# SIMPLE DOMAIN ADAPTATION FOR CROSS-DATASET ANALYSES OF BRAIN MRI DATA

*Christoph Hofer*[1], *Roland Kwitt*[1], *Yvonne Höller*[3,4], *Eugen Trinka*[2,3,4], *Andreas Uhl*[1] *and ADNI*

[1]Department of Computer Science, University of Salzburg, Austria
[2]Spinal Cord Injury & Tissue Regeneration Centre Salzburg, Paracelsus Medical University, Salzburg, Austria,
[3]Department of Neurology, Christian Doppler Medical Centre, Paracelsus Medical University, Salzburg, Austria,
[4]Centre for Cognitive Neuroscience, Paracelsus Medical University, Salzburg, Austria

## ABSTRACT

We consider the problem of domain shift in analyses of brain MRI data. While many different datasets are publicly available, most algorithms are still trained on a single dataset and often suffer the problem of limited and unbalanced sample sizes. In this work, we propose a surprisingly simple strategy to reduce the impact of domain shift – caused by different data sources and processing pipelines – that typically occurs in cross-dataset analyses. We experimentally evaluate our approach on the problem of using volumetric features to distinguish neurodegenerative diseases and report results using three datasets in two practically relevant scenarios: (1) cross-dataset learning and (2) leveraging pre-trained classifiers across different datasets. We show that our adaptation technique enables both scenarios with performance close to the single-dataset case.

## 1. INTRODUCTION

Structural magnetic resonance imaging (MRI) is one of todays primary imaging modalities in the diagnosis and study of neurodegenerative diseases, such as Alzheimer's (AD) (see, e.g., [1]). Throughout the past decade, both the number and the sizes of open-access datasets has grown substantially. Similarly, there is an abundance of automated methods to process this data, ranging from registration and segmentation to strategies for automated diagnosis of diseases. With the broad availability of computational power and the fact that many processing pipelines can be trivially parallelized on an image level, computationally demanding approaches (such as multi-atlas segmentation [2]) have become practical on a large scale and facilitate to leverage machine learning techniques and full-scale data mining strategies.

Since learning algorithms improve with the amount of available training data, it would be beneficial if we could leverage and combine already pre-processed data from multiple sources, possibly subjected to different processing pipelines. This raises the fundamental question, in which manner these datasets should be combined, such that differences along these processing pipelines have little effect.

In the context of MR brain image analysis, a substantial amount of different features for various tasks is at our disposal; see, e.g., [3] for an overview of the state-of-the-art in the classification of dementia diseases. Despite the diversity of features, the majority of approaches share a common component, namely the automated labeling of predefined brain regions. In the spirit of leveraging publicly available data, it seems reasonable to assume that this step is already completed and the results are available, e.g., in the form of already segmented regions. While, in this case, we apparently have no control over the segmentation process, we still have to demand a consistent label acquisition protocol (e.g., [4]); otherwise comparability would, a-priori, not be given.

**Related Work.** In the context of obtaining segmentations of the human brain, there has been a substantial amount of work devoted to multi-atlas segmentation strategies, subsuming steps such as registration, atlas selection, and label fusion, just to name a few; see [2] for a survey of these techniques. In the study of neuro-degenerative diseases, many strategies and features to discriminate between healthy controls and patients at various stages of mild cognitive impairment (MCI) and AD [5, 6, 7, 8, 9] have also been proposed. A well-established baseline is to investigate *volume changes*, i.e., cellular atrophy in predefined regions of the brain, such as parts of the temporal lobe which are indicative of AD and MCI, resp. [5, 6, 7, 8]. In other works, e.g., [9], features are extracted by observing the center of mass of different brain regions and by evaluating their connectivity, or network structure [10]. Nevertheless, the problem of cross-dataset learning in the context of neuro-degenerative diseases has gained little attention so far. While, in [11], the authors do assess cross-dataset performance of various features for random-forest based AD classification, the study primarily focuses on the impact of changes in specific parameters of the processing chain, such as different choices of atlases or feature selection/reduction strategies. In our work, we investigate a strategy to mitigate the effects of such differences, irrespective of its particular type.

**Practical considerations.** Lets consider two processing pipelines $P_1$ and $P_2$, where $P_i$ subsumes all processing steps required to obtain features that can later be used by a machine

learning algorithm. $P_i$ might include data acquisition, (multi-atlas) segmentation, and feature extraction. In other words, $P_i$ can be interpreted as some sort of mapping from the space of subjects to some feature space $\mathcal{F}$. As already mentioned, there is a broad variety of possible differences between $P_1$ and $P_2$. Hence, in most constellations $P_1(s) \neq P_2(s)$, for some subject $s$, even if the type of extracted feature is the same. This prevents us from directly fusing data in $\mathcal{F}$ to increase sample size for training some learning algorithm. However, under the assumption that there is some mapping $\varphi$, such that $P_1(s) = \varphi\big(P_2(s)\big)$, it is desirable to know as much as possible about $\varphi$, since this would facilitate a direct fusion of data, processed by $P_1$ and $P_2$. Without loss of generality, we let $S_1, S_2$ be two sets of subjects, each separated into two groups $A$ and $B$, i.e., $S_i = S_{i,\mathrm{A}} \cup S_{i,\mathrm{B}}$. For example, $S_1$ and $S_2$ could be two datasets from different hospitals containing *healthy controls* (A) as well as *non-healthy* (B) subjects. If subject data has already been processed, by $P_1$ and $P_2$, our goal is to work with $P_1(S_1)$ and $P_2(S_2)$, although large parts of the parametrization of $P_i$ might be unknown. Intuitively, to study (and possibly correct for) differences in $P_1$ and $P_2$, we could assess differences in $P_1(s)$ and $P_2(s)$ for some $s \in S_1 \cap S_2$ in order to learn more about $\varphi$. Unfortunately, however, a common situation in medical imaging is that $S_1 \cap S_2 = \emptyset$, since data was acquired by different sources. Hence, a direct comparison of the same subject under both $P_1$ and $P_2$ cannot be done. This raises the question of how (and to what extent) $\varphi$ can be approximated. In this work, we study that problem in a *classification setting*.

**Setup**. Given two datasets, $\mathtt{D} = P_1(S_1)$ and $\overline{\mathtt{D}} = P_2(S_2)$ in the aforementioned setting of $S_1 \cap S_2 = \emptyset$, we use statistical invariants to characterize the so called *domain shift* [12]. We assume that both datasets have *at least one common group*. This is realistic, since healthy controls are typically available in most cases, however, the number of controls might be (very) limited. Let $\mathtt{D}_\mathtt{h}$, $\overline{\mathtt{D}}_\mathtt{h}$ be those groups. We further assume that any reasonable processing pipeline will not change the (relative) statistical behavior of these groups in $\mathcal{F}$ with respect to samples of non-healthy subjects. Under this assumption (which we will assess later), we aim to find $\hat{\varphi}$ such that $\mathtt{D}_\mathtt{h} \sim \hat{\varphi}\,(\overline{\mathtt{D}}_\mathtt{h})$. We evaluate this approach on two scenarios, which reflect two common situations in practice.

*Scenario I (Control group substitution).* As gathering data for an upcoming study of some disease is typically a resource intensive procedure, it is desirable to reuse existing data from previous studies. This is particularly true in situations where data from patients with some disease is already available (as this data might be collected as part of clinical routine), but data from control patients is only available in a small amount. To avoid bias in classification results, typically we require a balanced control-to-disease ratio. Hence, in this scenario, we assume availability of a large (external) dataset of healthy controls, denoted as $\mathtt{H}$, and a large unbalanced dataset $\mathtt{D} = \mathtt{D}_\mathtt{h} \cup \mathtt{D}_\mathtt{d}$, where the number of diseased subjects, $|\mathtt{D}_\mathtt{d}|$, is far bigger than $|\mathtt{D}_\mathtt{h}|$. Given an estimated $\hat{\varphi}$, we could train on the combined dataset $\mathtt{H} \cup \hat{\varphi}(\mathtt{D})$.

*Scenario II (Pre-trained classifiers).* Here, we consider two balanced datasets $\mathtt{D} = \mathtt{D}_\mathtt{h} \cup \mathtt{D}_\mathtt{d}$ and $\overline{\mathtt{D}} = \overline{\mathtt{D}}_\mathtt{h} \cup \overline{\mathtt{D}}_\mathtt{d}$, with $|\mathtt{D}| >> |\overline{\mathtt{D}}|$. In this situation, it would be desirable to train a stable and well-generalizing classifier on the larger dataset $\mathtt{D}$ and later use it on $\hat{\varphi}(\overline{\mathtt{D}})$ *directly*.

## 2. ANALYSIS FRAMEWORK

We first have to define a common feature space $\mathcal{F}$, which is the endpoint of each processing pipeline that we consider in this work. While, in principle, our framework is agnostic to the particular type of feature that is used, we focus on volumetric changes in brain regions as an indicator for (neurodegenerative) diseases, as this is among the customary techniques in previous works. Specifically, we take a holistic view on volumetric changes over a collection of brain regions, i.e., we define a *relational volumetric signature (RVS)*

$$\mathcal{V}(s) := \frac{\big(\lambda(R_1), \ldots, \lambda(R_N)\big)}{||\big(\lambda(R_1), \ldots, \lambda(R_N)\big)||_1} \ , \tag{1}$$

where $\lambda(\cdot)$ denotes the natural volume in $\mathbb{R}^3$ and $R_i$ are the brain regions. This feature is motivated by the fact that volume relations between regions can provide useful information, especially if a disease does not influence every region $R_i$ in the same way. By taking the $L_1$ norm, we minimize the distorting influence of body/head size and project the volume vector to the positive face of the $N$-dimensional unit simplex. To simplify any subsequent statistical analysis [13], we further define a slightly modified RVS as

$$\tilde{\mathcal{V}}(s) := \left(\log\left(\frac{\lambda(R_1)}{\lambda(R_N)}\right), \ldots, \log\left(\frac{\lambda(R_{N-1})}{\lambda(R_N)}\right)\right) \ . \tag{2}$$

In summary, $\tilde{\mathcal{V}}$ is derived from $\mathcal{V}$ by smoothly stretching the positive face of the $N$-unit simplex to $\mathbb{R}^{N-1}$.
We remark that for specific neuro-degenerative diseases, a reduction of the feature vector to some particular regions may improve classification accuracy. However, in the sense of investigating a generic approach for handling the domain shift between MRI brain datasets, we decided to use all regions.

**Dataset(s).** We use three datasets in our study: two publicly-available datasets and one third-party dataset. Each dataset has its own characteristics which reflects the scenarios outlined in Sec. 1. *First*, `IXI`[1] contains brain MR images for a collection of (569) *healthy controls*; we use the T1-weighted scans in our experiments. *Second*, we use an already preprocessed portion of the Alzheimer's Disease Neuroimaging Initiative (ADNI), denoted as `ADNI`. This dataset contains

[1] http://brain-development.org/ixi-dataset

| | Control | LMCI | MCI | AD | TLE |
|---|---|---|---|---|---|
| IXI | 569 | – | – | – | – |
| ADNI | 284 | 178 | 307 | 226 | – |
| PMU | 22 | – | 19 | – | 17 |

**Table 1:** Meta-information for the datasets used in this study.

T1-weighted scans from 995 subjects, automatically segmented using MAPER [14]. We use all subjects labeled as healthy controls, MCI and AD and refer to these subsets as ADNI−CON, ADNI−MCI and ADNI−AD, resp. Our *third* dataset, denoted as PMU, is a small collection of T1-weighted scans of healthy controls, MCI and temporal lobe epilepsy (TLE) patients. For our purposes, we only use the subjects labeled as healthy controls or MCI, denoted as subsets PMU−CON and PMU−MCI. While ADNI is already pre-segmented, we processed IXI and PMU via a multi-atlas segmentation pipeline, using the same 30 atlases as in [14] (to guarantee a consistent labeling protocol). In detail, we performed brain extraction[2], and diffeomorphically[3] registered all atlases to each image. Meta information about the datasets is listed in Table 1. For all subjects from the three datasets, we obtain a parcellation of the brain into $N = 83$ regions. Consequently, $\mathcal{F} \subset \mathbb{R}^{82}$, as $\tilde{\mathcal{V}}(\cdot)$ uses the last coordinate for standardization. With respect to our setup of Sec. 1, $\mathtt{H} = \mathtt{IXI}$, $\mathtt{D}_\mathrm{h} \subset \mathtt{ADNI{-}CON}$ and $\overline{\mathtt{D}}_\mathrm{h} \subset \mathtt{PMU{-}CON}$.

**Assessing domain shift.** To assess domain shift, we first visualize the group of healthy controls of each dataset in feature space $\mathcal{F}$ via canonical 2D projections, an example is shown in Fig. 1. Qualitatively, the visualizations indicate that all three datasets are approximately Gaussian distributed in $\mathcal{F}$, but have different means and covariance structures. Since it is hard to test for multivariate normality in high-dimensional spaces, we conducted a series of univariate Kolmogorov-Smirnov tests on the marginal distributions at 5% significance (with less than 10% rejections).

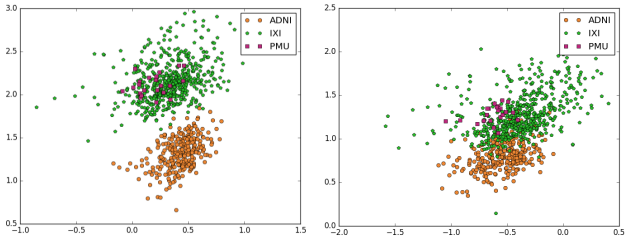


**Fig. 1:** Exemplary (canonical) 2D-projections of $\tilde{\mathcal{V}}(\cdot)$ for healthy controls. *We chose dimensions with obvious domain shift.*

This shift in the distribution of the data has an undesirable effect on classification as it is, e.g., possible to separate IXI from ADNI−CON with a simple discriminant classifier, such as an SVM. Nevertheless, according to our tests for normality and a first qualitative visual assessment, the distributions appear similar. To reduce the impact of the mean, it seems reasonable to consider mean-corrected data, i.e., $\mathtt{IXI'} = \mathtt{IXI} - \widehat{\mu}(\mathtt{IXI})$ and $\mathtt{ADNI'} = \mathtt{ADNI} - \widehat{\mu}(\mathtt{ADNI{-}CON})$, where $\widehat{\mu}$ is the sample mean. We then fitted a multivariate Gaussian $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ with diagonal covariance to the data, i.e., $\mathbf{\Sigma} = \mathrm{diag}([\sigma_1, \ldots, \sigma_N])$, and let $M_{\mathtt{IXI'}}$ and $M_{\mathtt{ADNI'}}$ be the corresponding models, estimated from IXI′ and ADNI′, respectively. This simple model can be easily estimated, especially in consideration of the sample sizes (cf. Table 1) and the dimensionality ($N = 82$) of the data.

To quantify our initial assumption (cf. Sec. 1) that different groups in the data show similar statistical behavior, regardless of the domain shift, we considered the (average) squared Mahalanobis distance

$$d_M(X) := \frac{1}{|X|} \sum_{\mathbf{x} \in X} (\mathbf{x} - \boldsymbol{\mu}_M)^T \mathbf{\Sigma}_M^{-1} (\mathbf{x} - \boldsymbol{\mu}_M). \quad (3)$$

of the control, MCI and AD group from ADNI with respect to the two Gaussian models, i.e., $M_{\mathtt{IXI'}}$ and $M_{\mathtt{ADNI{-}CON'}}$. The results, listed in Table 2, reveal that the relative behavior of the diseased groups to the controls is indeed similar, regardless of the particular model.

| | AD′/CON′ | MCI′/CON′ |
|---|---|---|
| $d_{M_{\mathtt{ADNI{-}CON'}}}$ | 1.65 | 1.32 |
| $d_{M_{\mathtt{IXI'}}}$ | 1.71 | 1.36 |

**Table 2:** (Average) of relative squared Mahalanobis distances.

**Correcting domain shift.** Considering the results of Table 2, we conjecture that a translation followed by a scaling in $\mathcal{F}$ is a sufficient approximation $\widehat{\varphi}$ of $\varphi$ such that $\widehat{\varphi}(\overline{\mathtt{D}}_\mathrm{h}) \sim \mathtt{D}_\mathrm{h}$. Without loss of generality, we set $\mathbb{E}[\mathtt{D}_\mathrm{h}] = \mathbf{0}$. Consequently, our estimate for the mean shift $t$ is the negative sample mean of $\overline{\mathtt{D}}_\mathrm{h}$. Hence, let $\widehat{\varphi}_t(x) := x - \widehat{\mu}(\overline{\mathtt{D}}_\mathrm{h})$ and thus $\mathtt{D}_\mathrm{h} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_\mathrm{h})$ and $\widehat{\varphi}_t(\overline{\mathtt{D}}_\mathrm{h}) \sim \mathcal{N}(\mathbf{0}, \overline{\mathbf{\Sigma}}_\mathrm{h})$. Next, we need to find $\lambda \in \mathbb{R}$ to define $\widehat{\varphi}_s(\mathbf{x}) := \lambda \mathbf{x}$, such that the distribution of $\widehat{\varphi_s} \circ \widehat{\varphi_t}(\overline{\mathtt{D}}_\mathrm{h})$ is similar to the distribution of $\mathtt{D}_\mathrm{h}$. To quantify similarity, we reuse Eq. (3), which is proportional to the average likelihood, i.e.,

$$\underbrace{\frac{1}{|\overline{\mathtt{D}}_\mathrm{h}|} \sum_{\mathbf{x} \in \overline{\mathtt{D}}_\mathrm{h}} (\lambda \cdot \mathbf{x})^\top \mathbf{\Sigma}_\mathrm{h}^{-1} (\lambda \cdot \mathbf{x})}_{= d_{M_\mathrm{h}}(\lambda \cdot \overline{\mathtt{D}}_\mathrm{h})} = \underbrace{\frac{1}{|\mathtt{D}_\mathrm{h}|} \sum_{\mathbf{x} \in \mathtt{D}_\mathrm{h}} \mathbf{x}^\top \mathbf{\Sigma}_\mathrm{h}^{-1} \mathbf{x}}_{= d_{M_\mathrm{h}}(\mathtt{D}_\mathrm{h})}. \quad (4)$$

where $M_\mathrm{h}$ is the model estimated from $\mathtt{D}_\mathrm{h}$ with covariance $\mathbf{\Sigma}_\mathrm{h}$. Hence, our estimate for $\lambda$ follows as

$$\lambda = \sqrt{d_{M_\mathrm{h}}(\mathtt{D}_\mathrm{h}) \cdot d_{M_\mathrm{h}}(\overline{\mathtt{D}}_\mathrm{h})^{-1}} \ . \quad (5)$$

In other words, $\lambda$ is determined such that the average likelihood of $\lambda \cdot \mathtt{ADNI{-}CON'}$ (a translated + scaled version of ADNI−CON) equals the average likelihood of IXI′ with respect to $M_{\mathtt{IXI'}}$. Equivalently, the covariance of the estimated

[2] using BET, available from http://fsl.fmrib.ox.ac.uk
[3] using ANTS, available from http://stnava.github.io/ANTs

Gaussian could be transformed by $\lambda^{-2}$. This would lift the adaptation from $\mathcal{F}$ to the model space where it can be interpreted as moving the model along a geodesic [15] until the mentioned criterion is fulfilled. With this *model adaptation* in mind, our approach can be extended to Fisher vectors [16]. We evaluated this approach and observed performance on par with our technique. We also remark that under non-Gaussian distributed data, we could leverage the approximative power of a Gaussian mixture model (GMM) combined with the Bayesian adaptation technique of [17, 18], to extended our approach to handle this scenario. The simplicity of our data, even in the 82-dim. space, does not support the use of GMMs, though, which tend to be unstable to estimate given limited samples sizes.

## 3. EXPERIMENTAL STUDY

*First*, it is imperative to understand the behavior of a classifier (SVM with RBF kernel, parameters not optimized) *without* domain shift. For that purpose, we only train and test on ADNI. Table 3 lists the results of this reference experiment with an increasing number of training samples (for AD and MCI vs. healthy controls; equally-sized samples per class, remaining samples used for testing). Not surprisingly, small sample sizes in the training data have a strong negative effect on the classification accuracy.

| $N_{\text{train}}$ | 5 | 10 | 15 | 50 | 100 | 150 |
|---|---|---|---|---|---|---|
| AD / CON | 65.4 | 67.7 | 69.2 | 75.8 | 79.9 | 82.5 |
| MCI / CON | 56.7 | 58.8 | 59.9 | 62.8 | 64.6 | 65.8 |

**Table 3:** Classification accuracy as a function of the number of training subjects per group.

In a *second* experiment, we assess whether our classifier is sensitive to the domain shift, i.e., the difference in distribution of ADNI−CON and IXI. For example, if we randomly partition ADNI−CON into two groups, we obtain the *expected* result of $\approx$ 50% (100 samples each, 100 runs). However, distinguishing IXI/ADNI−CON succeeds with an accuracy of $\approx$ 100% (150 samples each, 100 runs). This *artificial separability* is indicative of the domain shift. *Third*, we evaluate the impact of the size of the adaptation training set $\mathtt{T}_{\mathtt{adapt}}$ (i.e., the small subset of ADNI−CON used to estimate $\widehat{\varphi}$). Table 4 lists the results in separating IXI from $\widehat{\varphi}$(ADNI−CON). While we do not reach the desired accuracy

| $\lvert\mathtt{T}_{\mathtt{adapt}}\rvert$ | 2 | 5 | 10 | 15 | 30 | 60 |
|---|---|---|---|---|---|---|
| IXI / $\widehat{\varphi}$(ADNI−CON) | 98.8 | 92.1 | 81.8 | 74.8 | 67.3 | 61.6 |

**Table 4:** Accuracies for separating controls, IXI vs. ADNI−CON, *after adaptation* (std. dev. ranges from 1.3 to 5.7).

of 50%, we argue that $\widehat{\varphi}$ sufficiently reduces the distance between IXI and $\widehat{\varphi}$(ADNI−CON), even if $\lvert\mathtt{T}_{\mathtt{adapt}}\rvert$ is small.

**Scenario I .** To simulate *Scenario I* from Sec. 1, we create an unbalanced dataset by randomly choosing subjects from ADNI−CON, denoted as $\mathtt{T}_{\mathtt{adapt}}$, 150 from ADNI−AD and 150 from ADNI−MCI, resp., for training. We use $\mathtt{T}_{\mathtt{adapt}}$ to estimate $\widehat{\varphi}$ and map ADNI−AD and ADNI−MCI to the domain of IXI. Interestingly, as we can see from Table 5, classification accuracy becomes similar to our *reference* experiment of Table 3, even for small samples (colouring).

| $\lvert\mathtt{T}_{\mathtt{adapt}}\rvert$ | 2 | 5 | 10 | 15 | 60 |
|---|---|---|---|---|---|
| IXI / $\widehat{\varphi}$(ADNI−AD) | 53.8 | 71.8 | 78.8 | 80.8 | 81.3 |
| IXI / $\widehat{\varphi}$(ADNI−MCI) | 50.8 | 58.3 | 64.7 | 66.2 | 66.7 |

**Table 5:** Impact of adaptation set sizes on classification (std. dev. ranges from 1.8 to 5.9).

**Scenario II .** To assess *Scenario II* of Sec. 1, we use all data available from ADNI−CON and ADNI−MCI to train a SVM. Small subsets, denoted as $\mathtt{T}_{\mathtt{adapt}}$, from PMU−CON are then used to estimate $\widehat{\varphi}$ and balanced test sets are randomly drawn from the remaining subjects of PMU. Finally, we inject these test sets via $\widehat{\varphi}$ into the domain of ADNI, and try to classify them. Results are listed in Table 6 (top).

| $\lvert\mathtt{T}_{\mathtt{adapt}}\rvert$ | 2 | 5 | 10 | 15 |
|---|---|---|---|---|
| $\widehat{\varphi}$(MCI) / $\widehat{\varphi}$(CON) | $59.1_{\pm 6.4}$ | $60.58_{\pm 6.5}$ | $62.2_{\pm 7}$ | $64.2_{\pm 10.8}$ |
| $N_{\text{train}}$ | 2 | 5 | 10 | 15 |
| MCI / CON | $54.3_{\pm 15.5}$ | $56.0_{\pm 16.5}$ | $57.1_{\pm 15.9}$ | $55.5_{\pm 14.9}$ |

**Table 6:** Impact of adaptation set sizes drawn from PMU−CON on classification (*top*) in relation to the impact of training set size when using just PMU data (*bottom*).

As a reference, Table 6 (bottom) lists the accuracy we get if we train *only* on PMU with balanced training sets of size $N_{\text{train}} = 2 \cdot \lvert\mathtt{T}_{\mathtt{adapt}}\rvert$ per class. We remark that reasonable classification on PMU appears impossible, given only PMU data for training. However, under adaptation, classification performance on PMU becomes similar to the ADNI-only case (Table 3), demonstrating the utility of our approach.

## 4. DISCUSSION

Our experiments show that it is possible to combine different MRI brain datasets using a surprisingly simple strategy, if volumetric features are used. Our approach performs adaptation in feature space *directly*, but might as well be interpreted as *model adaptation* (cf. Sec. 2). Most notably, we can largely mitigate the negative effect of domain shift, observed in previous studies (e.g., [11]). While more complex models might allow for further compensation, estimation issues due to sample size become a problem. Our model, on the other hand, is simple and can be easily estimated. Nevertheless, additional

complexity might be needed in situations where the data cannot be modeled by single Gaussians.

## 5. ACKNOWLEDGEMENTS

This work was partially funded by the Austrian Science Fund FWF (KLI project 00012).

## 6. REFERENCES

[1] G.B. Frisoni, N.C. Fox, C.R. Jack Jr., P. Scheltens, and P.M. Thompson, "The clinical use of structural MRI in Alzheimer disease," *Nat. Rev. Neurol.*, vol. 6, no. 2, pp. 67–77, 2010.

[2] J.E. Iglesias and M.R. Sabuncu, "Multi-Atlas Segmentation of Biomedical Images: A survey," *Med. Image Anal.*, vol. 24, no. 1, pp. 205 – 219, 2015.

[3] E.E. Bron et al., "Standardized evaluation of algorithms for computer-aided diagnosis of dementia based on structural MRI: The CADDementia challenge," *NeuroImage*, vol. 111, pp. 562 – 579, 2015.

[4] A. Hammers, R. Allom, M.J. Koepp, S.L. Free, R. Myers, L. Lemieux, T.N. Mitchell, D.J. Brooks, and J.S. Duncan, "Three-dimensional maximum probability atlas of the human brain, with particular reference to the temporal lobe," *Hum. Brain Mapp.*, vol. 19, no. 4, pp. 224–247, 2003.

[5] L.G. Apostolova, A.E. Green, S. Babakchanian, K.S. Hwang, Y.Y. Chou, A.W. Toga, and P.M. Thompson, "Hippocampal atrophy and ventricular enlargement in normal aging, mild cognitive impairment (MCI), and Alzheimer disease," *Alzheimer Dis. Assoc. Disord.*, vol. 26, no. 1, pp. 17–27, 2012.

[6] C.D. Smith, A.H. Andersen, B.T. Gold, and ADNI, "Structural brain alterations before mild cognitive impairment in adni: Validation of volume loss in a predefined antero-temporal region," *J. Alzheimers Dis.*, vol. 31 (Suppl. 3), pp. 49–58, 2012.

[7] I. Driscoll, C. Davatzikos, Y. An, X. Wu, D. Shen, M. Kraut, and S.M. Resnick, "Longitudinal pattern of regional brain volume change differentiates normal aging from MCI," *Neurology*, vol. 72, no. 22, pp. 1906–1913, Jun 2009.

[8] S.G. Mueller, N. Schuff, K. Yaffe, C. Madison, B. Miller, and M.W. Weiner, "Hippocampal atrophy patterns in mild cognitive impairment and Alzheimer's disease," *Hum. Brain Mapp.*, vol. 31, no. 9, pp. 1339–1347, 2010.

[9] L. Lillemark, L. Sørensen, A. Pai, E. Dam, M. Nielsen, and ADNI, "Brain region's relative proximity as marker for Alzheimer's disease based on structural MRI," *BMC Medical Imaging*, vol. 14, no. 1, pp. 1–12, 2014.

[10] A. Alexander-Bloch, J.N. Giedd, and E. Bullmore, "Imaging structural co-variance between human brain regions," *Nat. Rev. Neurosci.*, vol. 14, no. 5, pp. 322–336, 2013.

[11] A.V. Lebedev, E. Westman, G.J.P. Van Westen, M.G. Kramberger, A. Lundervold, D. Aarsland, H. Soininen, I. Kloszewska, P. Mecocci, M. Tsolaki, B. Vellas, S. Lovestone, A. Simmons, and ADNI, "Random forest ensembles for detection and prediction of Alzheimer's disease with a good between-cohort robustness," *NeuroImage: Clinical*, vol. 6, pp. 115–125, 2014.

[12] H. Daumé III and D. Marcu, "Domain adaptation for statistical classifiers," *JAIR*, vol. 26, no. 1, pp. 101–126, 2006.

[13] J. Aitchison and S. M. Shen, "Logistic-normal distributions: Some properties and uses," *Biometrika*, vol. 67, no. 2, pp. 261–272, 1980.

[14] R.A. Heckemann, S. Keihaninejad, P. Aljabar, K.R. Gray C. Nielsen, D. Rueckert, J.V. Hajnal, and A. Hammers, "Automatic morphometry in Alzheimer's disease and mild cognitive impairment," *NeuroImage*, vol. 56, no. 4, pp. 2024 – 2037, 2011.

[15] S. I. R. Costa, Sandra A. Santos, and J. E. Strapasson, "Fisher information distance: A geometrical reading," *Discrete Applied Mathematics*, vol. 197, pp. 59–69, 2015.

[16] F. Perronnin and C. Dance, "Fisher kernels on visual vocabularies for image categorization," in *CVPR*, 2007.

[17] D. Reynolds, T.F. Quatieri, and R.B. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models," *Digital Signal Processing*, vol. 10, no. 13, pp. 19–41, 2000.

[18] M. Dixit, N. Rasiwasia, and N. Vasconcelos, "Adapted gaussian models for image classification," in *CVPR*, 2011.