)	86.4 (81.4–89.9)
8	*Wachinger-step1*	74.6 (70.8–78.1)	79.1 (73.5–83.1)	55.0 (48.5–61.4)	89.2 (85.3–92.3)
9.5	*Wachinger-step1Norm*	74.3 (70.5–77.9)	79.3 (74.1–83.5)	55.5 (48.5–61.6)	87.7 (83.7–91.1)
9.5	*Wachinger-man*	74.3 (70.9–77.9)	80.6 (75.7–84.9)	56.3 (49.7–63.0)	86.1 (81.7–90.0)
11	*Sensi*	73.8 (70.2–77.5)	81.7 (77.1–85.8)	55.0 (48.8–61.0)	83.9 (78.8–87.7)
12	*Ledig-CORT*	73.7 (69.9–77.2)	79.6 (75.0–84.2)	58.9 (52.9–64.9)	82.4 (76.7–87.3)
13	*Wachinger-step2*	72.7 (68.9–76.4)	79.3 (74.0–83.5)	51.9 (45.3–58.7)	86.5 (81.9–90.3)
14	*Ledig-VOL*	68.4 (64.5–72.5)	75.7 (70.3–81.0)	50.1 (44.1–56.4)	79.0 (73.3–83.5)
15	*Amoroso*	67.2 (63.3–71.3)	73.4 (67.8–78.7)	56.0 (49.7–61.9)	72.3 (66.2–77.5)
16	*Tangaro*	67.1 (63.2–71.0)	73.1 (67.8–78.0)	52.6 (45.9–58.6)	75.8 (70.2–80.6)
17	*Dolph*	63.0 (59.6–67.2)	66.2 (61.3–70.3)	55.4 (50.0–60.0)	65.8 (60.6–71.3)
18	*Cárdenas-Peña*	55.9 (51.2–59.9)	57.8 (51.6–63.4)	50.0 (43.9–57.1)	59.8 (53.5–65.7)
19	*Smith*	50.4 (46.7–54.6)	54.1 (48.0–60.0)	50.6 (45.0–57.1)	46.6 (40.0–53.6)

differed between the algorithms: 14 algorithms used regression, 7 algorithms used an SVM classifier, 6 used a random forest classifier, 2 used linear discriminant analysis (LDA) and 1 used a neural network for classification. Performance differences between the different classifiers seemed to be small. It should be noted, however, that one should be careful in drawing conclusions based on Table 4 or Fig. 5, as there are multiple differences between the algorithms.

Eight teams incorporated age effects in their algorithms, either by explicitly including age in the model (Franke and Gaser, 2014; Sarica et al., 2014; Smith et al., 2014) or by eliminating age effects using age-dependent normalization (Sørensen et al., 2014) or regression (Abdulkadir et al., 2014; Eskildsen et al., 2014; Moradi et al., 2014; Wachinger et al., 2014a). Three teams used the same strategy to correct for sex (Abdulkadir et al., 2014; Eskildsen et al., 2014; Sarica et al., 2014), two teams trained separate models for males and females (Franke and Gaser, 2014; Smith et al., 2014).

Training data

Most algorithms, except for *Dolph*, were trained on more training data than only the 30 provided data sets. Mainly data from ADNI and AIBL were used. Fig. 6 shows the relationship between the number of training data sets and the test set performance. Most algorithms used 600–800 data sets for training.

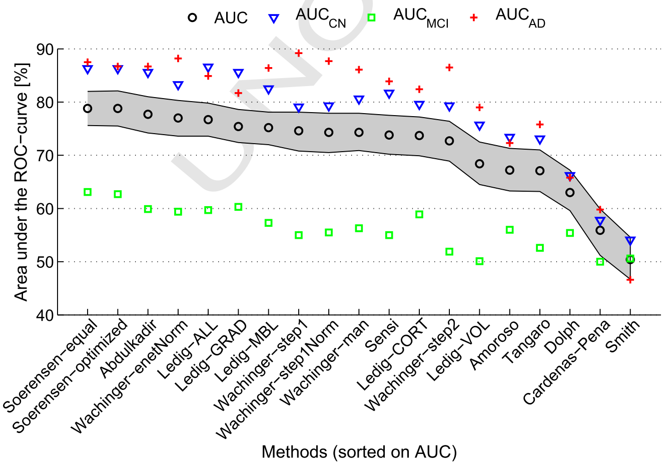


Fig. 2. Area under the ROC-curve (AUC) on the test data for the participating algorithms. For total AUC, the 95% confidence interval is shown in gray.

Fig. 7 shows the relationship between the accuracy of the algorithms on the test set and the accuracy on the 30 provided training data sets as reported in the workshop papers. The figure shows that almost all algorithms overestimated accuracy on the training set. However, some of the methods explicitly trained on the 30 provided data sets to ensure optimal performance on the test set. It should be noted that different strategies were used to evaluate the training set accuracy, i.e., train-test evaluation or cross-validation.

Discussion

Evaluation framework

Although the literature on computer-aided diagnosis of dementia has shown promising results, thorough validation of these algorithms for clinical use has rarely been performed. To enable proper validation of the algorithms, we addressed the following factors in our evaluation framework: comparability, generalizability and clinical applicability.

Comparability

Comparison of different state-of-the-art algorithms is difficult, as most studies use different evaluation data sets, validation strategies and performance measures. According to the literature, little has been

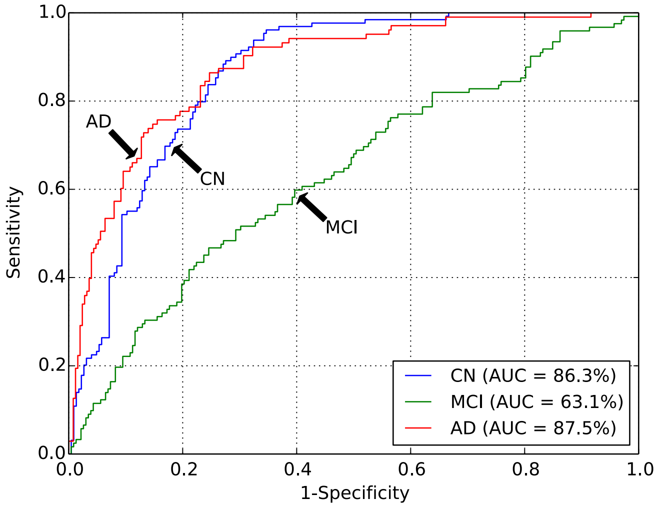
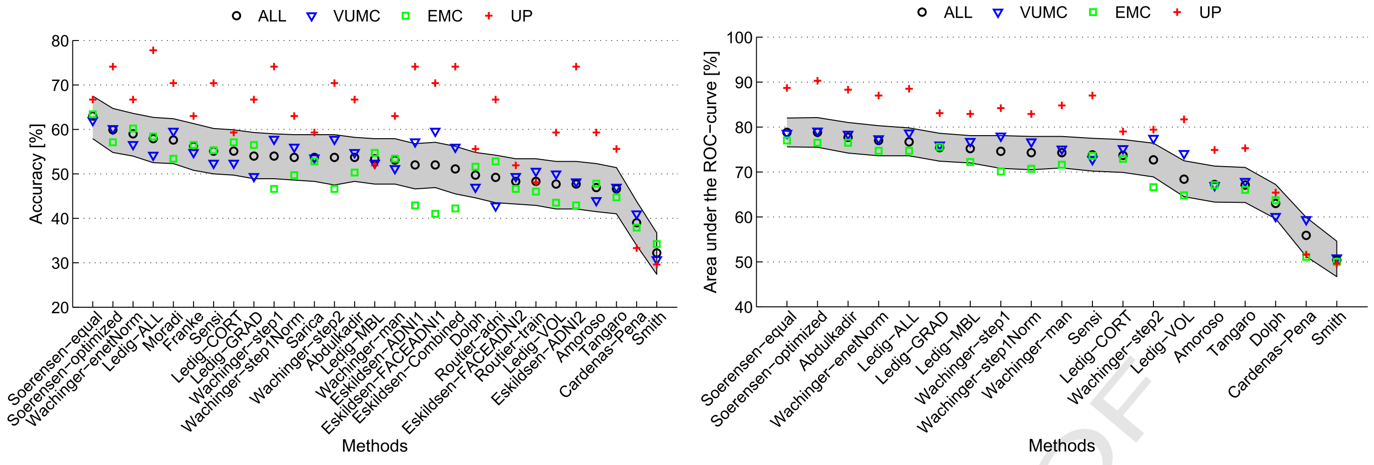


Fig. 3. The receiver-operating-characteristic (ROC) curve on all test data for the best performing algorithm: *Sørensen-equal*.

Please cite this article as: Bron, E.E., et al., Standardized evaluation of algorithms for computer-aided diagnosis of dementia based on structural MRI: The CADDementia challenge, NeuroImage (2015), http://dx.doi.org/10.1016/j.neuroimage.2015.01.048



(a) Accuracy

(b) AUC

Fig. 4. Accuracy (a) and area under the ROC-curve (AUC) (b) on the test data for the participating algorithms on all data (N = 354) and on the three subsets of test data from different centers: VUMC (N = 166), EMC (N = 161), UP (N = 27). For accuracy and AUC on all data, the 95% confidence interval is shown in gray.

done in comparing different algorithms using the same data and methodology. We found two studies that compared multiple algorithms (Cuingnet et al., 2011; Sabuncu and Konukoglu, 2014), of which the work of Cuingnet et al. (2011) does not allow addition of new methods to the comparison. For our evaluation framework, we aimed to increase comparability of the evaluated algorithms by making the testing data set and the validation scripts publicly available. Effort was made to compose a large multi-center data set and to define good evaluation criteria for multi-class classification. One of the main advantages of this evaluation framework is that it can be used by every researcher: anyone who developed a new algorithm can download the data and submit results via our web-based framework[13]. Both established and state-of-the-art algorithms can be evaluated and compared to algorithms evaluated by others. The framework remains open for new submissions.

Since the main question that we aimed to address with this framework is how well the current state-of-the-art methods would perform in clinical practice, we specifically chose to use few constraints for the participating methods. Therefore, the framework allows to compare algorithms performing the full analysis, from image to diagnosis. This introduces a lot of variation in the participating algorithms. Participants had a lot of freedom in their choices for the training data and the methods for image processing and classification. Therefore, in discussing the methods we were not able to completely explain the performance differences between methods in all cases. For example, a very good method that uses a small amount of training data may have the same performance as another method that is worse but uses more training data. With the chosen set-up, it is also not possible to assess which part of the algorithm led to the increase in performance. These include a multitude of aspects, such as feature extraction, feature selection, and classification.

At present, a similar challenge is running: the Alzheimer's Disease Big Data (ADBD) DREAM challenge #1[14], of which sub-challenge 3 is similar to the work presented in this paper. In the ADBD DREAM challenge, participants are asked to build a predictive model for MMSE and diagnosis based on T1w MRI data and other variables (i.e., age at baseline, years of education, sex, APOE4 genotype, imputed genotypes). One of the differences with our challenge is that the ADBD DREAM challenge supplies a fixed training set from the ADNI database, instead of leaving this open to the participants. Two test sets, both consisting of 107 subjects from the AddNeuroMed database (Lovestone et al., 2009) are provided. The ADBD DREAM challenge generally made the same choices for their evaluation framework, as they use the same diagnostic groups and reference standard. Preliminary results for the ADBD DREAM challenge are available from their web site. The best predictive model for MMSE yielded a Pearson correlation of 0.602, and the best model for diagnosis yielded an accuracy of 60.2%. The algorithm that was best ranked on average used Gaussian process regression with 20 image features, APOE4 and education (Fan and Guan, 2014).

Generalizability

For new methods, it is important to know how they would generalize to a new, clinically representative data set. Often cross-validation is used to validate the performance of machine learning algorithms (Falahati et al., 2014). Although cross-validation is very useful, especially in the situation when not many scans are available, it optimizes performance on a specific population and can therefore overestimate performance on the general population (Adaszewski et al., 2013). In addition, algorithms are often tuned to specific cohorts which limit their generalizability (Adaszewski et al., 2013). When generalizing an algorithm to other data, variability in the data acquisition protocol, the population or the reference standard can be problematic and can decrease performance (Sabuncu and Konukoglu, 2014). To evaluate generalizability of the algorithms, which is certainly required for clinical implementation, we used a large, new and unseen test set in this work. This data set consisted of scans acquired with GE (n = 354) and Siemens (n = 30) scanners, so we do not have information on the performance of the algorithms on data from other scanners. However, the data set had some differences in scanning parameters, which allows evaluation of the generalizability of the algorithms to different scanning protocols. The diagnostic labels of the test set were blinded to the authors of the algorithms, which is different from the benchmark papers by Cuingnet et al. (2011) and Sabuncu and Konukoglu (2014). The importance of an independent test is also confirmed by Fig. 7, which shows that all algorithms overestimated the performance by cross-validating or tuning on the training set.

Another factor providing insight into the generalizability of the performance results was the size of the test set. The test set was quite large, consisting of 354 subjects. Not many other studies used an unseen test set. For studies using cross-validation, usually 500–800 data sets from the ADNI database are used (Cuingnet et al., 2011; Falahati et al., 2014; Sabuncu and Konukoglu, 2014). The ADBD DREAM challenge uses an unseen test set, but much smaller than the one used here (107 subjects).

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

[14] http://www.synapse.org/#!Synapse:syn2290704/.

Please cite this article as: Bron, E.E., et al., Standardized evaluation of algorithms for computer-aided diagnosis of dementia based on structural MRI: The CADDementia challenge, NeuroImage (2015), http://dx.doi.org/10.1016/j.neuroimage.2015.01.048

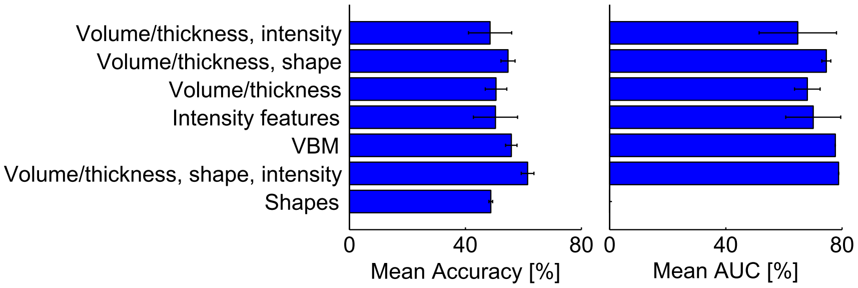


Fig. 5. Mean accuracy and area under the ROC-curve (AUC) on the test data for the different types of features used by the algorithms. The error bars show the standard deviation.

Clinical applicability

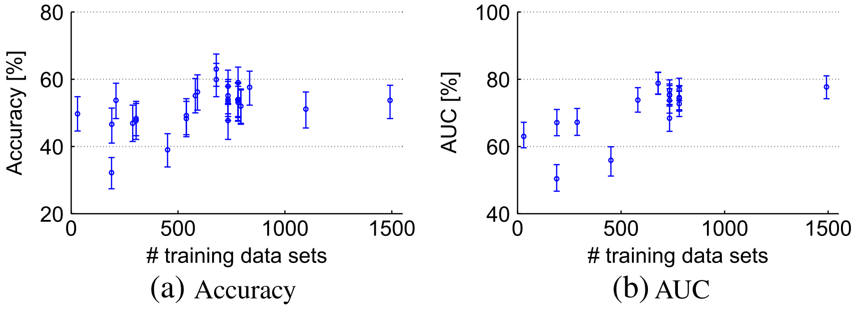
For this evaluation framework, the decision was made to split our multi-center data set into a small (n = 30) training set and a large test set. This choice resembles a clinical setting, where in a certain hospital only a small training data set is available. On the other hand, a lot of training data are available from publicly available databases like the ADNI and AIBL, which can be used for training the algorithms.

As reference standard for evaluation of the algorithms, the current clinical diagnosis criteria for AD (McKhann et al., 2011) and MCI (Petersen, 2004) were used, which is a common practice in studies of computer-aided diagnosis methods (Cuingnet et al., 2011; Klöppel et al., 2008; Falahati et al., 2014; Davatzikos et al., 2008a; Duchesne et al., 2008; Fan et al., 2008a,b; Gray et al., 2013; Koikkalainen et al., 2012; Magnin et al., 2009; Vemuri et al., 2008; Wolz et al., 2011). Ground truth diagnosis of dementia can only be assessed using autopsy and is therefore only rarely available. Of the previously mentioned papers, only one paper included one group of 20 AD patients with an autopsy confirmed diagnosis (Klöppel et al., 2008). Amyloid imaging (Klunk et al., 2004) has also proven to be a good biomarker for AD, as subjects with positive amyloid showed to have a more rapid disease progression (Jack et al., 2010). However, availability of these data is also very limited. The limitation of using clinical diagnosis as the ground truth is that it may be incorrect. In the literature, the reported accuracies of the clinical diagnosis of AD, based on the old criteria (McKhann et al., 1984), compared to postmortem neuropathological gold standard diagnosis were in the range of 70–90% (Mattila et al., 2012; Lim et al., 1999; Petrovitch et al., 2001; Kazee et al., 1993). Although the clinical diagnosis has limitations, we believe that it is the best available reference standard. One should also note that this challenge does not aim to assess the diagnostic accuracy of structural MRI, as MRI itself is also included in the criteria for clinical diagnosis. Instead, we focus on comparing computer-aided diagnosis algorithms on an unseen blinded test set with standardized evaluation methods using the clinical diagnosis as the best available reference standard.

This work interprets the differentiation of patients with AD, MCI and controls as a multi-class classification problem. This might not be optimal as there is an ordering of the classes, i.e., classification of an AD patient as an MCI patient might be less bad than classifying as a healthy person. However, addressing only binary problems, such as AD/CN classification, does not reflect the clinical diagnosis making and results in a too optimistic performance estimate. Because the current clinical diagnosis uses the three classes, we chose to focus on multi-class classification in this challenge and did not use the ordering in the evaluation.

According to the criteria of Petersen (2004) and similar to ADNI, only MCI patients with memory complaints, amnestic MCIs, were included in the data set. For classification, all MCI patients were considered to be a single group which is according to current clinical practice Petersen (2004). This is debatable, since MCI patients are known to be a clinically heterogeneous group with different patterns of brain atrophy (Misra et al., 2009), of which some cases will not progress to AD. From this point of view, it can be questioned whether MCI is a diagnostic entity or whether MCI describes a stage on a continuum from cognitively normal to AD. If MCI is actually an intermediate between the two other classes, the AD/CN border in three-class classification would be also subject to discussion. Although the usage of the MCI definition is advised for diagnosis in clinical practice (Petersen, 2004), the borders between AD/MCI and MCI/CN based on diagnostic criteria can be unclear. Because of those unclear borders and the heterogeneity in the MCI class, classification accuracies are expected to be reduced. The results of the evaluated algorithms confirmed that distinguishing MCI from AD and CN is difficult. The AUC for all algorithms was the lowest for the MCI class and in most cases also TPF was the lowest for MCI. Despite these limitations, the same choices for the reference standard, classification, and the MCI group were made in the ADBD DREAM challenge. Moreover, since MCI is still used as diagnostic label in current clinical practice, having an objective and automated algorithm that makes such diagnosis based on structural MRI, would already be useful, for example, as a second opinion.

For facilitating clinical implementation of the algorithms, it would be a great benefit to make the evaluated algorithms publicly available for enabling validation on other data without the need for reimplementation. In our evaluation framework, this is not yet possible. Instead, in our



(a) Accuracy (b) AUC

Fig. 6. The number of training data sets used plotted against the test set performance of every algorithm: (a) accuracy, (b) area under the ROC-curve (AUC). The error bars show the 95% confidence interval.

Please cite this article as: Bron, E.E., et al., Standardized evaluation of algorithms for computer-aided diagnosis of dementia based on structural MRI: The CADDementia challenge, NeuroImage (2015), http://dx.doi.org/10.1016/j.neuroimage.2015.01.048

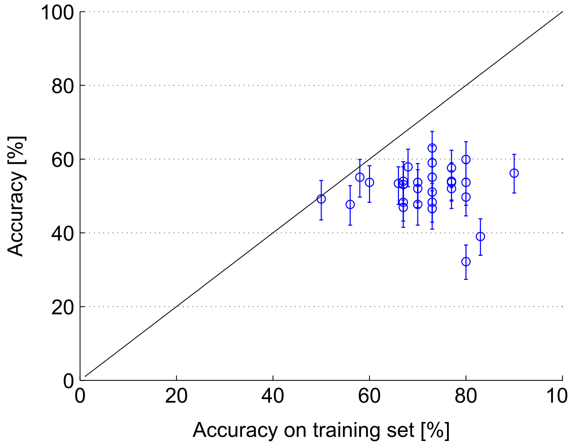


Fig. 7. Accuracies for each algorithm estimated on the provided training data plotted against the final accuracy. The error bars show the 95% confidence interval on the test data. The black line ($y = x$) indicates the expected relationship.

framework, all teams were encouraged to make a step-by-step implementation guide[15] to make it possible to run the submitted algorithms on other data sets.

Evaluated algorithms and results

The best performing algorithm (*Sørensen-equal*: accuracy = 63.0%, AUC = 78.8%) was based on a combination of features and used a simple linear classifier (LDA). Also, regarding the other top-ranked algorithms, the best performances were achieved by algorithms that incorporated features describing different properties of the scans. Although the performance differences between the different feature extraction strategies were small, algorithms that used shape or intensity features in addition to regional volumes and thickness performed slightly better than algorithms solely based on shape features or on volume features. The VBM-based methods also performed well. Different multivariate analysis techniques were used by the algorithms, mainly regression, SVM, and random forest classifiers. No trend in the best performing type of classifier could be found.

Since hardly any results for three-class classification have been reported, we cannot compare with representative results from the literature. The TPFs and AUCs for the AD and CN classes in this work are a bit lower than those reported previously for AD/CN classification (Falahati et al., 2014), but we expect that this is mainly due to the additional MCI class in the classification and its heterogeneity. The ADBD DREAM challenge also evaluated three-class classification, and it reported performances similar to those of this study (see Comparability section).

The methods *Sørensen-equal* and *Sørensen-optimized* were ranked highest both based on accuracy and AUC. In general, the rankings by the two performance measures were similar, but there were some exceptions. *Abdulkadir*, for example, ranked much higher based on AUC (rank = 3) than on accuracy (rank = 12.5), which means that this method was capable of distinguishing the classes with high sensitivity and specificity at different cut-off points. However, for measuring the accuracy, not the optimal cut-off point was chosen by the classifier. The accuracy of this method could be improved by optimizing the class priors used by the classifier. For classification, it is generally assumed that the training data and its class priors are representative for the test data. Depending on the class distributions of the training data used, this assumption on class priors might not always have been justified. On the other hand, it is difficult to correct for differences in class priors, as the distribution of the test set is often unknown. Of the participating teams, two specifically took the issue of class priors into account. Eskildsen et al. removed the class unbalance of the training set using a resampling technique (Eskildsen et al., 2014; Chawla et al., 2002). Sørensen et al. experimented with two sets of class priors: equal class priors and class priors optimized on the 30 training subjects (Sørensen et al., 2014). However, for most algorithms accuracy and AUC were similar, indicating that reasonable assumptions on the class priors were made.

The provided data set consisted of structural MRI scans from three centers. We noticed a small performance difference between the three subsets. The performance on the UP subset was the highest, but this might be explained by chance given the small size of the UP data set (n = 27 in the test set, n = 3 in the training set) and a slight selection bias towards more clinically clear-cut cases. Between the two other subsets, a minor performance difference could be noted. The performance differences might be caused by slight differences in inclusion criteria, used scanners and scanning protocols between the centers, emphasizing the importance of a multi-center test set.

The size of the training set is known to have a large influence on the performance of the classifier (Falahati et al., 2014). Although this study does not provide enough information to draw a valid conclusion, as we evaluated only 29 algorithms with the majority of training sets consisting of 600–800 subjects, we see a slight positive relation between the number of training data sets and the test set performance.

The mean age of AD patients in the used data set was 66.1 ± 5.2 years, whereas the age for AD patients in the ADNI cohorts that were used by many algorithms for training was about 10 years higher (Abdulkadir et al., 2014; Amoroso et al., 2014; Eskildsen et al., 2014; Ledig et al., 2014; Sarica et al., 2014; Sensi et al., 2014; Sørensen et al., 2014;Wachinger et al., 2014a). Although the same diagnosis criteria were used in both cohorts, this age difference is most probably due to selection bias. The used dataset consists of clinical data representing the outpatient clinic population, whereas ADNI consists of research data. For clinical practice, MRI may be used more conservatively. In addition, there is a referral bias towards younger patients because the VUMC and the EMC are tertiary centers specialized in presenile dementia. This age difference between training and test data might have had a negative effect on the performances found in this study. To take this into account, eight of the 15 teams incorporated age effects in their algorithms.

Recommendations for future work

This challenge provided insight on the best strategies for computer-aided diagnosis of dementia and on the performance of such algorithms on an independent clinically representative data set. However, for this challenge, specific choices for the evaluation framework were made. Therefore, for clinical implementation of such algorithms, more validation studies that explore variations of this challenge are necessary.

A limitation of this challenge is that the clinical diagnosis is used as reference standard. For the clinical diagnosis, MCI is used as a diagnostic entity; it could however be questioned whether this can exist as separate diagnosis next to AD. In addition, the accuracy of the clinical diagnosis is limited, but data sets with better reference standards are scarce. The best reference standard is the postmortem diagnosis based on pathology, which is the ground truth for AD diagnosis. A good alternative would be a reference standard based on the clinical diagnosis including amyloid biomarkers or a long-term follow-up. For a validation study, we strongly recommend to have an independent test set with blinded diagnostic labels to promote generalizability.

In this challenge, classification was based on structural MRI using subject age and sex as the only additional information. For a future challenge in which ground truth diagnosis is used for reference, it would be very interesting to use all available clinical data in addition to structural MRI as input for the computer-aided diagnosis algorithms. For the current challenge, this was not yet useful as the reference standard was

[15] http://caddementia.grand-challenge.org/wiki.

Please cite this article as: Bron, E.E., et al., Standardized evaluation of algorithms for computer-aided diagnosis of dementia based on structural MRI: The CADDementia challenge, NeuroImage (2015), http://dx.doi.org/10.1016/j.neuroimage.2015.01.048

based directly on these clinical data. For structural MRI, this is not a problem as it is only used qualitatively in clinical diagnosis making.

For the current work, we adopted hardly any constraints resulting in a wide range of participating algorithms. To aid the understanding of the influence of certain methodological choices on the algorithm performance, new projects could decide to focus on comparing specific elements of the algorithms.

We cannot be sure that the included algorithms are the best currently available. Although this challenge was broadly advertised, quite some effort from participants was required which may have kept some researchers from participating. Of the teams that submitted a proposal, two thirds did not participate in the challenge, possibly due to lack of time or resources. To reach a wider audience in future challenges, organizers could reduce the effort required from participants, for example by providing precomputed features.

Another interesting problem to address in a future challenge is that of differential diagnosis of AD and other types of dementia (e.g., frontotemporal dementia (Du et al., 2006; Davatzikos et al., 2008b; Raamana et al., 2014) or Lewy body dementia (Lebedev et al., 2013)). In addition, instead of evaluating diagnostic algorithms, evaluation of prognostic algorithms would be very useful. Future challenges could therefore evaluate the classification of MCI patients that convert to AD and MCI patients that do not convert to AD within a certain time period.

Lastly, new projects could request their participants to make their algorithms publicly available to facilitate clinical implementation of the algorithms for computer-aided diagnosis.

Conclusion

We presented a framework for the comparison of algorithms for computer-aided diagnosis of AD and MCI using structural MRI data and used it to compare 29 algorithms submitted by 15 research teams. The framework defines evaluation criteria and provides a previously unseen multi-center data set with the diagnoses blinded to the authors of the algorithms. The results of this framework therefore present a fair comparison of algorithms for multi-class classification of AD, MCI and CN. The best algorithm, developed by Sørensen et al., yielded an accuracy of 63% and an AUC of 78.8%. Although the performance of the algorithms was influenced by many factors, we noted that the best performance was generally achieved by methods that used a combination of features.

The evaluation framework remains open for new submissions to be added to the ranking. We refer interested readers to the web site http://caddementia.grand-challenge.org, where instructions for participation can be found.

We believe that public large-scale validation studies, such as this work, are an important step towards the introduction of high-potential algorithms for computer-aided diagnosis of dementia into clinical practice.

Acknowledgments

We would like to acknowledge the team members of the participating algorithms: Jessica Peter, Thomas Brox, Stefan Klöppel, Rosangela Errico, German Castellanos-Dominguez, Manar D Samad, D Louis Collins, Antonios Makropoulos, Rolf A Heckemann, Daniel Rueckert, Heikki Huttunen, Pietro Gori, Ana B Graciano Fouquier, Sophie Lecomte, Olivier Colliot, Mario Cannataro, J Douglas Saddy, Luca Rei, Gianluca Gemme, Paolo Bosco, Danica V Greetham, Peter Grindrod, Akshay Pai, Cecilie Anker, Ioana Balas, Martin Lillholm, Christian Igel, Rosalia Maglietta, Andrea Tateo, Kayhan Batmanghelich and Polina Golland.

nWe would like to thank Ronald van Schijndel (VU University Medical Center, Neuroscience Campus Amsterdam, The Netherlands) for his assistance in making the VUMC data available. In addition, many thanks to Sjoerd Kerkstra (Diagostic Image Analysis Group, Department of Radiology, Radboud University Medical Center Nijmegen, The Netherlands) for his assistance in building the web-based evaluation framework on http://grand-challenge.org.

This work was funded by an Erasmus MC grant on Advanced MR Q8 neuroimaging in presenile dementia.

Research of the VUMC Alzheimer Center is part of the neurodegeneration research program of the Neuroscience Campus Amsterdam. The VUMC Alzheimer Center is supported by Alzheimer Nederland and Stichting VUmc fonds. The clinical database structure was developed with funding from Stichting Dioraphte.

Data collection and sharing of the data used for training of most of the algorithms was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012).

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

Data used for training of some of the algorithms was obtained from the Australian Imaging Biomarkers and Lifestyle flagship study of aging (AIBL) funded by the Commonwealth Scientific and Industrial Research Organisation (CSIRO).

Q9

Appendix A. Confusion matrices of the algorithms

Sørensen-equal		True class			*Wachinger-step1Norm*		True class			*Routier-adni*		True class		
		CN	MCI	AD			CN	MCI	AD			CN	MCI	AD
Hypothesized class	CN	125	64	15	Hypothesized class	CN	82	49	7	Hypothesized class	CN	122	87	42
	MCI	3	35	25		MCI	47	58	46		MCI	7	14	23
	AD	1	23	63		AD	0	15	50		AD	0	21	38
Sørensen-optimized		True class			*Sarica*		True class			*Eskildsen-FACEADNI2*		True class		
		CN	MCI	AD			CN	MCI	AD			CN	MCI	AD
Hypothesized class	CN	91	37	5	Hypothesized class	CN	85	43	11	Hypothesized class	CN	63	31	6
	MCI	33	50	27		MCI	41	48	34		MCI	56	52	41
	AD	5	35	71		AD	3	29	57		AD	10	39	56

(continued on next page)

Please cite this article as: Bron, E.E., et al., Standardized evaluation of algorithms for computer-aided diagnosis of dementia based on structural MRI: The CADDementia challenge, NeuroImage (2015), http://dx.doi.org/10.1016/j.neuroimage.2015.01.048

Appendix A (*continued*)

Wachinger-enetNorm		True class			*Wachinger-step2*		True class			*Routier-train*		True class		
		CN	MCI	AD			CN	MCI	AD			CN	MCI	AD
Hypothesized class	CN	93	44	6	Hypothesized class	CN	86	51	4	Hypothesized class	CN	62	17	2
	MCI	36	63	44		MCI	41	47	42		MCI	42	26	18
	AD	0	15	53		AD	2	24	57		AD	25	79	83
Ledig-ALL		True class			*Abdulkadir*		True class			*Ledig-VOL*		True class		
		CN	MCI	AD			CN	MCI	AD			CN	MCI	AD
Hypothesized class	CN	115	57	16	Hypothesized class	CN	59	19	2	Hypothesized class	CN	86	53	11
	MCI	14	50	47		MCI	69	80	50		MCI	41	45	54
	AD	0	15	40		AD	1	23	51		AD	2	24	38
Moradi		True class			*Ledig-MBL*		True class			*Eskildsen-ADNI2*		True class		
		CN	MCI	AD			CN	MCI	AD			CN	MCI	AD
Hypothesized class	CN	74	30	2	Hypothesized class	CN	107	66	13	Hypothesized class	CN	77	36	7
	MCI	52	73	44		MCI	20	53	61		MCI	49	47	51
	AD	3	19	57		AD	2	3	29		AD	3	39	45
Franke		True class			*Wachinger-man*		True class			*Amoroso*		True class		
		CN	MCI	AD			CN	MCI	AD			CN	MCI	AD
Hypothesized class	CN	76	48	12	Hypothesized class	CN	79	39	5	Hypothesized class	CN	87	58	32
	MCI	44	53	21		MCI	50	74	63		MCI	36	52	44
	AD	9	21	70		AD	0	9	35		AD	6	12	27
Sensi		True class			*Eskildsen-ADNI1*		True class			*Tangaro*		True class		
		CN	MCI	AD			CN	MCI	AD			CN	MCI	AD
Hypothesized class	CN	92	45	9	Hypothesized class	CN	84	30	7	Hypothesized class	CN	88	62	18
	MCI	36	49	40		MCI	33	39	35		MCI	31	46	54
	AD	1	28	54		AD	12	53	61		AD	10	14	31
Ledig-CORT		True class			*Eskildsen-FACEADNI1*		True class			*Cárdenas-Peña*		True class		
		CN	MCI	AD			CN	MCI	AD			CN	MCI	AD
Hypothesized class	CN	88	49	18	Hypothesized class	CN	84	29	8	Hypothesized class	CN	65	51	36
	MCI	32	55	33		MCI	38	44	39		MCI	30	35	29
	AD	9	18	52		AD	7	49	56		AD	34	36	38
Ledig-GRAD		True class			*Eskildsen-Combined*		True class			*Smith*		True class		
		CN	MCI	AD			CN	MCI	AD			CN	MCI	AD
Hypothesized class	CN	113	59	19	Hypothesized class	CN	83	33	7	Hypothesized class	CN	62	51	44
	MCI	15	46	52		MCI	39	43	41		MCI	39	25	32
	AD	1	17	32		AD	7	46	55		AD	28	46	27
Wachinger-step1		True class			*Dolph*		True class							
		CN	MCI	AD			CN	MCI	AD					
Hypothesized class	CN	88	57	7	Hypothesized class	CN	109	73	46					
	MCI	40	50	43		MCI	14	28	18					
	AD	1	15	53		AD	6	21	39					

References

Abdulkadir, A., Peter, J., Brox, T., Ronneberger, O., Klöppel, S., 2014. Voxel-based multi-class classification of AD, MCI, and elderly controls: blind evaluation on an independent test set. Proc MICCAI Workshop Challenge on Computer-Aided Diagnosis of Dementia Based on Structural MRI Data, pp. 8–15.

Adaszewski, S., Dukart, J., Kherif, F., Frackowiak, R., Draganski, B., 2013. How early can we predict Alzheimer's disease using computational anatomy? Neurobiol. Aging 34 (12), 2815–2826.

Albert, M.S., DeKosky, S.T., Dickson, D., Dubois, B., Feldman, H.H., Fox, N.C., Gamst, A., Holtzman, D.M., Jagust, W.J., Petersen, R.C., Snyder, P.J., Carrillo, M.C., Thies, B., Phelps, C.H., 2011. The diagnosis of mild cognitive impairment due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimers Dement. 7 (3), 270–279.

Alzheimer's Association, 2014. 2014 Alzheimer's disease facts and figures. Alzheimers Dement. 10 (2), e47–e92.

Amoroso, N., Errico, R., Bellotti, R., 2014. PRISMA-CAD: Fully automated method for Computer-Aided Diagnosis of Dementia based on structural MRI data. Proc MICCAI Workshop Challenge on Computer-Aided Diagnosis of Dementia Based on Structural MRI Data, pp. 16–23.

Binnewijzend, M.A.A., Kuijer, J., Benedictus, M.R., van der Flier, W.M., Wink, A.M., Wattjes, M.P., van Berckel, J., Scheltens, P., Barkhof, F., 2013. Cerebral blood flow measured with 3D pseudocontinuous arterial spin labeling MR imaging in Alzheimer disease and mild cognitive impairment: a marker for disease severity. Radiology 267 (1), 221–230.

Bron, E.E., Smits, M., van Swieten, J.C., Niessen, J., Klein, S., 2014. Proc MICCAI Workshop Challenge on Computer-Aided Diagnosis of Dementia Based on Structural MRI Data.

Cárdenas-Peña, D., Álvarez Meza, A., Castellanos-Dominguez, G., 2014. CADDementia based on structural MRI using supervised kernel-based representations. Proc MICCAI Workshop Challenge on Computer-Aided Diagnosis of Dementia Based on Structural MRI Data, pp. 24–30.

Chawla, N., Bowyer, K., Hall, L.O., Kegelmeyer, W.P., 2002. SMOTE: synthetic minority over-sampling technique. J. Artif. Intell. Res. 16, 321–357.

Cuingnet, R., Gerardin, E., Tessieras, J., Auzias, G., Lehéricy, S., Habert, J., Chupin, M., Benali, J., Colliot, O., 2011. 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.

Davatzikos, C., Resnick, S.M., Wu, X., Parmpi, P., Clark, C.M., 2008a. Individual patient diagnosis of AD and FTD via high-dimensional pattern classification of MRI. Neuroimage 41 (4), 1220–1227.

Davatzikos, C., Fan, Y., Wu, X., Shen, D., Resnick, S.M., 2008b. Detection of prodromal Alzheimer's disease via pattern classification of magnetic resonance imaging. Neurobiol. Aging 29 (4), 514–523.

Dietterich, T., 1996. Statistical tests for comparing supervised classification learning algorithms. Oregon State University Technical Report. 1, pp. 1–24.

Dolph, C.V., Samad, M.D., Iftekharuddin, K.M., 2014. Classification of Alzheimer's disease using structural MRI. Proc MICCAI Workshop Challenge on Computer-Aided Diagnosis of Dementia Based on Structural MRI Data, pp. 31–37.

Du, A., Jahng, G., Hayasaka, S., Kramer, J., 2006. Hypoperfusion in frontotemporal dementia and Alzheimer disease by arterial spin labeling MRI. Neurology 67 (7), 1215–1220.

Duchesne, S., Caroli, A., Geroldi, C., Barillot, C., Frisoni, G.B., Collins, D.L., 2008. MRI-based automated computer classification of probable AD versus normal controls. IEEE Trans. Med. Imaging 27 (4), 509–520.

Durrleman, S., Prastawa, M., Charon, N., Korenberg, J.R., Joshi, S., Gerig, G., Trouvé, A., 2014. Morphometry of anatomical shape complexes with dense deformations and sparse parameters. Neuroimage 101, 35–49.

Ellis, K.A., Bush, A.I., Darby, D., De Fazio, D., Foster, J., Hudson, P., Lautenschlager, N.T., Lenzo, J., Martins, R.N., Maruff, J., Masters, C., Milner, J., Pike, K., Rowe, C., Savage, G., Szoeke, C., Taddei, K., Villemagne, V., Woodward, M., Ames, D., 2009. 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. Int. Psychogeriatr. 21 (4), 672–687.

Eskildsen, S.F., Coupé, P., Fonov, V., Collins, D.L., 2014. Detecting Alzheimer's disease by morphological MRI using hippocampal grading and cortical thickness. Proc MICCAI Workshop Challenge on Computer-Aided Diagnosis of Dementia Based on Structural MRI Data, pp. 38–47.

Eskildsen, S.F., Coupé, P., Fonov, V.S., Pruessner, J.C., Collins, D.L., 2015. Structural imaging biomarkers of Alzheimer's disease: predicting disease progression. Neurobiol. Aging 36 (Suppl. 1), S23–S31.

Falahati, F., Westman, E., Simmons, A., 2014. Multivariate data analysis and machine learning in Alzheimer's disease with a focus on structural magnetic resonance imaging. J. Alzheimers Dis. 41 (3), 685–708.

Fan, Z., Guan, Y., 2014. GuanLab — Alzheimer's disease prediction. Alzheimers Disease Big Data DREAM Challenge.

Please cite this article as: Bron, E.E., et al., Standardized evaluation of algorithms for computer-aided diagnosis of dementia based on structural MRI: The CADDementia challenge, NeuroImage (2015), http://dx.doi.org/10.1016/j.neuroimage.2015.01.048

Fan, Y., Batmanghelich, N., Clark, C.M., Davatzikos, C., 2008a. Spatial patterns of brain atrophy in MCI patients, identified via high-dimensional pattern classification, predict subsequent cognitive decline. Neuroimage 39 (4), 1731–1743.

Fan, Y., Resnick, S.M., Wu, X., Davatzikos, C., 2008b. Structural and functional biomarkers of prodromal Alzheimer's disease: a high-dimensional pattern classification study. Neuroimage 41 (2), 277–285.

Fawcett, T., 2006. An introduction to ROC analysis. Pattern Recogn. Lett. 27 (8), 861–874.

Fischl, B., 2012. FreeSurfer. Neuroimage 62 (2), 774–781.

Franke, K., Gaser, C., 2014. Dementia classification based on brain age estimation. Proc MICCAI Workshop Challenge on Computer-Aided Diagnosis of Dementia Based on Structural MRI Data, pp. 48–54.

Frisoni, G.B., Jack, C.R., 2011. Harmonization of magnetic resonance-based manual hippocampal segmentation: a mandatory step for wide clinical use. Alzheimers Dement. 7 (2), 171–174.

Gaonkar, B., Davatzikos, C., 2013. Analytic estimation of statistical significance maps for support vector machine based multi-variate image analysis and classification. Neuroimage 78, 270–283.

Gray, K.R., Aljabar, P., Heckemann, R.A., Hammers, A., Rueckert, D., 2013. Random forest-based similarity measures for multi-modal classification of Alzheimer's disease. Neuroimage 65, 167–175.

Hand, D.J., Till, R.J., 2001. A simple generalisation of the area under the ROC curve for multiple class classification problems. Mach. Learn. 45, 171–186.

Jack, C.R., Bernstein, M., Fox, N.C., Thompson, P., Alexander, G., Harvey, D., Borowski, B., Britson, P.J., Whitwell, J.L., 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., 2008. The Alzheimer's Disease Neuroimaging Initiative (ADNI): MRI methods. J. Magn. Reson. Imaging 27 (4), 685–691.

Jack, C.R., Wiste, H.J., Vemuri, P., Weigand, S.D., Senjem, M.L., Zeng, G., Bernstein, M.a., Gunter, J., Pankratz, V.S., Aisen, J., Weiner, M.W., Petersen, J., Shaw, L.M., Trojanowski, J.Q., Knopman, J., 2010. Brain beta-amyloid measures and magnetic resonance imaging atrophy both predict time-to-progression from mild cognitive impairment to Alzheimer's disease. Brain 133 (11), 3336–3348.

Jack, C.R., Albert, M.S., Knopman, D.S., McKhann, G.M., Sperling, R.A., Carrillo, J., Thies, B., Phelps, J., 2011. Introduction to the recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimers Dement. 7 (3), 257–262.

Jack, C.R., Vemuri, P., Wiste, H.J., Weigand, S.D., Lesnick, T.G., Lowe, V., Kantarci, K., Bernstein, J., Senjem, M.L., Gunter, J., Boeve, B.F., Trojanowski, J.Q., Shaw, J., Aisen, P.S., Weiner, J., Petersen, R.C., Knopman, D.S., the Alzheimers Disease Neuroimaging, 2012. Shapes of the trajectories of 5 major biomarkers of Alzheimer disease. Arch. Neurol. 69 (7), 856–867.

Jack, C.R., Knopman, D.S., Jagust, W.J., Petersen, R.C., Weiner, M.W., Aisen, P.S., Shaw, L.M., Vemuri, P., Wiste, H.J., Weigand, S.D., Lesnick, T.G., Pankratz, J., Donohue, M.C., Trojanowski, J.Q., 2013. Tracking pathophysiological processes in Alzheimer's disease: an updated hypothetical model of dynamic biomarkers. Lancet Neurol. 12 (2), 207–216.

Kazee, A.M., Eskin, T.A., Lapham, L.W., Gabriel, K.R., McDaniel, K.D., Hamill, R.W., 1993. Clinicopathologic correlates in Alzheimer disease: assessment of clinical and pathologic diagnostic criteria. Alzheimer Dis. Assoc. Disord. 7 (3), 152–164.

Klöppel, S., Stonnington, J., Chu, C., Draganski, J., Scahill, I., Rohrer, J., Fox, N.C., Jack Jr., J., Ashburner, J., Frackowiak, R.S.J., Scahill, J., Jack, C.R., 2008. Automatic classification of MR scans in Alzheimer's disease. Brain 131 (3), 681–689.

Klöppel, S., Abdulkadir, J., Jack, C.R., Koutsouleris, J., Mourão Miranda, J., Vemuri, P., 2012. Diagnostic neuroimaging across diseases. Neuroimage 61 (2), 457–463.

Klunk, W.E., Engler, H., Nordberg, A., Wang, Y., Blomqvist, G., Holt, D.P., Bergstro, M., Savitcheva, I., Debnath, M.L., Barletta, J., Price, J.C., Sandell, J., Lopresti, B.J., Wall, A., Koivisto, P., Antoni, G., Mathis, C.A., Långström, B., 2004. Imaging brain amyloid in Alzheimer's disease with Pittsburgh compound B. Ann. Neurol. 55, 306–319.

Koikkalainen, J., Pölönen, H., Mattila, J., van Gils, M., Soininen, J., Lötjönen, J., 2012. Improved classification of Alzheimer's disease data via removal of nuisance variability. PLoS One 7 (2) (e31112-e31118).

Konukoglu, E., Glocker, B., Zikic, D., Criminisi, A., 2013. Neighbourhood approximation using randomized forests. Med. Image Anal. 17 (7), 790–804.

Lebedev, A.V., Westman, E., Beyer, M.K., Kramberger, J., Aguilar, C., Pirtosek, J., Aarsland, D., 2013. Multivariate classification of patients with Alzheimer's and dementia with Lewy bodies using high-dimensional cortical thickness measurements: an MRI surface-based morphometric study. J. Neurol. 260 (4), 1104–1115.

Ledig, C., Guerrero, R., Tong, T., Gray, K., Makropoulos, A., Heckemann, J., Rueckert, D., 2014. Alzheimer's disease state classification using structural volumetry, cortical thickness and intensity features. Proc MICCAI Workshop Challenge on Computer-Aided Diagnosis of Dementia Based on Structural MRI Data, pp. 55–64.

Leung, K.Y.E., van der Lijn, J., Vrooman, H.A., Sturkenboom, M.C.J.M., Niessen, J., 2014. IT infrastructure to support the secondary use of routinely acquired clinical imaging data for research. Neuroinformatics.

Lim, A., Tsuang, D., Kukull, W., Nochlin, D., Leverenz, J., McCormick, W., Bowen, J., Teri, L., Thompson, J., Peskind, E.R., Raskind, M., Larson, E.B., 1999. Clinico-neuropathological correlation of Alzheimer's disease in a community-based case series. J. Am. Geriatr. Soc. 47 (5), 564–569.

Lovestone, S., Francis, P., Kloszewska, I., Mecocci, P., Simmons, A., Soininen, H., Spenger, C., Tsolaki, M., Vellas, B., Wahlund, L.O., Ward, M., 2009. AddNeuroMed—the European collaboration for the discovery of novel biomarkers for Alzheimer's disease. Ann. N. Y. Acad. Sci. 1180, 36–46.

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

Mattila, J., Soininen, H., Koikkalainen, J., Rueckert, J., Wolz, R., Waldemar, G., Lötjönen, J., 2012. Optimizing the diagnosis of early Alzheimer's disease in mild cognitive impairment subjects. J. Alzheimers Dis. 32 (4), 969–979.

McKhann, G., Drachman, D., Folstein, M., Katzman, R., Price, D., Stadlan, E.M., 1984. Clinical diagnosis of Alzheimer's disease: report of the NINCDS-ADRDA work group under the auspices of Department of Health and Human Services Task Force on Alzheimer's disease. Neurology 34 (7), 939–944.

McKhann, G.M., Knopman, D.S., Chertkow, H., Hyman, B.T., C.R.J. Jr., Kawas, C.H., Klunk, W.E., Koroshetz, W.J., Manly, J.J., Mayeux, R., Mohs, R.C., Morris, J.C., Rossor, M.N., Scheltens, P., Carillo, M.C., Thies, B., Weintraub, S., Phelps, C.H., 2011. 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. Alzheimers Dement. 7, 263–269.

Misra, C., Fan, Y., Davatzikos, C., 2009. Baseline and longitudinal patterns of brain atrophy in MCI patients, and their use in prediction of short-term conversion to AD: results from ADNI. Neuroimage 44 (4), 1415–1422.

Moradi, E., Gaser, C., Huttunen, H., Tohka, J., 2014. MRI based dementia classification using semi-supervised learning and domain adaptation. Proc MICCAI Workshop Challenge on Computer-Aided Diagnosis of Dementia Based on Structural MRI Data, pp. 65–73.

Papma, J.M., de Groot, M., de Koning, I., Mattacel-Raso, J., van der Lugt, A., Vernooij, M.W., Niessen, J., van Swieten, J.C., Koudstaal, P.J., Prins, N.D., Smits, M., 2014. Cerebral small vessel disease affects white matter microstructure in mild cognitive impairment. Hum. Brain Mapp. 35 (6), 2836–2851.

Paquerault, S., 2012. Battle against Alzheimer's disease: the scope and potential value of magnetic resonance imaging biomarkers. Acad. Radiol. 19, 509–511.

Petersen, R.C., 2004. Mild cognitive impairment as a diagnostic entity. J. Intern. Med. 256 (3), 183–194.

Petrovitch, H., White, L.R., Ross, G.W., Steinhorn, S.C., Li, C.Y., Masaki, K.H., Davis, D.G., Nelson, J., Hardman, J., Curb, J.D., Blanchette, P.L., Launer, J., Yano, K., Markesbery, J., 2001. Accuracy of clinical criteria for AD in the Honolulu–Asia aging study, a population-based study. Neurology 57 (2), 226–234.

Prince, M., Bryce, R., Ferri, C., 2011. World Alzheimer Report 2011, the Benefits of Early Diagnosis and Intervention. Alzheimer's Disease International.

Prince, M., Bryce, R., Albanese, E., Wimo, A., Ribeiro, W., Ferri, C.P., 2013. The global prevalence of dementia: a systematic review and metaanalysis. Alzheimers Dement. 9 (1), 63–75.e2.

Provost, F., Domingos, P., 2001. Well-trained PETs: improving probability estimation trees. Technical Report; CeDER Working Paper #IS-00-04. Stern School of Business, New York University, New York, NY, USA.

Raamana, P.R., Rosen, H., Miller, B., Weiner, M.W., Wang, L., Beg, M.F., 2014. Three-class differential diagnosis among Alzheimer disease, frontotemporal dementia, and controls. Front. Neurol. 5 (71), 1–15.

Reuter, M., Wolter, F.E., Peinecke, N., 2006. Laplace–Beltrami spectra as 'shape-DNA' of surfaces and solids. Comput. Aided Des. 38 (4), 342–366.

Routier, A., Gori, P., Graciano Fouquier, A.B., Lecomte, S., Colliot, O., Durrleman, S., 2014. Evaluation of morphometric descriptors of deep brain structures for the automatic classification of patients with Alzheimer's disease, mild cognitive impairment and elderly controls. Proc MICCAI Workshop Challenge on Computer-Aided Diagnosis of Dementia Based on Structural MRI Data, pp. 74–81.

Sabuncu, M.R., Konukoglu, E., 2014. Clinical prediction from structural brain MRI scans: a large-scale empirical study. Neuroinformatics.

Sabuncu, M.R., Van Leemput, J., 2012. The relevance voxel machine (RVoxM): a self-tuning Bayesian model for informative image-based prediction. IEEE Trans. Med. Imaging 31 (12), 2290–2306.

Sarica, A., di Fatta, G., Smith, G., Cannataro, M., Saddy, J.D., 2014. Advanced feature selection in multinominal dementia classification from structural MRI data. Proc MICCAI Workshop Challenge on Computer-Aided Diagnosis of Dementia Based on Structural MRI Data, pp. 82–91.

Sensi, F., Rei, L., Gemme, G., Bosco, P., Chincarini, A., 2014. Global disease index, a novel tool for MTL atrophy assessment. Proc MICCAI Workshop Challenge on Computer-Aided Diagnosis of Dementia Based on Structural MRI Data, pp. 92–100.

Smith, G.M., Stoyanov, Z.V., Greetham, D.V., Grindrod, P., Saddy, J.D., 2014. Disease A., Initiative N. Towards the Computer-aided Diagnosis of Dementia based on the geometric and network connectivity of structural MRI data. Proc MICCAI Workshop Challenge on Computer-Aided Diagnosis of Dementia Based on Structural MRI Data, pp. 101–110.

Sørensen, L., Pai, A., Anker, C., Balas, I., Lillholm, M., Igel, C., Nielsen, M., 2014. Dementia diagnosis using MRI cortical thickness, shape, texture, and volumetry. Proc MICCAI Workshop Challenge on Computer-Aided Diagnosis of Dementia Based on Structural MRI Data, pp. 111–118.

Tangaro, S., Inglese, P., Maglietta, R., Tateo, A., 2014. MIND-BA: fully automated method for Computer-Aided Diagnosis of Dementia based on structural MRI data. Proc MICCAI Workshop Challenge on Computer-Aided Diagnosis of Dementia Based on Structural MRI Data, pp. 119–128.

Tustison, N.J., Avants, B.B., Cook, P.A., Zheng, Y., Egan, A., Yushkevich, P.A., Gee, J.C., 2010. N4ITK: improved N3 bias correction. IEEE Trans. Med. Imaging 29 (6), 1310–1320.

van der Flier, W.M., Pijnenburg, J., Prins, N., Lemstra, J., Bouwman, F.H., Teunissen, C.E., van Berckel, J., Stam, C.J., Barkhof, J., Visser, P.J., van Egmond, J., Scheltens, P., 2014. Optimizing patient care and research: the Amsterdam dementia cohort. J. Alzheimers Dis. 41 (1), 313–327.

Vemuri, P., Gunter, J.L., Senjem, M.L., Whitwell, J.L., Kantarci, K., Knopman, D.S., Boeve, B.F., Petersen, R.C., Jack Jr., C.R., 2008. Alzheimer's disease diagnosis in individual subjects using structural MR images: validation studies. Neuroimage 39 (3), 1186–1197.

Wachinger, C., Batmanghelich, J., Golland, P., Reuter, J., 2014a. BrainPrint in the Computer-Aided Diagnosis of Alzheimer's disease. Proc MICCAI Workshop Challenge on Computer-Aided Diagnosis of Dementia Based on Structural MRI Data, pp. 129–138.

Wachinger, C., Golland, P., Reuter, M., 2014b. BrainPrint: identifying subjects by their brain. Proc Intl Conf Med Image Comput Comp Ass Intervent. Lecture Notes in Computer Science. vol. 8675, pp. 41–48.

Wolz, R., Julkunen, V., Koikkalainen, J., Niskanen, J., Zhang, D.P., Rueckert, J., Soininen, H., Lötjönen, J., 2011. Multi-method analysis of MRI images in early diagnostics of Alzheimer's disease. PLoS One 6 (10) (e25446-e25446).

Please cite this article as: Bron, E.E., et al., Standardized evaluation of algorithms for computer-aided diagnosis of dementia based on structural MRI: The CADDementia challenge, NeuroImage (2015), http://dx.doi.org/10.1016/j.neuroimage.2015.01.048