Domain Adaptation for Alzheimer's Disease Diagnostics

Christian Wachinger[a,b,c*], Martin Reuter[b,c]
for the Alzheimer's Disease Neuroimaging Initiative
and the Australian Imaging Biomarkers and Lifestyle flagship study of ageing[*]

[a] *Department of Child and Adolescent Psychiatry, Psychosomatic and Psychotherapy, Ludwig-Maximilian-University, Munich, Germany*
[b] *Athinoula A. Martinos Center for Biomedical Imaging, Department of Radiology, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA*
[c] *Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA, USA*

Abstract

With the increasing prevalence of Alzheimer's disease, research focuses on the early computer-aided diagnosis of dementia with the goal to understand the disease process, determine risk and preserving factors, and explore preventive therapies. By now, large amounts of data from multi-site studies have been made available for developing, training, and evaluating automated classifiers. Yet, their translation to the clinic remains challenging, in part due to their limited generalizability across different datasets. In this work, we describe a compact classification approach that mitigates overfitting by regularizing the multinomial regression with the mixed ℓ_1/ℓ_2 norm. We combine volume, thickness, and anatomical shape features from MRI scans to characterize neuroanatomy for the three-class classification of Alzheimer's disease, mild cognitive impairment and healthy controls. We demonstrate high classification accuracy via independent evaluation within the scope of the CADDementia challenge. We, furthermore, demonstrate that variations between source and target datasets can substantially influence classification accuracy. The main contribution of this work addresses this problem by proposing an approach for supervised domain adaptation based on instance weighting. Integration of this method into our classifier allows us to assess different strategies for domain adaptation. Our results demonstrate (i) that training on only the target training set yields better results than the naïve combination (union) of source and target training sets, and (ii) that domain adaptation with instance weighting yields the best classification results, especially if only a small training component of the target dataset is available. These insights imply that successful deployment of systems for computer-aided diagnostics to the clinic depends not only on accurate classifiers that avoid overfitting, but also on a dedicated domain adaptation strategy.

Keywords: Computer-Aided Diagnosis, Alzheimer's Disease, Classification, Domain Adaptation

1. Introduction

Alzheimer's disease (AD) is the most common form of dementia with incidence rates further increasing in the future due to increasing life expectancy. Early and accurate diagnosis of AD is a key objective as it can help patients to access supportive therapies earlier allowing them to maintain independence for longer (Paquerault, 2012). When treatment options that directly interfere with disease pathways finally become available, intervention will likely be most effective in early preclinical or presymptomatic disease stages. Furthermore, early identification of high-risk individuals can already support selection into promising drug trials, inform patient stratification, as well as aid the identification of risk and preserving factors. Magnetic resonance imaging (MRI) is an important tool for AD diagnosis because the atrophy measured in MRI correlates with neuron loss and can indicate the onset of the impairment in close temporal proximity (Jack et al., 2013). Computer-aided diagnosis of dementia based on MRI is an active research field as indicated by 50 articles reviewed on this topic by Falahati et al. (2014). The deployment of automated system for diagnosis of AD in the clinic promises several advantages: (i) the improvement of diagnosis in places with limited neuroradiological

[*] Corresponding Author. Address: Waltherstr. 23, 81369 München, Germany; Email: christian.wachinger@med.uni-muenchen.de

[**] Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: `http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf`

Data used in the preparation of this article was obtained from the Australian Imaging Biomarkers and Lifestyle flagship study of ageing (AIBL) funded by the Commonwealth Scientific and Industrial Research Organisation (CSIRO) which was made available at the ADNI database (www.loni.usc.edu/ADNI). The AIBL researchers contributed data but did not participate in analysis or writing of this report. AIBL researchers are listed at www.aibl.csiro.au.

know-how, (ii) a faster diagnosis without compromising accuracy by avoiding lengthy specialist investigations, and (iii) a more objective diagnostic assessment based increasingly on quantitative information in contrast to traditionally more subjective diagnostic impression (Klöppel et al., 2012). Computational diagnostics promise to be particularly useful for screening purposes to identify individuals with preclinical disease.

Large, multi-center datasets are available for studying Alzheimer's disease and for supporting the training of complex classification models. A challenge for such models is generalizability, i.e., the ability to transfer a model that is trained on one dataset to another dataset while retaining high prediction accuracy. In an attempt to provide an objective assessment of state-of-the-art methods for AD classification, the CADDementia challenge has been organized recently (Bron et al., 2015). The task was to differentiate between patients with Alzheimer's disease, mild cognitive impairment (MCI), and healthy controls (CN) based on T1-weighted MRI data. Classification accuracy of a variety of submissions was evaluated on an independent test dataset with hidden diagnosis. Intriguingly, the study showed that all participating groups overestimated the accuracy of their method. One of the main reasons for the overestimation may be overfitting to the training data. Neuroimaging applications are susceptible to overfitting due to a potentially large number of features extracted from images and a restricted number of samples available for training. Overfitting is further aggravated by complex classification models with many degrees-of-freedom that easily fine-tune to a specific population but overestimate the performance on the general population (Adaszewski et al., 2013; Mwangi et al., 2014). In our classifier we employ methods that mitigate overfitting by (i) using sparsity constraints to estimate a compact model and by (ii) choosing a linear classification model based on multinomial regression to further limit the number of free parameters. Yet, in spite of these efforts, the bias towards overestimating performance on the training set still prevails, indicating that overfitting may not solely be responsible. Here, we identify another cause for reduced classification accuracy on the final test set: the differences in the distribution between training and test data.

The main contributions of this work are twofold: We introduce a compact classifier for Alzheimer's disease that incorporates shape information and evaluate its performance on an independent test setting. We further demonstrate that variations in source and target datasets have a large impact on classification accuracy and present a novel algorithm for domain adaptation that re-weights samples from the source dataset.

1.1. Computer-Aided Diagnosis of Dementia

Predicting or classifying dementia based on structural MRI is an active field of research. Cuingnet et al. (2011) compare several approaches for the discrimination of AD and MCI patients using the cortical thickness, the hippocampus and voxel-based methods. Falahati et al. (2014) review the literature for the classification of individuals with dementia. The extensive list of articles discussed in the review illustrates the wide interest in the research field. In this work, we introduce an algorithm for AD classification that is based on *BrainPrint* (Wachinger et al., 2015) for quantifying brain morphology, which naturally extends the region of interest (ROI)-based volume and thickness analysis with shape information (Reuter et al., 2006). Anatomical shape features contribute valuable information to the characterization of brain structures, which are only coarsely represented by their volume. Finding representative and descriptive features is crucial for automatic classification as it is well known in pattern recognition that the prediction accuracy is primarily driven by the representation (Dickinson, 2009).

Both of the review articles mentioned above refer to a total of only three publications that employ shape information, indicating that shape is not commonly used. Most previous work that includes shape analysis, typically focus on a single structure, predominantly the hippocampus. More precisely, Gerardin et al. (2009) approximate the hippocampal shape by a series of spherical harmonics. Ferrarini et al. (2009) use permutation tests to extract surface locations that are significantly different among patients with AD and controls. Costafreda et al. (2011) incorporate shape information by deriving thickness measurements of the hippocampus from a medical representation. Shen et al. (2012) use statistical shape models to detect hippocampal shape changes. Bates et al. (2011) investigated spectral signatures for AD classification, with a focus on right hippocampus, right thalamus and right putamen. Other structures of interest for shape analysis were the cortex and ventricles: Kim et al. (2014) use multiresolution shape features with non-Euclidean wavelets for the analysis of cortical thickness, King et al. (2010) analyze the fractal dimension of the cortical ribbon, and Gutman et al. (2013) model surface changes of the ventricles in a longitudinal setup with a medial representation. In contrast to all these studies, we incorporate an ensemble of both cortical and subcortical structures. This extensive characterization of brain anatomy is promising in diagnosing Alzheimer's disease, which is associated with widespread atrophy across the entire brain.

1.2. Domain Adaptation

As described above, differences between source and target datasets can significantly reduce classification accuracy. In traditional cross-validation, where a single dataset is split into subsets, such variations are negligible, as the subsets tend to represent the data well. However, when an independent dataset is used for testing, differences in the distributions can have a dramatic impact on the classification accuracy. Such problems are studied in domain adaptation (Pan and Yang, 2010), where the model is learned on a source dataset and then transferred to a target dataset with different properties. In fact, we believe that domain adaptation is crucial for the translation of computer-aided diagnostic methods to the clinic, where the source dataset usually consists of large, possibly multi-center, data and the target dataset is the (limited) data acquired at the specific hospital, where the system is deployed. There are clearly several factors that can contribute to variations between source and target datasets arising from location and selection biases.

Here, we assume a supervised domain adaptation scenario, where a subset of the target dataset is available for training, replicating the situation that a small, local dataset from the clinic is available to support training. Based on this small target training set we weight samples from the source dataset to match distributional properties of the target dataset. The proposed instance weighting presents a general framework, where naïve strategies for combining source and target training data (e.g. the union or selecting one vs. the other) can be derived by setting the weights to appropriate constants. We measure a variation in classification accuracy of more than 20% across strategies, highlighting the importance of domain adaptation. Domain adaptation with instance weighting has previously been described in the machine learning literature (Bickel et al., 2007; Jiang and Zhai, 2007). An unsupervised domain adaptation strategy for AD classification was used by Moradi et al. (2014). This strategy applies discriminative clustering on the source and target domain, where a feature weighting is learned by optimizing the mutual information (Shi and Sha, 2012). In contrast, we use a supervised domain adaptation strategy and do not weight features but instances. Further related are approaches that assume a semi-supervised classification setting (Zhao et al., 2014; Adeli-Mosabbeb et al., 2015), yet they operate on the same domain.

Domain adaptation has previously been successfully used in medical image analysis. van Opbroek et al. (2015) proposed transfer learning for image segmentation across scanners and image protocols with support vector machines and AdaBoost. Heimann et al. (2013) used domain adaptation for the localization of ultrasound transducers in X-ray images with probabilistic boosting trees. Schlegl et al. (2014) applied domain adaptation for lung tissue classification with convolutional neural networks.

2. Methods

In this section, we introduce our approach to AD classification with domain adaptation. The classification task is to predict the diagnostic label y of an individual based on image and non-image data summarized in the vector $\mathbf{x}$. First, we introduce the multinomial classifier used for the prediction. We then derive our approach to multinomial regression with domain adaptation, and finally describe the extraction of image-based features from MRI scans with a focus on shape features from the *BrainPrint*.

2.1. Elastic-Net Multinomial Regression for Alzheimer's Classification

We employ multinomial regression with a generalized linear model for the classification of subjects in three diagnostic groups (controls, MCI, and AD). The high-dimensional characterization of an individual may cause overfitting on the training dataset. We therefore select a subset of the features to establish a compact model. The elastic-net regularizes multinomial regression: During the estimation of a model for predicting diagnostic label y from observation $\mathbf{x}$ it identifies the most predictive variables (Friedman et al., 2010). The categorical response variable y has $K = 3$ levels with AD, MCI, and CN. We use the multi-logit model, which is a generalization of linear logistic regression to the multi-class situation. The conditional probability for label ℓ is

$$p(y = \ell|\mathbf{x}) = \frac{e^{\beta_{0\ell}+\mathbf{x}^\top \beta_\ell}}{\sum_{k=1}^{K} e^{\beta_{0k}+\mathbf{x}^\top \beta_k}} \tag{1}$$

with regression coefficients β. The model is fitted by regularized maximum multinomial likelihood with a penalty on the regression coefficients $R(\beta)$

$$\max_{\{\beta_{0\ell},\beta_\ell\}_1^K \in \mathbb{R}^{K(p+1)}} \left[\frac{1}{N} \sum_{i=1}^{N} \log p(y_i|\mathbf{x}_i) - \kappa \sum_{\ell=1}^{K} R_\alpha(\beta_\ell) \right] \tag{2}$$

with N the number of training samples. The parameter κ balances the data fit term with the penalty term.

The regularizer in elastic-net combines lasso ℓ_1 and ridge-regression ℓ_2 penalties, modulated by the parameter α

$$R_\alpha(\beta) = (1-\alpha)\frac{1}{2}\|\beta\|_2^2 + \alpha\|\beta\|_1. \tag{3}$$

For correlated predictors, ridge-regression shrinks their coefficients and allows them to borrow strength from each other. In contrast, lasso will tend to pick one and ignore the rest. Increasing α from 0 to 1 will monotonically increase the sparsity of the solution until the lasso solution is reached. In a comparison of methods for model selection (Wachinger et al., 2014), we obtained the best results for the elastic-net, when compared to manual selection or the stepwise selection with the Akaike information criterion.

2.2. Domain Adaptation for AD Classification

Domain adaptation distinguishes between a *source* domain and a *target* domain (Quionero-Candela et al., 2009; Pan and Yang, 2010; Margolis, 2011). The setup for supervised domain adaptation is schematically illustrated in Fig. 1. The source domain is the training domain with labeled data $D_s = \{(\mathbf{x}_i^s, y_i^s)\}_{i=1}^{N_s}$, while the target domain is the test domain with only a fraction of labeled data. The labeled data in the target domain is denoted as $D_t = \{(\mathbf{x}_i^t, y_i^t)\}_{i=1}^{N_t}$ and the unlabeled data in the target domain as $D_u = \{(\mathbf{x}_i^u, y_i^u)\}_{i=1}^{N_u}$. As common in supervised

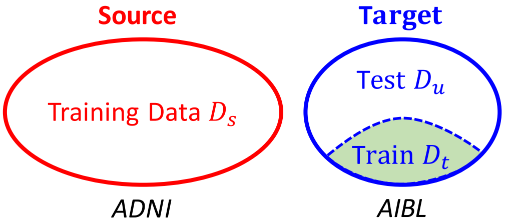

Figure 1: Supervised domain adaptation with source (red) and target (blue) domains, where a part of the target data (green) is available for training. In the experiments, we use the ADNI data as source data and AIBL data as target data.

domain adaptation, we assume $N_s \gg N_t$ and further that the training subset in the target domain is representative of the entire target dataset.

For the formulation of the problem, we consider supervised learning as empirical risk minimization. Abstractly, the optimal model θ^* in the model family Θ is inferred by minimizing the loss function L

$$\theta^* = \arg\min_{\theta\in\Theta} \sum_{(\mathbf{x},y)\in\mathcal{X}\times\mathcal{Y}} p(\mathbf{x},y)\cdot L(\mathbf{x},y,\theta) \tag{4}$$

with the joint distribution over observations and labels $p(\mathbf{x},y)$, the input space $\mathcal{X}$ and label set $\mathcal{Y}$. In this work, we use the negative log-likelihood function from the multilogit model (first term in Eq. (2)) as loss function, but the formulation is of general nature and therefore also extends to other loss functions. Since the joint distribution $p(\mathbf{x},y)$ is unknown, we use the empirical approximation with the training data $\{(\mathbf{x}_i,y_i)\}_{i=1}^N$

$$\hat\theta = \arg\min_{\theta\in\Theta} \sum_{i=1}^{N} L(\mathbf{x}_i,y_i,\theta). \tag{5}$$

This setup considers a single dataset. For domain adaptation, we want to infer the model with minimal loss on the target dataset D_t, while the source training sample D_s is randomly sampled from the source distribution p_s. In domain adaptation with *instance weighting*, this problem is addressed by re-weighting the elements of the source training dataset. The weight is dependent on the probability of source samples in the target domain, which can be derived as follows

$$\begin{aligned}\theta_t^* &= \arg\min_{\theta\in\Theta} \sum_{(\mathbf{x},y)\in\mathcal{X}\times\mathcal{Y}} p_t(\mathbf{x},y)\cdot L(\mathbf{x},y,\theta) && (6)\\ &= \arg\min_{\theta\in\Theta} \sum_{(\mathbf{x},y)\in\mathcal{X}\times\mathcal{Y}} \frac{p_t(\mathbf{x},y)}{p_s(\mathbf{x},y)} p_s(\mathbf{x},y)\cdot L(\mathbf{x},y,\theta) && (7)\\ &\approx \arg\min_{\theta\in\Theta} \sum_{i=1}^{N_s} \frac{p_t(\mathbf{x}_i^s,y_i^s)}{p_s(\mathbf{x}_i^s,y_i^s)}\cdot L(\mathbf{x}_i^s,y_i^s,\theta). && (8)\end{aligned}$$

Considering further that training data is also available

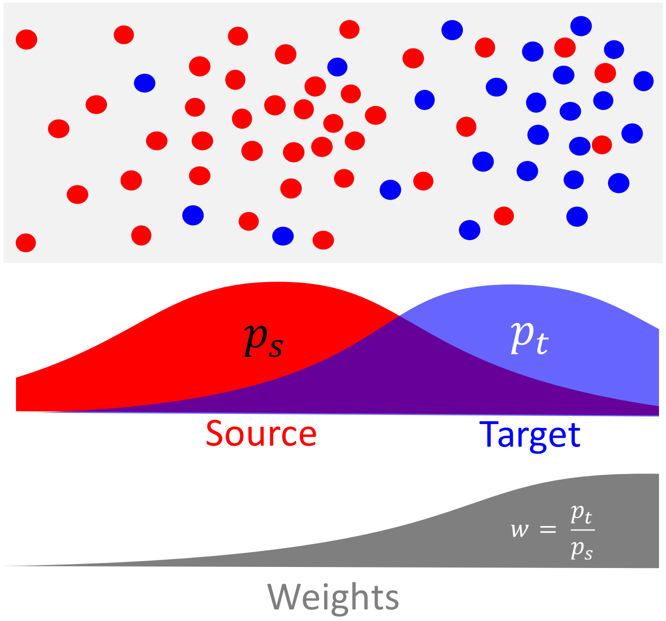

Figure 2: Schematic illustration of data points (subjects) from the source (blue) and target (red) domain (top). Marginal distributions of the source p_s and target p_t domain (center) show the variation between both domains. With instance weighting by the term $w = p_t/p_s$ (bottom), red source data points on right are assigned higher weights in the estimation of the regression coefficients, while source data points on the left are assigned lower weights.

from the target domain, the model estimation is

$$\theta_t^* = \arg\min_{\theta\in\Theta} \sum_{i=1}^{N_s} \frac{p_t(\mathbf{x}_i^s,y_i^s)}{p_s(\mathbf{x}_i^s,y_i^s)}\cdot L(\mathbf{x}_i^s,y_i^s,\theta) + \sum_{i=1}^{N_t} L(\mathbf{x}_i^t,y_i^t,\theta). \tag{9}$$

The ratio

$$w_i = \frac{p_t(\mathbf{x}_i^s,y_i^s)}{p_s(\mathbf{x}_i^s,y_i^s)} \tag{10}$$

is the weighting factor for the sample $(\mathbf{x}_i^s,y_i^s)$. It evaluates the probability of the source data sample $(\mathbf{x}_i^s,y_i^s)$ under the target distribution p_t, normalized by the probability under the the source distribution. The challenging part is the estimation of the probability of the source data under the target distribution $p_t(\mathbf{x}_i^s,y_i^s)$. For supervised domain adaptation, we have training data in the target domain D_t, which we will use for estimating the target distribution. Fig. 2 illustrates source and target datasets. The probability density functions highlight the different distribution of both datasets. The weight emphasizes points that lie in high density regions of the target domain and low density regions of the source domain, in the figure these are the red dots towards the right. With this re-weighting of the source domain, we infer a model that is better adapted to the target domain and consequently yields higher classification accuracy.

In contrast to covariate shift that only considers variations in the marginal distributions of the observations

$p_s(\mathbf{x}) \neq p_t(\mathbf{x})$ (Shimodaira, 2000) and class imbalance that only considers variations in the marginal distributions of the labels $p_s(y) \neq p_t(y)$ (Japkowicz and Stephen, 2002), we consider variations in the joint distribution $p_s(\mathbf{x}, y) \neq p_t(\mathbf{x}, y)$. To facilitate the estimation, we use the factorization $p(\mathbf{x}, y) = p(\mathbf{x})p(y)$, which implies the assumption of independence. The factorization is continued on the multivariate variable $\mathbf{x}$, yielding a product of univariate densities. An alternative would be the direct estimation of the high-dimensional joint density, however, a reliable estimation would require the number of samples to grow exponentially with the number of dimensions. Given the limited number of elements in the target training dataset, a reliable estimation in higher dimensions seems not feasible. In this work, we select diagnostic information, age, sex, and the number APOE4 allele for instance weighting, because they capture important characteristics about an individual. The distributions are estimated with histogramming for discrete variables and kernel density estimation for continuous variables.

The integration of instance weighting for domain adaptation in the multinomial regression in Eq.(2) results in

$$\max_{\{\beta_{0\ell}, \beta_\ell\}_1^K \in \mathbb{R}^{K(p+1)}} \left[\sum_{i=1}^{N_s} w_i \log p(y_i^s | \mathbf{x}_i^s) + \sum_{i=1}^{N_t} \log p(y_i^t | \mathbf{x}_i^t) \right. \\ \left. - \kappa \sum_{\ell=1}^{K} P_\alpha(\beta_\ell) \right], \tag{11}$$

where the regression coefficients β correspond to the model parameters θ and training samples in the log-likelihood function are weighted. The weights w_i are defined in Eq.(10) for training samples of the source dataset. Samples from the target training dataset have constant weight one. For solving the optimization problem in Eq.(11), a coordinate descent scheme is used (Friedman et al., 2010). In nested loops over the parameter κ and the classes, partial quadratic approximations to the log-likelihood are computed, where regression coefficients only vary for a single class at a time. Coordinate descent is then used to solve the penalized weighted least-squares problem.

In addition to setting the weights according to the equation for instance weighting, we can enforce different domain adaptation strategies by setting the weights to a constant, $w_i = c$. The strategies differ in the data used for training:

I. $c \gg 1$: only the source training dataset D_s,
II. $c = 0$: only the target training dataset D_t,
III. $c = 1$: the union of both training sets $D_s \cup D_t$.

Selecting either one of the datasets or combining both are straightforward approaches that do not take the challenges of domain adaptation into account. We show in the results' section that the chosen strategy substantially influences the classification result.

2.3. Image-Based Features for Classification

As mentioned above, the selection of descriptive features for the classifier is essential for the performance. For the extraction of image-based features, we process the scans with FreeSurfer (Dale and Sereno, 1993; Dale et al., 1999; Fischl et al., 1999a,b, 2002). FreeSurfer automatically segments cortical and subcortical structures in the image. Based on the segmentation, we use the thickness of 70 cortical regions and the volume 39 brain structures. Next to volume and thickness, we use shape features that are derived from the *BrainPrint*, further explained in Section 2.3.1. The shape features include 14 lateral shape differences and 44 PCA shape variations. The total number of image features for the classification is 167. According to the recent analysis of the normalization of variables for AD classification (Westman et al., 2013), we normalize volumetric measures by the intracranial volume (ICV) but do not normalize cortical thickness measures. Age residualization with linear regression was performed for each feature to remove the confounding effect of age in the analysis. After ICV normalization and age residualization, age and gender was not used in multinomial regression.

2.3.1. The Brain Descriptor BrainPrint

We create surface and volumetric meshes from cortical and subcortical segmentations. Based on these meshes, we compute compact shape representations for all structures, constituting the *BrainPrint* (Wachinger et al., 2015). The *shapeDNA* (Reuter et al., 2006) is used as shape descriptor, which performed among the best in a comparison of methods for non-rigid 3D shape retrieval (Lian et al., 2012). The Laplace-Beltrami spectrum is computed on the intrinsic geometry of the object to form the *shapeDNA*. Considering the Laplace-Beltrami operator Δ, one obtains the spectrum by solving the Laplacian eigenvalue problem (Helmholtz equation)

$$\Delta f = -\lambda f. \tag{12}$$

The solution consists of eigenvalue $\lambda_i \in \mathbb{R}$ and eigenfunction f_i pairs, sorted by eigenvalues, $0 \leq \lambda_1 \leq \lambda_2 \leq \ldots$ (a positive diverging sequence). The first l non-zero eigenvalues computed using the finite element methods, form the *shapeDNA*: $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_l)$. To achieve scale independence, we normalize the eigenvalues, $\lambda' = \text{vol}^{\frac{2}{D}} \lambda$, where vol is the Riemannian volume of the D-dimensional manifold (Reuter et al., 2006), i.e., the surface area for 2D manifolds. Fig. 3 illustrates the first three eigenfunctions of the left white matter surface. The eigenfunctions show natural vibrations of the shape when oscillating at a frequency specified by the square root of the eigenvalue. We also map the same eigenfunctions on the inflated surface to highlight the characteristics of the eigenfunctions, not obscured by the complex cortical folding patterns.

The eigenvalues are isometry invariant with respect to the Riemannian manifold, meaning that length-preserving

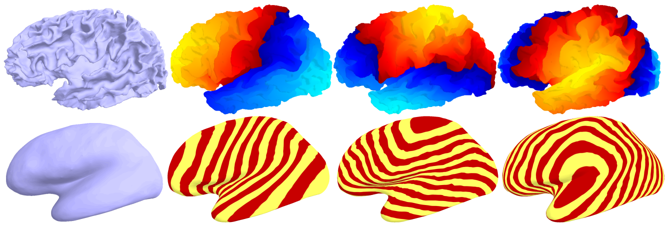

Figure 3: Left white matter surface and first three non-constant eigenfunctions of the Laplace-Beltrami operator calculated on the surface, where shape features are derived from natural frequencies of eigenfunctions. Top: Eigenfunctions shown with color gradient. Bottom: Inflated white matter surface with eigenfunctions shown as level sets. The main directions of variation are anterior-posterior, superior-inferior, and lateral-medial, respectively.

deformations will not change the spectrum. Isometry invariance includes rigid body motion and therefore permits the comparison of shapes across subjects by directly comparing the *shapeDNA*. A second property is that the spectrum continuously changes with topology-preserving deformations of the boundary of the object. These properties make the *shapeDNA* well suited for comparing shapes, as initial alignment of the shapes can be completely avoided.

We compute the spectra for all cortical and subcortical structures on the 2D boundary surfaces (triangle meshes (Reuter et al., 2009; Niethammer et al., 2007)) and additionally for cortical structures (white and pial surfaces in both hemispheres) also on the full 3D solid (tetrahedra meshes (Reuter et al., 2007)), forming the *BrainPrint* $\Lambda = (\boldsymbol{\lambda}_1, \dots, \boldsymbol{\lambda}_\eta)$. The *BrainPrint* contains 36 subcortical structures and 8 descriptors for cortical structures (left/right, white/gray matter, 2D/3D), yielding $\eta = 44$.

2.3.2. Features from BrainPrint

Depending on the number of eigenvalues l computed, we can easily end up with thousands of values in the *BrainPrint*, which may make the approach more susceptible to overfitting. To decrease the number of variables and increase robustness, we (i) use a 1D asymmetry measure (the distance of *BrainPrint* across hemispheres), and (ii) employ principal component analysis. An important aspect of the *BrainPrint* is that the eigenvalues form an increasing sequence with the variance increasing as well. Depending on the distance measure, this can cause higher eigenvalues to dominate the similarity measure between two shapes, although these components do not necessarily contain the most important geometric information. To account for these issues we normalize the *BrainPrint* and employ appropriate distance computations as described next.

Asymmtery: As a first shape feature, we measure the asymmetry of lateralized brain structures in *BrainPrint*. Since *shapeDNA* is invariant to mirroring, we directly compute the Mahalanobis distance between the descriptors of a lateralized brain structure s

$$d_s = \|\boldsymbol{\lambda}_s^{\text{left}} - \boldsymbol{\lambda}_s^{\text{right}}\|_{\Sigma_s}, \tag{13}$$

with Σ_s the covariance matrix across all subjects for structure s. The lateralized structures that we use are white matter and pial surfaces with triangular and tetrahedral meshes, as well as triangular meshes for cerebellum white matter and gray matter, striatum, lateral ventricles, hippocampus, amygdala, thalamus, caudate, putamen, and accumbens.

Alternative distance functions that have been proposed for *shapeDNA* are the Euclidean distance (or any p-norm), Hausdorff distances, the Euclidean distance on re-weighted eigenvalues $\hat{\lambda}_i = \lambda_i / i$ (Reuter et al., 2006; Reuter, 2006), and the weighted spectral distance (Konukoglu et al., 2013). The weighted distances (latter two approaches) are motivated by the need to reduce the impact of higher eigenvalues on the distance. The linear re-weighting is based upon the observation that the eigenvalues demonstrate a linear growth pattern (Weyl's law) and therefore yields an approximately equal contribution of each eigenvalue. The weighted spectral distance is similar to a division by the squared eigenvalue number and therefore functions like a low-pass filter. Here, we use the Mahalanobis distance because it accounts for the covariance pattern in the data and supports an equal contribution of all eigenvalues in the sequence.

Principal Component Analysis: We derive a second set of features from *BrainPrint* by computing principal components for each of the 44 brain structures. Projecting the *shapeDNA* on the principal component retains most of the variance in the dataset, while reducing the dimensionality. Problematic in this regard is once again that higher eigenvalues show most variance, so that they will dominate the identification of the principal component. We have experimented with (i) linear re-weighting, $\hat{\lambda}_i = \lambda_i / i$, and (ii) the normalization of each eigenvalue to unit variance across the dataset. Evaluation of both approaches yielded similar results, so that we employ the simpler linear re-weighting.

Software: *shapeDNA* and the *BrainPrint* software is available at `http://reuter.mit.edu/software/` and `https://github.com/reuter-lab/BrainPrint`. We integrated the domain adaptation in the glmnet[1] package in the statistical computing environment R.

2.4. Data

We use data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu), the Australian Imaging Biomarkers and Lifestyle Study of Aging (AIBL, Ellis et al. (2009)), and the CADDementia challenge (Bron et al., 2015). Table 1 summarizes the datasets used for the dementia prediction. All datasets were processed with FreeSurfer.

For the CADDementia data, we only have access to the diagnostic information for the 30 subjects in the training dataset. For the larger test dataset with 354 subjects,

[1] https://cran.r-project.org/web/packages/glmnet/index.html

	Subjects	Diagnosis			Gender		Age quantiles			APOE4		
		CN	MCI	AD	Male	Female	1st	2nd	3rd	0	1	2
ADNI	751	213	364	174	437(58%)	314(42%)	71.1	75.3	79.8	383	287	81
AIBL	215	142	39	34	114(53%)	101(47%)	67.5	73.0	79.0	107	90	18
CADDementia-Train	30	12	9	9	17(57%)	13(43%)	59.3	65.0	68.0	-	-	-
CADDementia-Test	354	-	-	-	213(60%)	141(40%)	59.0	64.0	71.0	-	-	-

Table 1: Demographic, diagnostic, and genetic information of datasets used in this study.

only the scans with gender and age information is provided. The CADDementia data is composed of imaging data from three medical centers: VU University Medical Center, Amsterdam, the Netherlands; Erasmus MC, Rotterdam, the Netherlands; University of Porto / Hospital de Sao Joao, Porto, Portugal.

AIBL study methodology has been reported previously (Ellis et al., 2009) and AIBL data was collected by the AIBL study group. The study was launched in 2006 and focuses on the early detection of AD, towards lifestyle interventions. The data is collected at two centers (40% subjects from Perth in Western Australia, 60% from Melbourne, Victoria).

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 assessment 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. The Principal Investigator of this initiative is Michael W. Weiner, MD, VA Medical Center and University of California - San Francisco. ADNI is the result of efforts of many coinvestigators from a broad range of academic institutions and private corporations, and subjects have been recruited from over 50 sites across the U.S. and Canada. The follow up duration of each group is specified in the protocol for ADNI. For up-to-date information, see www.adni-info.org.

3. Results

3.1. CADDementia Challenge

For an independent evaluation of the AD classification method, we report results of our method from the CADDementia challenge (Bron et al., 2015). The task of the challenge was to differentiate between patients with Alzheimer's disease, mild cognitive impairment, and healthy controls given T1-weighted MRI data. Within the scope of the challenge, 384 multi-center scans were released, where for 30 of the 384 scans the diagnosis was also provided. In addition to the 30 scans also other data, *e.g.*, from ADNI could be used for training. The prediction results for the 354 test cases were submitted to the challenge organizers for the evaluation.

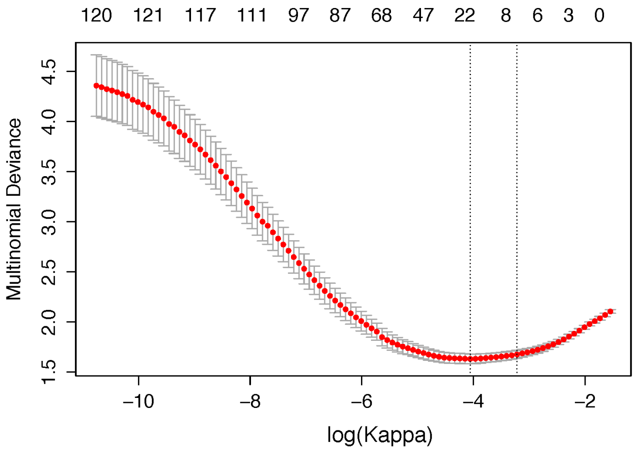

Figure 4: Multinomial deviance of elastic-net computed with cross-validation for different parameters κ (bottom) and the corresponding number of features (top). The plot shows the mean deviance together with upper and lower standard deviation.

Based on a comparison in (Wachinger et al., 2014), we selected $l = 40$ eigenvalues. We set $\alpha = 1$ and the parameter κ in the elastic-net that balances the data fit and penalty term with cross-validation. Fig. 4 shows the cross-validation results, where the lowest multinomial deviance of the model is roughly for $\kappa = \exp(-3.5)$. The figure also shows the number of selected features on top, where smaller κ yields the selection of more features and larger κ yields the selection of fewer features. We merged the 751 ADNI subjects with the 30 CADDementia training subjects for the training, which corresponds to strategy III. Fig. 5 shows the receiver operating characteristic curve for the prediction on the CADDementia test data. Table 2 lists the prediction accuracy together with the true positive fraction across the different diagnostic categories and the area under the curve. The table also shows the confidence interval, which is computed using bootstrapping

Accuracy	TPF-CN	TPF-MCI	TPF-AD	AUC
59.0 (54.0 - 63.6)	72.1 (63.4 - 79.2)	51.6 (43.5 - 61.3)	51.5 (41.5 - 61.2)	77.0 (73.6 - 80.3)

Table 2: Classification results on the CADDemetia test data, reported values are in %, confidence intervals are shown in parenthesis. (TPF = true positive fraction, AUC = area under the receiver operating characteristic curve).

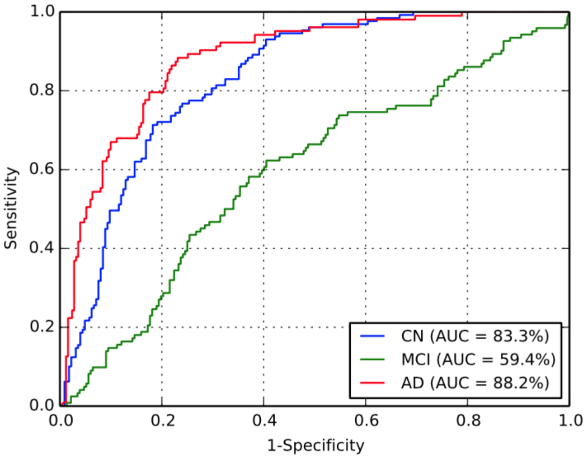

Figure 5: Receiver operating characteristic (ROC) curve for the classification in control (CN), mild cognitive impairment (MCI), and Alzheimer's disease (AD) on the CADDementia test data. The plot was created by the challenge organizers.

		True		
		CN	MCI	AD
Pred.	CN	93	44	6
	MCI	36	63	44
	AD	0	15	53

Table 3: Confusion matrix for the classification results on the CADDementia test data. Rows represent the predicted diagnostic classes, columns represent the true classes. Table entries were computed by the challenge organizers.

on the test set (1000 resamples) (Bron et al., 2015). Table 3 shows the classification results as confusion matrix. With these results, the proposed classifier was ranked on the second place after the algorithms of Soerensen in the challenge[2].

The results show that MCI is the most difficult class to predict. The trained elastic-net classifier selected 9 volume, 15 thickness, and 17 shape features (3 asymmetry, 14 principal components). Notable is the age difference between the challenge and ADNI data. The first quantile of ADNI is higher than the third quantile of the challenge data meaning that 75% of ADNI cases are older than 75% of challenge cases. This mismatch may have a detrimental effect on classification accuracy.

3.2. Domain Adaptation

For the evaluation of the domain adaptation strategies in Sec. 2.2, we use the AIBL data as independent target dataset, because the diagnostic information of the CADDementia-Test dataset has not been disclosed by the organizers in order to allow continuation of the challenge. To obtain the target training data D_t, we randomly sample a subset of the AIBL data. The remaining AIBL data serves as test data D_u. We vary the subset size between 0% and 30% of the original dataset and set $\alpha = 0.7$. Further, the number of apolipoprotein-E-ϵ4 (APOE4) allele is added to the variables because it is available for both ADNI and AIBL .

Fig. 6 depicts the three-class prediction accuracy on the AIBL test data for the domain adaptation strategies. The plot shows the mean accuracy and standard error over randomly sampling subsets 50 times from AIBL. The lowest accuracy is achieved with strategy I, where the classifier was only trained on the source data (ADNI). The low accuracy stems from the variations across the two datasets. When merging the ADNI data with training data from the target domain (strategy III), we see an accuracy improvement of about 5%. The accuracy increases about 10% when more data from the target dataset becomes available for training, i.e. when employing 30% instead of only 5% of the target data.

Surprisingly, only using the small fraction of the target training data for the estimation of the classifier (strategy II) yields a steep improvement in classification accuracy relative to the combination with the source training data (strategy III). Since these datasets are fairly small for low fractions, we note a larger variance of the results, as shown by the standard error. The consistently higher accuracies for strategy II in comparison to strategy III suggest that the common approach to simply merge source and target training set is is suboptimal. The larger source dataset dominates the smaller one and prevents optimal adaptation to the target data. As an intermediary between strategy II and strategy III, we use the weighting with a constant weight of $c = 0.5$. This strategy decreases the influence of all the source samples by half and therefore emphasizes the target samples. As expected, the accuracy of this strategy lies between using only target data and the union of source and target data.

The highest classification accuracy in this comparison is achieved for the proposed domain adaptation with instance weighting. The improvement is most prominent for small training fractions of the target data (5%), which is probably the most realistic scenario for the deployment in

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

the hospital. As mentioned previously, we use diagnostic information, age, sex, and the number APOE4 allele for instance weighting. Note, that the standard error for instance weighting is lower than for strategy II (only AIBL), because we also include the ADNI data, making the classification more robust. Increasing the amount of target data available for training above 30% reported in Fig. 6 yields a further increase in classification accuracy for all strategies, where strategy II and instance weighting have a similar accuracy after about 60%. Fig. 7 illustrates the receiver operating characteristic curve for the prediction on the AIBL test data with weighting and 30% of target data available for training. As for the previous results on the CADDementia, the lowest accuracy is for the MCI class.

Fig. 8 illustrates the distribution of weights w_i for instance weighting the source dataset. There is one mode at about 1.0, which causes an equal weighting of source and target samples. The mode with highest frequency is at about 0.3, which decreases the impact of source samples on the estimation. The third mode is at about 1.9 and therefore roughly doubles the weight of such samples. We see from the distribution that only a small subset of the source data receives high weights and that the impact of the majority of elements is reduced.

To illustrate the variation in the selected features across strategies and the fraction of training data, we report the Bhattacharyya coefficient in Table 4. For each strategy and selected fraction, we computed the probability of features being selected across 50 runs. The Bhattacharyya coefficient is a measure for the amount of overlap, where higher values signify more overlap. For the computation, we selected the weighting strategy with 30% training data as reference. The table shows that there is only a slight variation in the selected features for varying the percentage of training data. The features selected with strategy II are more similar to the weighting than the features selected with strategy III, which is similar to the reported classification accuracy. The features that have been selected most frequently across the 50 runs for different subsets of the AIBL dataset are: hippocampus volume/shape, ventricle volume/shape, amygdala shape, entorhinal thickness, parahippocampal thickness, and middletemporal thickness. These features are mainly consistent with previously reported structures associated to Alzheimer's disease.

4. Discussion

In this work, we present an approach for computer-aided diagnostics of Alzheimer's disease. Our classifier combines volume, thickness, and shape features to obtain an accurate characterization of brain morphology. In particular, the use of neuroanatomical shape descriptors based on the spectrum of the Laplace-Beltrami operator for AD classification is novel. Inclusion of shape information shows great promise, as demonstrated by the high ranking of our approach on the independent test dataset within the scope of the CADDementia challenge. To reduce susceptibility to overfitting, we limit complexity by choosing a linear model. We further construct a compact model by setting sparsity constraints with the mixed ℓ_1/ℓ_2 norm. This automatic model selection approach is advantageous in comparison to manual selection and stepwise refinement as evaluated previously using the Akaike information criterion. Despite these efforts, our results show strong remaining variations in the classification accuracy on the target dataset that are likely not attributed to overfitting. Instead, we demonstrate that differences in source and target data have a substantial impact on classification accuracy, indicating the need for dedicated domain adaptation approaches in

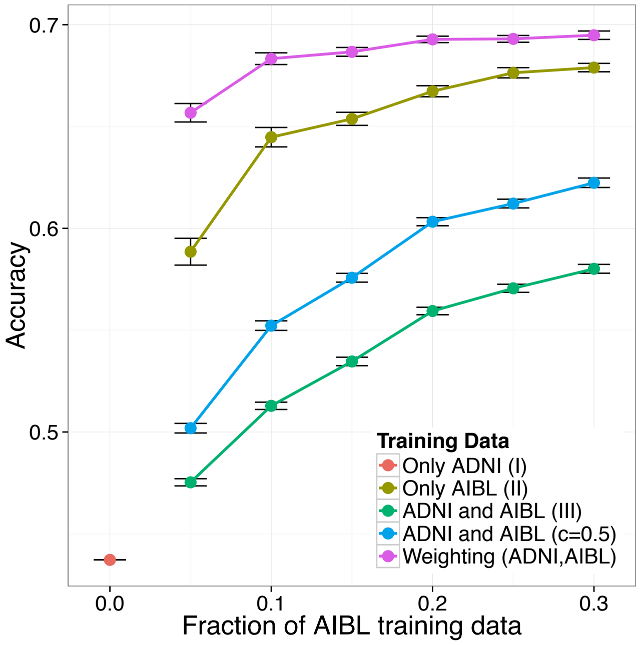

Figure 6: Results for the domain adaptation with ADNI (source) and AIBL (target) data. In addition to the proposed domain adaptation with instance weighting, we evaluate four strategies that use either one of the datasets (I and II), merge both (III), and combine both with $c = 0.5$. Discs show mean classification accuracy over 50 samples and bars indicate standard error. We vary the size of the training dataset from AIBL (x-axis), where the remaining AIBL data is used for testing.

	Fraction of AIBL training data					
	0.05	0.10	0.15	0.20	0.25	0.30
Strategy II	0.76	0.81	0.82	0.83	0.83	0.82
Strategy III	0.57	0.60	0.58	0.58	0.59	0.58
Weighting	0.94	0.98	0.98	0.99	0.99	1.00

Table 4: Bhattacharyya coefficients to express the overlap between selected features for different strategies and the size of the target training data. The weighting strategy with 30% target data is used as reference. Higher values indicate a more similar selection of features than for the reference.

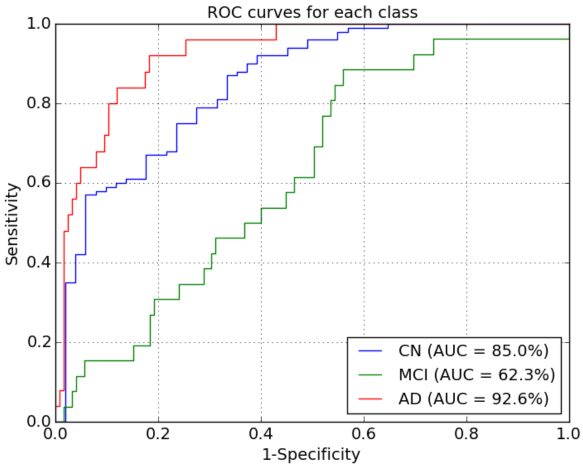

Figure 7: Receiver operating characteristic (ROC) curve for the classification in control (CN), mild cognitive impairment (MCI), and Alzheimer's disease (AD) on the AIBL test data for the weighting strategy with 30% training data.

computer-aided diagnosis.

To address this challenge, we present an approach to domain adaptation based on instance weighting and its integration into the multinomial elastic-net classification. Samples in the source training dataset are weighted according to the ratio of the target and source probabilities. For estimating the target probability, we assume that a subset of the target data is available for training, yielding a supervised approach to domain adaptation. In our experiments, we selected random subsets from the target dataset for training. In the clinical use, best results are to be expected if the target training data is representative of the full target dataset. The weighting can be based on several characteristics of the subject, where we include the diagnostic information, age, sex, and APOE4. This could be extended to additional variables, also including image features. Next to single image features, brain distance functions between subjects could be used as proposed by Gerber et al. (2010) for manifold learning. Low-dimensional embeddings have shown that major variations stem from age and sex differences, so that it needs to be further investigated if brain distances add additional value over directly using age and sex. The proposed instance weighting presents a general framework to domain adaptation, where we derived naïve approaches for combining source and target data by setting the weights to different constants, yielding the selection of either the source, or the target data, or the union of both.

Our results demonstrate that the simple union of source and target data, which we used in the CADDementia challenge, is not the optimal strategy for obtaining a high classification accuracy. Due to the much larger cardinality of the source dataset, the impact of the valuable target training data is lost. Surprisingly, the classifier yields better results when employing only the target training dataset and ignoring the source dataset completely, even if only 5% of the target data is available for training. Optimal results are achieved for the proposed domain adaptation with instance weighting, where source and target training data are combined but the importance of each source sample is weighted based on the target and source distributions.

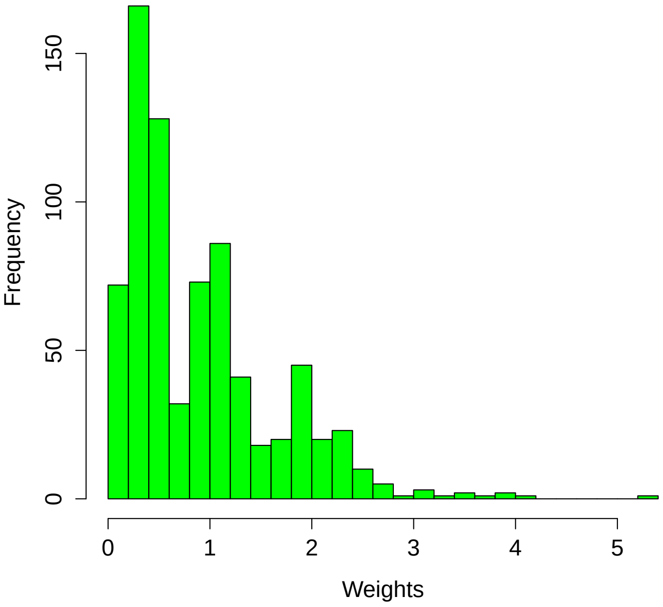

Figure 8: Histogram of weights w_i for weighting source samples. Weights above 1.0 increase the importance of elements in the inference and weights below 1.0 decrease the importance.

As an alternative strategy to instance weighting, we could select source elements with the highest weights and therefore extract a subset of the source dataset that is most similar to the target training data. Such an approach would require an additional threshold parameter to define the cut-off on the weights. The selection of a subset of the source data that matches covariates on the target training data is related to propensity score matching (Rosenbaum and Rubin, 1983).

With respect to classifier performance, we achieve the lowest classification accuracy for the MCI group. This is not surprising as it can be questioned whether MCI forms a separate diagnostic entity. Some of the individuals in this group convert to AD, while others remain stable. MCI is a clinically heterogeneous group of individuals with varying patterns of brain atrophy (Misra et al., 2009). The results of other challenge participants (Bron et al., 2015) demonstrate a similar low performance for this group. In this work, we considered Alzheimer's diagnostics by classifying AD, MCI, and CN subjects. Another important topic is Alzheimer's prognosis, by predicting if and when a subject converts to AD. The early identification of individuals with an elevated risk for developing dementia is of great value for treatments. For predicting the conversion, we face a similar domain adaptation challenge as for the classification, so that the presented strategies may also be of great interest, which remains to be shown in the future.

The algorithm from Sørensen et al. (2014) achieved the highest accuracy in the CADDementia challenge. There are certain similarities to our algorithm; FreeSurfer volume and thickness measurements are used. But also several differences: manual feature selection was used, texture features were integrated, shape features based on surface landmarks were computed, and regularized linear discriminant analysis was used. Further, the training for the challenge was performed on ADNI and AIBL data. Since there are numerous differences between our approach and the one from Sørensen et al. (2014), it is difficult to ascertain what drives the variations in classification accuracy.

Our results provide valuable insights for the optimal implementation of computer-aided diagnosis approaches:

(i) Simply relying on large datasets for training is not sufficient for obtaining a classifier that generalizes well. This is due to the large impact of the data distribution on the estimation of model parameters, since the loss minimization is driven by high density areas.

(ii) Large source datasets are, however, beneficial in supporting the target training set for model estimation. Since large datasets capture a wider portion of the population, it is likely that they also contain instances that are similar to the target samples. These instances can be identified by domain adaptation and emphasized in the model estimation.

(iii) Availability of training data from the target domain is essential for tailoring the model to each target, e.g., an individual hospital. The benefit of supervised domain adaptation is most pronounced for small training fractions of the target data (5%), which may the most realistic scenario when translating a model to the clinic.

5. Conclusions

We highlight the importance of domain adaptation for the classification of Alzheimer's disease and present an approach based on instance weighting. We introduce a classifier based on volume, thickness, and shape features, where the *BrainPrint* is used for the shape representation. A compact model is estimated by regularization of the regression coefficients with the mixed ℓ_1/ℓ_2 norm. This classifier is evaluated on the independent dataset of the CADDementia challenge and used for testing different strategies for domain adaptation. Our results demonstrate that using only data from either the source or target domain, or the union of both, are sub-optimal strategies. We achieved the best results with instance weighting, which compensates for differences in source and target distributions.

6. Acknowledgements

Support for this research was provided in part by the Humboldt foundation, the National Cancer Institute (1K25CA181632-01), the Massachusetts Alzheimer's Disease Research Center (5P50AG005134), the MGH Neurology Clinical Trials Unit, the Harvard NeuroDiscovery Center, the Genentech Foundation, and an NVIDIA hardware award. Further support was provided by the A. A. Martinos Center for Biomedical Imaging (P41RR014075, P41EB015896, U24RR021382), and Shared Instrumentation Grants (1S10RR023401, 1S10RR019307, 1S10RR023043).

Data used in the preparation of this article were obtained from the CADDementia challenge (`http://caddementia.grand-challenge.org`). Data used in the preparation of this article were obtained from the Australian Imaging Biomarkers and Lifestyle Flagship Study of Ageing (AIBL) funded by the Commonwealth Scientific and Industrial Research Organisation (CSIRO), which was made available at the ADNI database (www.loni.ucla.edu/ADNI).

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and 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.

7. References

Adaszewski, S., Dukart, J., Kherif, F., Frackowiak, R., Draganski, B., Initiative, A. D. N., et al., 2013. How early can we predict alzheimer's disease using computational anatomy? Neurobiology of aging 34 (12), 2815–2826.

Adeli-Mosabbeb, E., Thung, K.-H., An, L., Shi, F., Shen, D., 2015. Robust feature-sample linear discriminant analysis for brain disorders diagnosis. In: Advances in Neural Information Processing Systems. pp. 658–666.

Bates, J., Pafundi, D., Kanel, P., Liu, X., Mio, W., 2011. Spectral signatures of point clouds and applications to detection of alzheimer's disease through neuroimaging. In: IEEE International Symposium on Biomedical Imaging. pp. 1851–1854.

Bickel, S., Brückner, M., Scheffer, T., 2007. Discriminative learning for differing training and test distributions. In: Proceedings of the 24th international conference on Machine learning. ACM, pp. 81–88.

Bron, E. E., Smits, M., van der Flier, W. M., Vrenken, H., Barkhof, F., Scheltens, P., Papma, J. M., Steketee, R. M., Orellana, C. M., Meijboom, R., Pinto, M., Meireles, J. R., Garrett, C., Bastos-Leite, A. J., Abdulkadir, A., Ronneberger, O., Amoroso, N., Bellotti, R., Cardenas-Pena, D., Alvarez-Meza, A. M., Dolph, C. V., Iftekharuddin, K. M., Eskildsen, S. F., Coupe, P., Fonov, V. S., Franke, K., Gaser, C., Ledig, C., Guerrero, R., Tong, T., Gray, K. R., Moradi, E., Tohka, J., Routier, A., Durrleman, S., Sarica, A., Fatta, G. D., Sensi, F., Chincarini, A., Smith, G. M., Stoyanov, Z. V., S¿rensen, L., Nielsen, M., Tangaro, S., Inglese, P., Wachinger, C., Reuter, M., van Swieten, J. C., Niessen, W. J., Klein, S., 2015. Standardized evaluation of algorithms for computer-aided diagnosis of dementia based on structural mri: The {CADDementia} challenge. NeuroImage 111, 562 – 579.

Costafreda, S. G., Dinov, I. D., Tu, Z., Shi, Y., Liu, C.-Y., Kloszewska, I., Mecocci, P., Soininen, H., Tsolaki, M., Vellas, B., et al., 2011. Automated hippocampal shape analysis predicts the onset of dementia in mild cognitive impairment. Neuroimage 56 (1), 212–219.

Cuingnet, R., Gerardin, E., Tessieras, J., Auzias, G., Lehéricy, S., Habert, M.-O., Chupin, M., Benali, H., Colliot, O., Initiative, A. D. N., et al., 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.

Dale, A. M., Fischl, B., Sereno, M. I., 1999. Cortical surface-based analysis: I. segmentation and surface reconstruction. Neuroimage 9 (2), 179–194.

Dale, A. M., Sereno, M. I., 1993. Improved localizadon of cortical activity by combining eeg and meg with mri cortical surface reconstruction: A linear approach. Journal of cognitive neuroscience 5 (2), 162–176.

Dickinson, S. J., 2009. Object categorization: computer and human vision perspectives. Cambridge University Press.

Ellis, K. A., Bush, A. I., Darby, D., De Fazio, D., Foster, J., Hudson, P., Lautenschlager, N. T., Lenzo, N., Martins, R. N., Maruff, P., et al., 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. International Psychogeriatrics 21 (04), 672–687.

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. Journal of Alzheimer's Disease.

Ferrarini, L., Frisoni, G. B., Pievani, M., Reiber, J. H., Ganzola, R., Milles, J., 2009. Morphological hippocampal markers for automated detection of alzheimer's disease and mild cognitive impairment converters in magnetic resonance images. Journal of Alzheimer's Disease 17 (3), 643–659.

Fischl, B., Salat, D. H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., van der Kouwe, A., Killiany, R., Kennedy, D., Klaveness, S., Montillo, A., Makris, N., Rosen, B., Dale, A. M., 2002. Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. Neuron 33 (3), 341–355.

Fischl, B., Sereno, M. I., Dale, A. M., 1999a. Cortical surface-based analysis: Ii: Inflation, flattening, and a surface-based coordinate system. Neuroimage 9 (2), 195–207.

Fischl, B., Sereno, M. I., Tootell, R. B., Dale, A. M., et al., 1999b. High-resolution intersubject averaging and a coordinate system for the cortical surface. Human brain mapping 8 (4), 272–284.

Friedman, J., Hastie, T., Tibshirani, R., 2010. Regularization paths for generalized linear models via coordinate descent. Journal of statistical software 33 (1), 1.

Gerardin, E., Chételat, G., Chupin, M., Cuingnet, R., Desgranges, B., Kim, H.-S., Niethammer, M., Dubois, B., Lehéricy, S., Garnero, L., et al., 2009. Multidimensional classification of hippocampal shape features discriminates alzheimer's disease and mild cognitive impairment from normal aging. Neuroimage 47 (4), 1476–1486.

Gerber, S., Tasdizen, T., Fletcher, P. T., Joshi, S., Whitaker, R., 2010. Manifold modeling for brain population analysis. Medical Image Analysis 14 (5), 643 – 653.

Gutman, B. A., Hua, X., Rajagopalan, P., Chou, Y.-Y., Wang, Y., Yanovsky, I., Toga, A. W., Jack Jr, C. R., Weiner, M. W., Thompson, P. M., 2013. Maximizing power to track Alzheimer's disease and MCI progression by LDA-based weighting of longitudinal ventricular surface features. Neuroimage 70, 386–401.

Heimann, T., Mountney, P., John, M., Ionasec, R., 2013. Learning without labeling: Domain adaptation for ultrasound transducer localization. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2013. Springer, pp. 49–56.

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., et al., 2013. Tracking pathophysiological processes in alzheimer's disease: an updated hypothetical model of dynamic biomarkers. The Lancet Neurology 12 (2), 207–216.

Japkowicz, N., Stephen, S., 2002. The class imbalance problem: A systematic study. Intelligent data analysis 6 (5), 429–449.

Jiang, J., Zhai, C., 2007. Instance weighting for domain adaptation in nlp. In: ACL. Vol. 7. pp. 264–271.

Kim, W. H., Singh, V., Chung, M. K., Hinrichs, C., Pachauri, D., Okonkwo, O. C., Johnson, S. C., 2014. Multi-resolutional shape features via non-Euclidean wavelets: Applications to statistical analysis of cortical thickness. NeuroImage 93, 107 – 123.

King, R. D., Brown, B., Hwang, M., Jeon, T., George, A. T., 2010. Fractal dimension analysis of the cortical ribbon in mild Alzheimer's disease. Neuroimage 53 (2), 471–479.

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

Konukoglu, E., Glocker, B., Criminisi, A., Pohl, K. M., 2013. WESD–Weighted spectral distance for measuring shape dissimilarity. Pattern Analysis and Machine Intelligence, IEEE Transactions on 35 (9), 2284–2297.

Lian, Z., Godil, A., Bustos, B., Daoudi, M., Hermans, J., Kawamura, S., Kurita, Y., Lavoué, G., Van Nguyen, H., Ohbuchi, R., et al., 2012. A comparison of methods for non-rigid 3D shape retrieval. Pattern Recognition 46, 449–461.

Margolis, A., 2011. A literature review of domain adaptation with unlabeled data. Rapport Technique, University of Washington, 35.

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. In: CADDementia.

Mwangi, B., Tian, T. S., Soares, J. C., 2014. A review of feature reduction techniques in neuroimaging. Neuroinformatics 12 (2), 229–244.

Niethammer, M., Reuter, M., Wolter, F.-E., Bouix, S., Peinecke, N., Koo, M.-S., Shenton, M., 2007. Global medical shape analysis using the Laplace-Beltrami spectrum. In: Ayache, N., Ourselin, S., Maeder, A. J. (Eds.), MICCAI 2007. Vol. 4791 of LNCS. Springer, Heidelberg, pp. 850–857.

Pan, S. J., Yang, Q., 2010. A survey on transfer learning. Knowledge and Data Engineering, IEEE Transactions on 22 (10), 1345–1359.

Paquerault, S., 2012. Battle against alzheimer's disease: The scope and potential value of magnetic resonance imaging biomarkers.

Academic radiology 19 (5), 509–511.

Quionero-Candela, J., Sugiyama, M., Schwaighofer, A., Lawrence, N. D., 2009. Dataset shift in machine learning. The MIT Press.

Reuter, M., 2006. Laplace Spectra for Shape Recognition. Books on Demand GmbH.

Reuter, M., Niethammer, M., Wolter, F.-E., Bouix, S., Shenton, M., 2007. Global medical shape analysis using the volumetric Laplace spectrum. In: International Conference on Cyberworlds, NASA-GEM Workshop. pp. 417–426.

Reuter, M., Wolter, F.-E., Peinecke, N., 2006. Laplace-Beltrami spectra as "Shape-DNA" of surfaces and solids. Computer-Aided Design 38 (4), 342–366.

Reuter, M., Wolter, F.-E., Shenton, M., Niethammer, M., 2009. Laplace-Beltrami eigenvalues and topological features of eigenfunctions for statistical shape analysis. Computer-Aided Design 41 (10), 739–755.

Rosenbaum, P. R., Rubin, D. B., 1983. The central role of the propensity score in observational studies for causal effects. Biometrika 70 (1), 41–55.

Schlegl, T., Ofner, J., Langs, G., 2014. Unsupervised pre-training across image domains improves lung tissue classification. In: Medical Computer Vision: Algorithms for Big Data. Springer, pp. 82–93.

Shen, K.-k., Fripp, J., Mériaudeau, F., Chételat, G., Salvado, O., Bourgeat, P., 2012. Detecting global and local hippocampal shape changes in alzheimer's disease using statistical shape models. Neuroimage 59 (3), 2155–2166.

Shi, Y., Sha, F., 2012. Information-theoretical learning of discriminative clusters for unsupervised domain adaptation. In: International Conference on Machine Learning. pp. 1079–1086.

Shimodaira, H., 2000. Improving predictive inference under covariate shift by weighting the log-likelihood function. Journal of statistical planning and inference 90 (2), 227–244.

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. In: Proc MICCAI Workshop Challenge on Computer-Aided Diagnosis of Dementia Based on Structural MRI Data. pp. 111–118.

van Opbroek, A., Ikram, M. A., Vernooij, M. W., de Bruijne, M., 2015. Transfer learning improves supervised image segmentation across imaging protocols. IEEE Transactions on Medical Imaging 34 (5), 1018–1030.

Wachinger, C., Batmanghelich, K., Golland, P., Reuter, M., 2014. Brainprint in the computer-aided diagnosis of alzheimer's disease. In: Challenge on Computer-Aided Diagnosis of Dementia, MICCAI.

Wachinger, C., Golland, P., Kremen, W., Fischl, B., Reuter, M., 2015. Brainprint: A discriminative characterization of brain morphology. NeuroImage 109, 232 – 248.

Westman, E., Aguilar, C., Muehlboeck, J.-S., Simmons, A., 2013. Regional magnetic resonance imaging measures for multivariate analysis in Alzheimer's disease and mild cognitive impairment. Brain topography 26 (1), 9–23.

Zhao, M., Chan, R. H., Chow, T. W., Tang, P., 2014. Compact graph based semi-supervised learning for medical diagnosis in alzheimer?s disease. Signal Processing Letters, IEEE 21 (10), 1192–1196.