Analyzing hierarchical multi-view MRI data with StaPLR: an application to Alzheimer's disease classification

Loon, W. van; Vos, F. de; Fokkema, M.; Szabo, B.; Koini, M.; Schmidt, R.; Rooij, M. de

Analyzing hierarchical multi-view MRI data with StaPLR: An application to Alzheimer's disease classification

Wouter van Loon[1,*], Frank de Vos[1,2,3], Marjolein Fokkema[1], Botond Szabo[4], Marisa Koini[5], Reinhold Schmidt[5], and Mark de Rooij[1,3]

[1]Department of Methodology and Statistics, Leiden University, Wassenaarseweg 52, 2333 AK Leiden, the Netherlands
[2]Department of Radiology, Leiden University Medical Center, Albinusdreef 2, 2333 ZA Leiden, the Netherlands
[3]Leiden Institute for Brain and Cognition, Albinusdreef 2, 2333 ZA Leiden, the Netherlands
[4]Department of Mathematics, VU Amsterdam, De Boelelaan 1111, 1081 HV Amsterdam, the Netherlands
[5]Department of Neurology, Division of Neurogeriatrics, Medical University of Graz, Auenbruggerplatz 22, A-8036 Graz, Austria
[*]Corresponding author: Wouter van Loon, w.s.van.loon@fsw.leidenuniv.nl

November 18, 2021

Abstract

Multi-view data refers to a setting where features are divided into feature sets, for example because they correspond to different sources. Stacked penalized logistic regression (StaPLR) is a recently introduced method that can be used for classification and automatically selecting the views that are most important for prediction. We introduce an extension of this method to a setting where the data has a hierarchical multi-view structure. We also introduce a new view importance measure for StaPLR, which allows us to compare the importance of views at any level of the hierarchy. We apply our extended StaPLR algorithm to Alzheimer's disease classification where different MRI measures have been calculated from three scan types: structural MRI, diffusion-weighted MRI, and resting-state fMRI. StaPLR can identify which scan types and which MRI measures are most important for classification, and it outperforms elastic net regression in classification performance.

keywords *multimodal MRI, machine learning, stacked generalization, penalized regression, feature selection*

1 Introduction

In biomedical research, the integration of data from different sources into a single classification model is becoming increasingly common [1, 2]. This is fueled by the increasing availability of multi-source data, for example through the UK Biobank [3, 4], the Alzheimer's Disease Neuroimaging Initiative (ADNI) [5], and various dementia registries around the world [6] such as the Prospective Registry on Dementia (PRODEM) [7]. Training a model on multiple data sources has been found to increase accuracy in the prediction of brain-age [8] and the classification of Alzheimer's disease (AD) [9, 10, 11].

A general term for data in which the features are divided into feature sets (for example, by source or modality) is *multi-view data*, and the field of developing algorithms for such data is known as *multi-view (machine) learning* [12, 13]. Of particular interest to this study is the multi-view learning framework known as *multi-view stacking* [11, 14, 15]. The general idea of multi-view stacking is to first train a model on each feature set (also called a *view*) separately. Then, each of these models is cross-validated to obtain a set of predictions of the outcome. Finally, another algorithm (called the *meta-learner*) is trained on these cross-validated predictions. The meta-learner thus learns how to best combine the predictions from the view-specific models. Several methods used in previous neuroscience studies can be considered a form of multi-view stacking, where they performed better than single-view or non-stacked approaches [8, 10, 11, 16, 17, 18, 19, 20]. However, these methods are generally ad-hoc approaches tailored specifically to the data at hand, and there is little consistency between the used methods.

Most earlier research has focused on improving classification accuracy but it is also important to identify which views are relevant for prediction. For example, if a certain scan modality turns out to be irrelevant for prediction of a disease, it may not have to be measured at all. Recently, a variant of multi-view stacking called *stacked penalized logistic regression* (StaPLR) has been developed specifically for this purpose [15]. StaPLR essentially integrates the penalized logistic regression models which are already commonly used in neuroimaging classification, such as ridge regression [21, 22] and the lasso [23, 24], into a single unified multi-view stacking methodology. StaPLR can be used to select the feature sets that are most relevant for prediction, and has been shown to have several advantages over earlier methods, including a decreased false positive rate in view selection and a large reduction in computation time, while maintaining good classification accuracy [15]. In addition StaPLR is, to our knowledge, the only multi-view learning method which can be extended to a hierarchical multi-view structure with an arbitrary number of levels while keeping computational feasibility. By hierarchical multi-view data we mean that feature sets are nested in other feature sets. Consider, for example, data collected from different domains, such as genetics, neuroimaging, and cognitive tests. Each of these domains could be considered a different view of the patients under consideration. These views could then be further divided into subsets. For example, the higher-level neuroimaging feature set could be further divided into lower-level feature sets corresponding to different scan types.

In this study we will show a proposed extension of the StaPLR method and its

application to an Alzheimer's disease classification problem based on three MRI scan types: structural MRI, diffusion-weighted MRI, and resting state functional MRI. For each of these scan types, different MRI measures were computed, where each measure is represented by multiple features. This yields a hierarchical multi-view structure with three levels: the *features* (base level) are nested in the *MRI measures* (intermediate level), which in turn are nested in the different *scan types* (top level). Parts of this multi-view data set, which consists of data collected as part of PRODEM [7] and the Austrian Stroke Prevention Study (ASPS) [25, 26], have been used in previous studies [9, 27, 28], but this is the first time these features are all included into a single analysis. Previous applications of StaPLR have focused solely on a setting with two levels [15, 29]. In this paper we will therefore adapt StaPLR to the hierarchical structure of the data. We will show how StaPLR can be used to both perform classification and identify the views that are most important for prediction. To provide a 'benchmark' for the classification performance and interpretability of the model we additionally perform logistic elastic net regression [30], which is a method that has been used in many previous multi-view neuroimaging classification studies [9, 16, 31, 32, 33, 34].

In addition to its advantages in view selection and computation time [15], the proposed extension of StaPLR has important advantages in terms of the interpretability of the resulting classifier. First, measures of view and feature relevance are readily available in the form of coefficients in a logistic regression model. This is in contrast to previous multi-view stacking methods focused on prediction accuracy, such as those using random forests as a meta-learner [8, 19]. Second, extending StaPLR to match the hierarchical multi-view structure of the data allows us to calculate such measures of importance *at each level of the hierarchy*. Thus, we can easily obtain estimates of the contribution of each scan type, but also of each MRI measure within those scan types. Finally, we show in section 2.4.1 how the proposed extension of StaPLR allows us to compare the contribution of different MRI measures even if they correspond to different scan types.

The application to the current data set aims to provide an example of a more general class of applications within neuroimaging and biomedical science as a whole. Since our focus is on demonstrating the methodology rather than on the specific data set, we will refrain from any interpretation regarding the specific clinical meaning of our findings with respect to the target population of those who originally collected the data.

2 Materials and methods

2.1 Participants

Our data set consisted of 76 patients clinically diagnosed with probable AD, and 173 cognitively normal elderly controls, for a total of 249 observations. The AD patients were scanned at the Medical University of Graz as part of PRODEM [7]. The elderly controls were scanned at the same scanning site, with the same scanning protocol, and over the same time period as part of the ASPS [25, 26]. We only included patients for which anatomical MRI, diffusion MRI and rs-fMRI were available.

2.2 MRI analysis

The scanning protocols, and an elaborate description of the MRI analyses are provided in the supplementary materials. For each scan type, several MRI measures were computed; below we provide a brief description. An overview of the features included in our analyses is presented in Table 1.

The strutural MRI scans were used to calculate five different MRI measures. Grey matter density refers to the percentage of grey matter in a certain area of the brain. The 48 features correspond to the 48 regions of the probabilistic Harvard-Oxford cortical atlas [35]. Subcortical volumes describe the size of several subcortical brain structures. The 14 features correspond to the thalamus, caudate, putamen, pallidium, hippocampus, amygdala and accumbens of both hemispheres. The neocortex was parcellated into the 68 regions of the Desikan-Killiany atlas [36]. For each of these regions, the mean cortical thickness, mean cortical curvature, and the total area of the region's cortical surface ("cortical area") was calculated.

The diffusion-weighted MRI scans were used to calculate fractional anisotropy, mean diffusivity, axial diffusivity, and radial diffusivity for the 20 white matter regions of the JHU white-matter tractography atlas [37].

The resting state fMRI scans were used to calculate multiple types of functional connectivity (FC) estimates. First, temporal concatenation independent component analysis (ICA) was used to extract both 20 and 70 components. For both of these configurations, FC matrices were calculated using either full or sparse partial correlations, resulting in four different FC matrices for each participant. These four matrices were further used to calculate FC dynamics using a sliding window approach. The FC matrices were calculated for each time window, and the standard deviation of these matrices over time reflect the FC dynamics. In addition, the sliding window matrices of all participants were clustered into five connectivity states using k-means clustering. The number of sliding window matrices assigned to each of these five states was calculated for each participant. The four FC matrices were also used to calculate several common graph metrics. Additionally, voxel-wise connectivity with 10 different resting state networks was calculated using dual regression, as well as seed based connectivity with both the left and right hippocampus as seed regions. Furthermore, a voxel-wise eigenvector centrality map was calculated. Eigenvector centrality attributes a value to each voxel in the brain such that a voxel receives a large value if it is strongly correlated with many other voxels that are themselves central within the network. Lastly, the amplitude of low frequency fluctuations (ALFF), and its weighted variant the fractional ALFF (fALFF), were calculated for each voxel. Details can be found in the supplementary materials and in de Vos et al. [16].

(s) scan type	(v) MRI measure	number of features
(1) structural MRI	(1) grey matter density	48
	(2) subcortical volumes	14
	(3) cortical thickness	68
	(4) cortical area	68
	(5) cortical curvature	68
(2) diffusion MRI	(6) fractional anisotropy	20
	(7) mean diffusivity	20
	(8) axial diffusivity	20
	(9) radial diffusivity	20
(3) resting state fMRI	(10) full FC correlation matrix (20 × 20)	190
	(11) full FC correlation matrix (70 × 70)	2,415
	(12) sparse partial FC correlation matrix (20 × 20)	190[*]
	(13) sparse partial FC correlation matrix (70 × 70)	2,415[*]
	(14) SD of full FC matrix (20 × 20)	190
	(15) SD of full FC matrix (70 × 70)	2,415
	(16) SD of sparse partial FC matrix (20 × 20)	190
	(17) SD of sparse partial FC matrix (70 × 70)	2,415[*]
	(18) FC states of full FC matrix (20 × 20)	5
	(19) FC states of full FC matrix (70 × 70)	5
	(20) FC states of sparse partial FC matrix (20 × 20)	5
	(21) FC states of sparse partial FC matrix (70 × 70)	5
	(22) Graph metrics of full FC matrix (20 × 20)	124
	(23) Graph metrics of full FC matrix (70 × 70)	424
	(24) Graph metrics of sp. par. FC matrix (20 × 20)	124[*]
	(25) Graph metrics of sp. par. FC matrix (70 × 70)	424[*]
	(26) FC with visual network 1	190,981
	(27) FC with visual network 2	190,981
	(28) FC with visual network 3	190,981
	(29) FC with default mode network	190,981
	(30) FC with the cerebellum	190,981
	(31) FC with sensorimotor network	190,981
	(32) FC with auditory network	190,981
	(33) FC with executive control network	190,981
	(34) FC with frontoparietal network 1	190,981
	(35) FC with frontoparietal network 2	190,981
	(36) FC with left hippocampus	190,981
	(37) FC with right hippocampus	190,981
	(38) Fast eigenvector centrality mapping	190,981
	(39) ALFF	190,981
	(40) fALFF	190,981
total		2,876,515[**]

Table 1: Overview of the scan types, MRI measures, and corresponding number of features used in this study. The indices s and v are used to refer to the different scan types and MRI measures in Algorithm 1. [*]Some of these features were removed due to having a variance of zero; the total number of features after removal for each of these MRI measures is 189, 2337, 2414, 123 and 423, respectively. [**]The removal of features due to zero variance is already reflected in this total.

2.3 Logistic elastic net regression (benchmark)

To provide a reference value for the accuracy and area under the receiver operating characteristic curve (AUC) we performed logistic elastic net regression [30]. Elastic net regression employs a mixture of L_1 and L_2 penalties on the vector of regression coefficients [30]. The L_1 penalty can perform feature selection by setting some coefficients to zero, while the L_2 penalty encourages groups of correlated features to be selected together. Elastic net regression operates at the level of the features and thus ignores the multi-view structure of the data completely. Elastic net regression has two tuning parameters, one which determines the mixture of L_1 and L_2 penalties (α), and one which determines the amount of penalization (λ). We selected a value for both penalties through 10-fold cross-validation. To prevent overfitting, we assessed the models classification performance using a double (nested) cross-validation approach [38]: an inner loop is used to determine the values of the tuning parameters, and an outer loop is used for determining classification accuracy and AUC. For both the inner and outer loop we used 10 folds. Additionally, we repeat this nested cross-validation approach 10 times to average out the effects of different allocations of the subjects to the folds. Elastic net regression was performed in R 4.0.2 [39], using the package `glmnet` 1.9-8 [24].

Since the elastic net ignores the multi-view structure of the data, it is hard to infer the importance of a MRI measure or scan type. After all, a single MRI measure can be represented by anywhere from 5 to over 190,000 regression coefficients. With a total of over 2,8 million features, showing the results for each feature individually is infeasible. In order to summarize the results at the MRI measure level we calculated for each measure the following: (1) the number of non-zero coefficients, and (2) the L_2-norm (i.e. the square root of the sum of squared values) of the associated regression coefficient vector.

2.4 Stacked penalized logistic regression

From each of the three scan types several MRI measures are derived. In turn, each MRI measure consists of multiple features, as shown in Table 1. We therefore apply an extension of the StaPLR algorithm to 3 levels (Algorithm 1). We start by training a logistic ridge regression model on each of the 40 MRI measures under consideration (line 1). For each of these models we use 10-fold cross-validation to choose an appropriate value for the penalty parameter. The reason we use ridge regression at this step is that we are not interested in selecting individual features, only entire MRI measures. We refer to the classifiers that were obtained for each of the 40 MRI measures as $\hat{f}_1, \ldots, \hat{f}_{40}$. Since these classifiers are probabilistic, they give predicted values in $[0, 1]$.

For each scan type s, we want to obtain an intermediate classifier $\hat{f}^{(s)}_{\text{inter}}$ that combines the predictions of the classifiers trained on the corresponding MRI measures. For example, for the structural MRI scan type, we want to obtain an intermediate classifier $\hat{f}^{(1)}_{\text{inter}}$ that combines the predictions of $\hat{f}_1$ through $\hat{f}_5$, which are the classifiers corresponding to grey matter density, subcortical volumes, cortical thickness, cortical area, and cortical curvature. In order to train such a combination model, we need a vector of predictions for each of the classifiers $\hat{f}_1$ through $\hat{f}_5$. We could simply use the fitted values for each of

these classifiers, but this would yield overly optimistic estimates of predictive accuracy, because the same data would be used for fitting the model and generating predictions. Instead, we would like to obtain a vector of estimated out-of-sample predictions. We obtain such estimates through 10-fold cross-validation (line 2). We divide the observations into 10 folds, train each classifier on 9 folds, then generate predictions for the observations in the left-out fold. We repeat this procedure to obtain predictions for each of the folds. Note that "training the classifier" includes the selection of penalty parameters; the cross-validation loop used to select the penalty parameter is nested within the loop used to generate the predictions. This means the predictions are truly made on data which the classifier has never seen.

Once we obtained a vector of cross-validated predictions for each of the 40 classifiers, we collect them into 3 separate matrices, one for each scan type (line 3). These matrices then become the training data for the next step in the StaPLR algorithm, where we train a nonnegative logistic lasso model on each of the 3 matrices of predictions (line 4). We apply the nonnegative lasso at this step because we would like to select a subset of the available MRI measures. The nonnegativity constraints have previously been shown to improve performance; see van Loon et al. [15] for empirical evidence and theoretical support. We end up with 3 intermediate classifiers, one for each scan type.

In order to train the meta-learner, we need to obtain a vector of estimated out-of-sample predictions for each of the 3 intermediate classifiers. We again do this using 10-fold cross-validation (line 5). We then collect these in another matrix (line 6), and train another logistic nonnegative lasso model on this matrix (line 7). The model tranining is now complete, and the final stacked classifier can be used by applying the classifiers $\hat{f}_1, \ldots, \hat{f}_{40}$ to the 40 MRI measures, aggregating their predictions for each scan type using the intermediate classifiers $\hat{f}^{(1)}_{\text{inter}}$, $\hat{f}^{(2)}_{\text{inter}}$, $\hat{f}^{(3)}_{\text{inter}}$, and then combining the output of each scan type using the meta-classifier $\hat{f}_{\text{meta}}$ (line 8).

StaPLR was performed in R using the package `multiview` 0.3.1 [40]. The scripts used for model fitting and evaluation are available in a public code repository [41] For a more general discussion of the original StaPLR algorithm we refer to van Loon et al. [15].

Algorithm 1: StaPLR with 3 levels, as applied to the current data set

Data: $\boldsymbol{X}^{(v)}$, $v = 1 \ldots 40$, the 40 different MRI measures as shown in Table 1, and $\boldsymbol{y}$ the binary outcome variable, where a value of 1 indicates probable Alzheimer disease, and a value of 0 indicates a healthy control

1 Train a logistic ridge regression classifier (including cross-validation for λ) on the pairs $(\boldsymbol{X}^{(v)}, \boldsymbol{y})$, $v = 1, \ldots, 40$, to obtain view-specific classifiers $\hat{f}_1, \ldots, \hat{f}_{40}$.

2 Apply 10-fold cross-validation to obtain a vector of predictions $\boldsymbol{z}^{(v)}$ for each of the $\hat{f}_v$, $v = 1, \ldots, 40$.

3 For each of the three scan types $s = 1, 2, 3$, collect the predictions $\boldsymbol{z}^{(v)}$ which correspond to that scan type column-wise into the matrix $\boldsymbol{Z}^{(s)}$.

4 Train a logistic nonnegative lasso classifier (including cross-validation for λ) on the pairs $(\boldsymbol{Z}^{(s)}, \boldsymbol{y})$, $s = 1, 2, 3$, to obtain the intermediate classifiers $\hat{f}^{(1)}_{\text{inter}}$, $\hat{f}^{(2)}_{\text{inter}}$, $\hat{f}^{(3)}_{\text{inter}}$.

5 Apply 10-fold cross-validation to obtain a vector of predictions $\boldsymbol{z}^{(s)}_{\text{inter}}$ for each of the $\hat{f}^{(1)}_{\text{inter}}$, $\hat{f}^{(2)}_{\text{inter}}$, $\hat{f}^{(3)}_{\text{inter}}$.

6 Collect the predictions $\boldsymbol{z}^{(1)}_{\text{inter}}$, $\boldsymbol{z}^{(2)}_{\text{inter}}$, $\boldsymbol{z}^{(3)}_{\text{inter}}$ column-wise into the matrix $\boldsymbol{Z}_{\text{inter}}$.

7 Train a logistic nonnegative lasso classifier (including cross-validation for λ) on the pair $(\boldsymbol{Z}_{\text{inter}}, \boldsymbol{y})$ to obtain a meta-classifier $\hat{f}_{\text{meta}}$.

8 Define the final stacked classifier as:

$$\begin{aligned}\hat{f}_{\text{stacked}}(\boldsymbol{X}) \quad = \quad & \hat{f}_{\text{meta}}(\hat{f}^{(1)}_{\text{inter}}(\hat{f}_1(\boldsymbol{X}^{(1)}), \ldots, \hat{f}_5(\boldsymbol{X}^{(5)})), \\ & \hat{f}^{(2)}_{\text{inter}}(\hat{f}_6(\boldsymbol{X}^{(6)}), \ldots, \hat{f}_9(\boldsymbol{X}^{(9)})), \\ & \hat{f}^{(3)}_{\text{inter}}(\hat{f}_{10}(\boldsymbol{X}^{(10)}), \ldots, \hat{f}_{40}(\boldsymbol{X}^{(40)}))).\end{aligned}$$

2.4.1 Quantifying feature set relevance across scan types

One of the advantages of StaPLR is that at each level the method fits a logistic regression model. Thus, at each level, the classifiers can be interpreted as regular logistic regression models. This way one can easily determine the relative importance of the different scan types, or the different MRI measures within a scan type. However, if one wishes to compare feature sets corresponding to different scan types, for example an MRI measure from structural and one from functional MRI, an issue arises. Because the models used at each level are logistic regression models, and the logistic function is nonlinear, the final stacked classifier cannot be obtained by simply multiplying the regression weights of the different levels. If one wishes to compare the relative importance of feature sets across the different scan types, a different approach is needed.

In StaPLR, at the base level a separate classifier is trained on each MRI measure separately. Consider as an analogy a human committee: each base-level classifier can

be considered a member of a committee providing a prediction of the outcome. The intermediate classifiers and meta-classifier then assign weights to the predictions of the committee members and combine them into a single predicted outcome. Now consider the possibility that one committee member makes a different prediction than all the others. In human committees such a dissenting opinion is sometimes called a *minority report* [42, 43]. We can measure the impact of such a minority report by quantifying how the final predicted outcome changes as a single member changes its prediction, while the predictions of all the other members are kept constant. We will call this quantification the *minority report measure* (MRM). Since in our case, each committee member is a classifier trained on a specific MRI measure, the MRM can be considered a measure of importance of this MRI measure in determining the final prediction.

The MRM measures the change in the outcome of the stacked classifier when the prediction corresponding to the ith MRI measure derived from scan type s changes from value a to value b, while all other predictions are kept constant at value c. Different choices for a, b and c are possible. In our analysis, we choose $a = 0$ and $b = 1$, which are the theoretical minimum and maximum, and $c = \bar{y}$, which is the proportion of observations corresponding to class 1. In this case the MRM measures the maximum possible change in final prediction attributable to the view $\boldsymbol{X}^{[s,i]}$, while the predictions corresponding to all other MRI measures are set to the sample mean of y. Other possible choices for a and b include the empirical minimum and maximum, respectively.

In the context of StaPLR applied to the current data set, the MRM can be formalized as follows: Denote by $\boldsymbol{X}^{[s,i]}, i = 1 \dots m_s,\ s = 1 \dots S$, the ith MRI measure of scan type s, with m_s the total number of measures corresponding to scan type s. Denote by $\hat{\beta}_0^{[s]}, s = 1 \dots S$ the intercept of the intermediate classifier corresponding to scan type s. Denote by $\hat{\beta}_i^{[s]}, i = 1 \dots m_s, s = 1 \dots S$ the coefficient of the ith MRI measure of scan type s. Denote by $\hat{\omega}_0$ the intercept of the meta-classifier, and by $\hat{\omega}^{[s]}, s = 1 \dots S$ the weight of scan type s. Then for the ith MRI measure corresponding to scan type s, we define the MRM as:

$$\text{MRM}(\boldsymbol{X}^{[s,i]}, a, b, c) = g(\boldsymbol{X}^{[s,i]}, b, c) - g(\boldsymbol{X}^{[s,i]}, a, c), \tag{1}$$

with $a, b, c \in [0, 1]$, $b > a$, and

$$g(\boldsymbol{X}^{[s,i]}, b, c) = \psi\left(\hat{\omega}_0 + \hat{\omega}_s\psi\left(\hat{\beta}_0^{[s]} + \hat{\beta}_i^{[s]}b + \sum_{j\neq i}\hat{\beta}_j^{[s]}c\right) + \sum_{k\neq s}\hat{\omega}_k\psi\left(\hat{\beta}_0^{[k]} + \sum_{j=1}^{m_k}\hat{\beta}_j^{[k]}c\right)\right), \tag{2}$$

where ψ denotes the logistic function, i.e.

$$\psi(x) = \frac{\exp(x)}{1 + \exp(x)}. \tag{3}$$

Note that, given a, b and c, the value of the MRM depends only on the estimated parameters of the stacked model. The MRM can thus be readily calculated without any need for resampling or refitting of the model, unlike many model-agnostic measures of feature importance such as permutation feature importance [44, 45] or SHAP values [46].

2.5 Data and code availability statement

The source code of the R package 'multiview' is publicly available [40], as are the R scripts used for model fitting and evaluation [41]. The MRI data used in this study is available upon direct request to R. Schmidt; a formal data sharing agreement is mandatory.

2.6 Ethics statement

The original PRODEM study was approved by the ethics committees of the Medical University of Graz, the Medical University of Innsbruck, the Medical University of Vienna, the Konventhospital Barmherzige Brüder Linz, the Province of Upper Austria, the Province of Lower Austria and the Province of Carinthia [7]. Written informed consent was obtained from all patients and their caregivers [7]. For collection of the original ASPS data, standard protocol approvals, registrations, and patient consents were obtained [26]. The study was approved by the standard ethics committee of the Medical University of Graz for experiments using human participants, and written informed consent was obtained from all study participants [26]. Permission to use the data for the current study was obtained from M. Koini and R. Schmidt at the Medical University of Graz.

3 Results

3.1 Elastic net regression

The mean AUC of the model was 0.922 (SD = 0.008). The mean test accuracy of the model was 0.848 (SD = 0.012). The selected value of the tuning parameter α varied from 0.2 to 1, with an average of 0.788. On average, the model contained 168.17 (SD = 113.72) features. On average, the selected features were spread out over 24.07 (SD = 3.15) different views. Thus, elastic net regression provides classifiers which are fairly sparse at the feature level, but not at the MRI measure level. Consider Figure 1, which shows the distribution of the number of selected features for each MRI measure across the 10×10 fitted models. It can be observed that among the MRI measures with the largest median number of selected features are those which correspond to the voxel-wise functional connectivity with various RSNs (measures (v) 26 through 37). For all of these measures, the median number of selected features is greater than zero. It should be noted that these measures, along with ALFF and fALFF, are also those which contain by far the largest number of features. Each of these measures contains over 190,000 features, but the median number of selected features from each of them is typically around 5 to 15 (see also Figure 1).

We can therefore raise several issues with the output of the elastic net regression. Elastic net regression tends to select a small number of features among a large number of MRI measures. This is not very useful from a data collection point of view, since one would typically collect or calculate an entire MRI measure. For example, one would have to perform the process of calculating the RSNs through ICA regardless of how many

features were selected among measures 26 through 37. It is also not very useful from the viewpoint of model interpretation, since the functional connectivity of a resting state network with a handful of voxels scattered throughout the brain is unlikely to be very informative to a clinician. Additionally, comparing Figure 1 with Table 1 shows that the MRI measures with the largest number of selected features are also the measures which contain the largest number of features to begin with. However, these are not necessarily the most important measures for predicting the outcome. Thus, given two views which are similarly predictive of the outcome, a view with a much larger number of features will likely have a much larger number of selected features.

Elastic net regression does not provide a direct measure of the importance of an MRI measure, since it operates at the feature rather than the view level. However, we can obtain a measure of the importance of an MRI measure by calculating the L_2-norm (i.e. the square root of the sum of squared values) of the corresponding regression coefficient vector. The results are shown in Figure 2, where it can be observed that it is actually the structural MRI measures of grey matter density and cortical thickness which have the largest L_2-norm. Although Figures 1 and 2 allow us to summarize the outcome at the MRI measure level, it is difficult to use the results of the elastic net regression to draw conclusions about which MRI measure is the most important for classification, or which MRI measures do not need to be measured in the future, since at least some features were selected from a large number of measures. Furthermore, in order to draw conclusions about the different scan types, one would have to re-aggregate the results at that level.

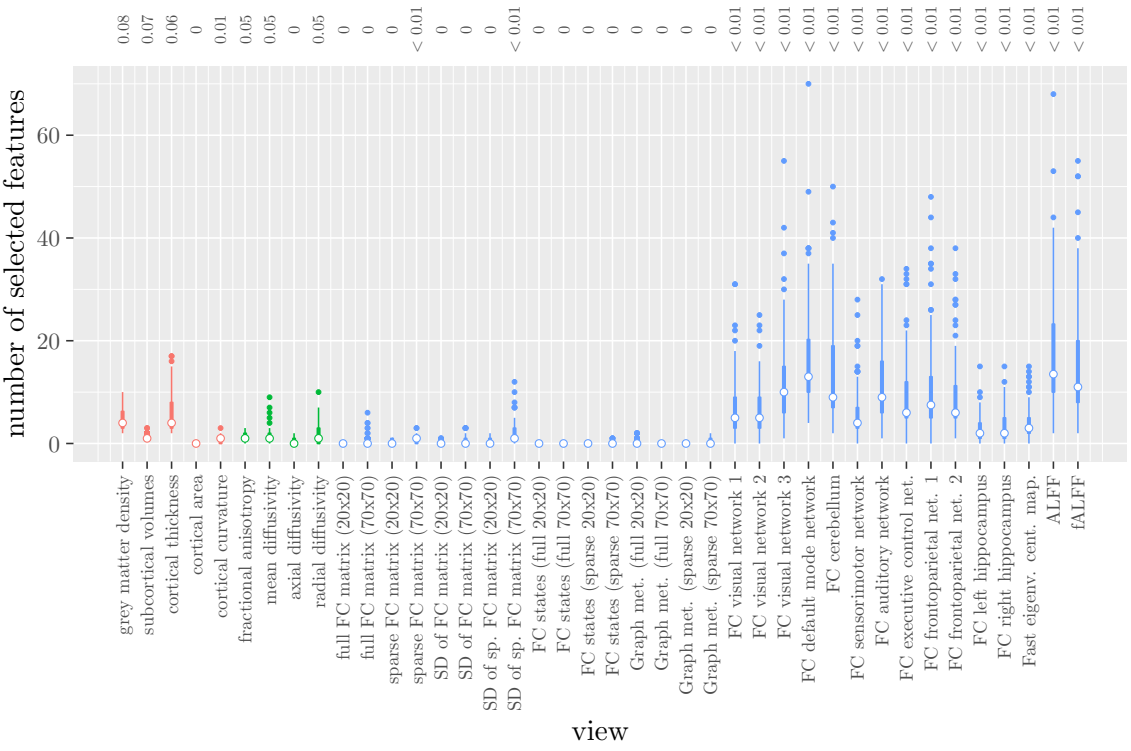

Figure 1: Boxplots of the number of selected features for each MRI measure resulting from the elastic net regression, colored by scan type (red = structural MRI, green = diffusion MRI, blue = resting state fMRI). The numbers at the top of the graph denote the proportion of times the coefficient was nonzero.

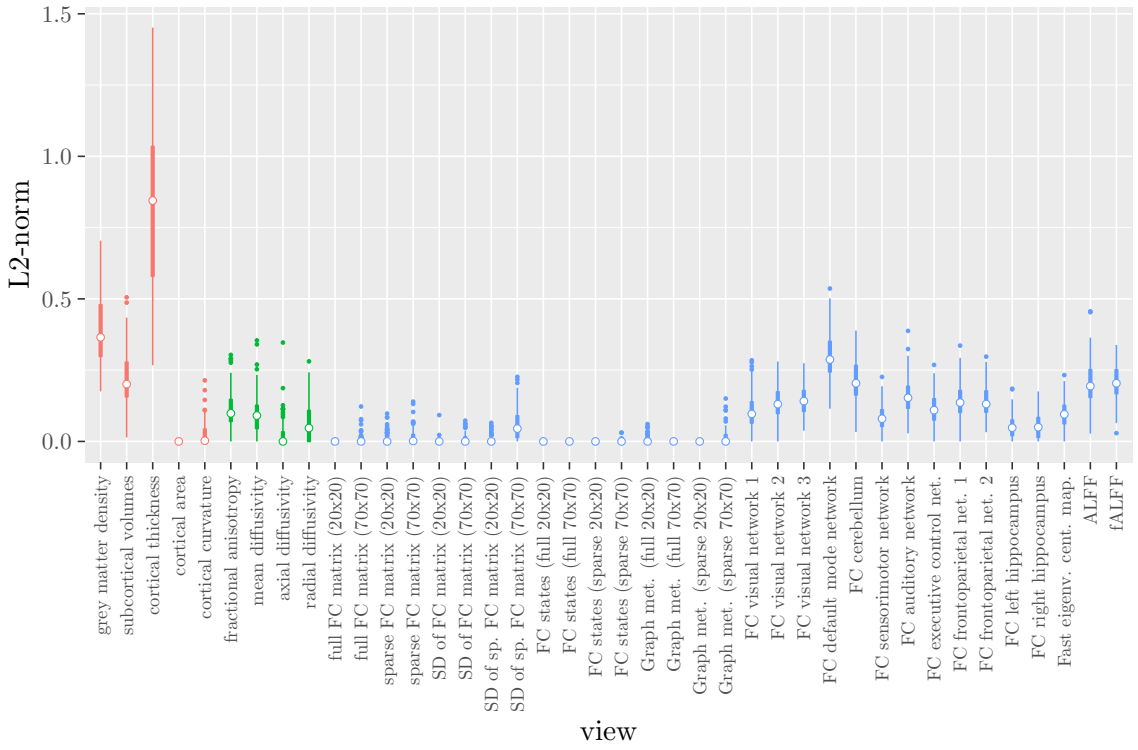

Figure 2: Boxplots of the L_2-norm of the regression coefficient vector for each MRI measure resulting from the elastic net regression, colored by scan type (red = structural MRI, green = diffusion MRI, blue = resting state fMRI).

3.2 StaPLR

The mean AUC of StaPLR was 0.942 (SD = 0.006), which was higher than that of the elastic net (0.922, SD = 0.008). The mean accuracy was 0.893 (SD = 0.008), which was also higher than that of the elastic net (0.848, SD = 0.012). Across the 10×10 fitted stacked classifiers, the structural scan was selected 100% of the time, the diffusion weighted scan 83% of the time, and the RS-fMRI scan 90% of the time. The median regression coefficient for each scan type, as well as their distribution across the 10×10 fitted stacked classifiers can be observed in Figure 3. These are simply the regression coefficients in a logistic regression classifier. The input to this classifier is the output of the classifiers corresponding to each scan type, which are all predicted probabilities between zero and one. Taking the median values shown in Figure 3, the final predicted probability of Alzheimer's disease is given by an intercept plus 5.12 times the prediction from structural MRI, plus 1.02 times the prediction from diffusion-weighted MRI, plus 1.25 times the prediction from resting state fMRI. The final classification is thus largely determined by the classifier corresponding to the structural scan, with smaller contributions from the diffusion-weighted and resting state fMRI scans. The contribution of each MRI measure within a given scan type can be compared in the same way. Figure 4 shows that within the structural MRI scan type, the measures of cortical thickness and grey matter density contributed the most to the prediction. Subcortical volumes provided a much smaller contribution, and was not always selected. Cortical curvature was generally not selected and only provided a small contribution in 5% of the fitted classifiers, while cortical area was never selected. Figures 5 and 6 show the contributions of the measures within the diffusion-weighted and resting state fMRI scan types, respectively.

One important thing to consider when interpreting a StaPLR model with more than two levels, is that the coefficients shown in Figures 4, 5, and 6, are coefficients of three different intermediate classifiers. Thus, we cannot simply compare coefficients across these figures. Doing so would lead us to conclude that ALFF (median coefficient of 4.05) is more important than grey matter density (median coefficient of 3.48). However, this would be an erroneous conclusion, since the structural scan type has a much larger weight than the resting state functional scan type (see Figure 3). To compare MRI measures across the different scan types we can use the minority report measure (MRM) introduced in Section 2.4.1. Because the MRM measures the effect of the MRI measure-specific models on the final predicted outcome, it is suitable for comparing the importance of MRI measures even if they correspond to different scan types. We calculated the MRM for each measure, for each of the repetitions. As shown in Figure 7, the MRM properly reflects the high importance of the structural scan type compared with the diffusion and functional scan types.

If we compare the results of StaPLR with the results of elastic net regression (Figures 1 and 2), we can observe both similarities and differences. In terms of the overall importance of the different scan types, the results are similar, with the structural MRI providing the MRI measures with the largest contribution, both in StaPLR (in terms of MRM and meta-level regression coefficient) and in the elastic net (in terms of the L_2

norm of the regression coefficients). In terms of the MRI measures within the structural scan type, the results are also similar, with cortical thickness being the most important measure, followed by grey matter density, and lastly subcortical volumes. The fact that both methods agree on the same MRI measures being the most important for the classification of Alzheimer's disease provides somewhat of a 'sanity check' in terms of the models providing sensible results. Within the scan types which have a smaller contribution, i.e. diffusion-weighted MRI and resting state fMRI, we can see some differences between the methods. For example, in StaPLR mean diffusivity is not considered important, while it is of some importance in the elastic net model. The largest differences, however, are seen within the functional scan type. StaPLR generally included only 4 resting state fMRI measures, whereas elastic net generally included features from 17 fMRI measures. Features from ALFF are included by both methods. Although the StaPLR model is much sparser in terms of the MRI measures which are included, this did not lead to a reduction in accuracy. In fact, the accuracy of the StaPLR model compares favorably to that of the elastic net.

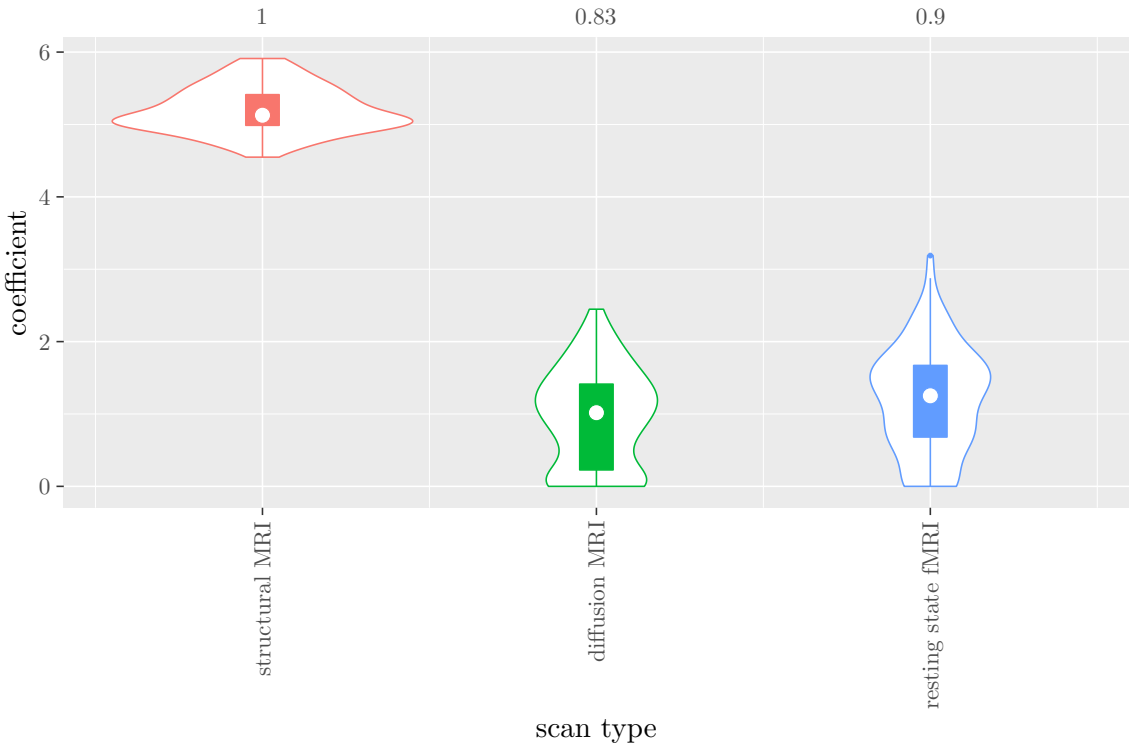

Figure 3: Box-and-violin plots of the StaPLR meta-level regression coefficients for each scan type. The numbers at the top of the graph denote the proportion of times the coefficient was nonzero.

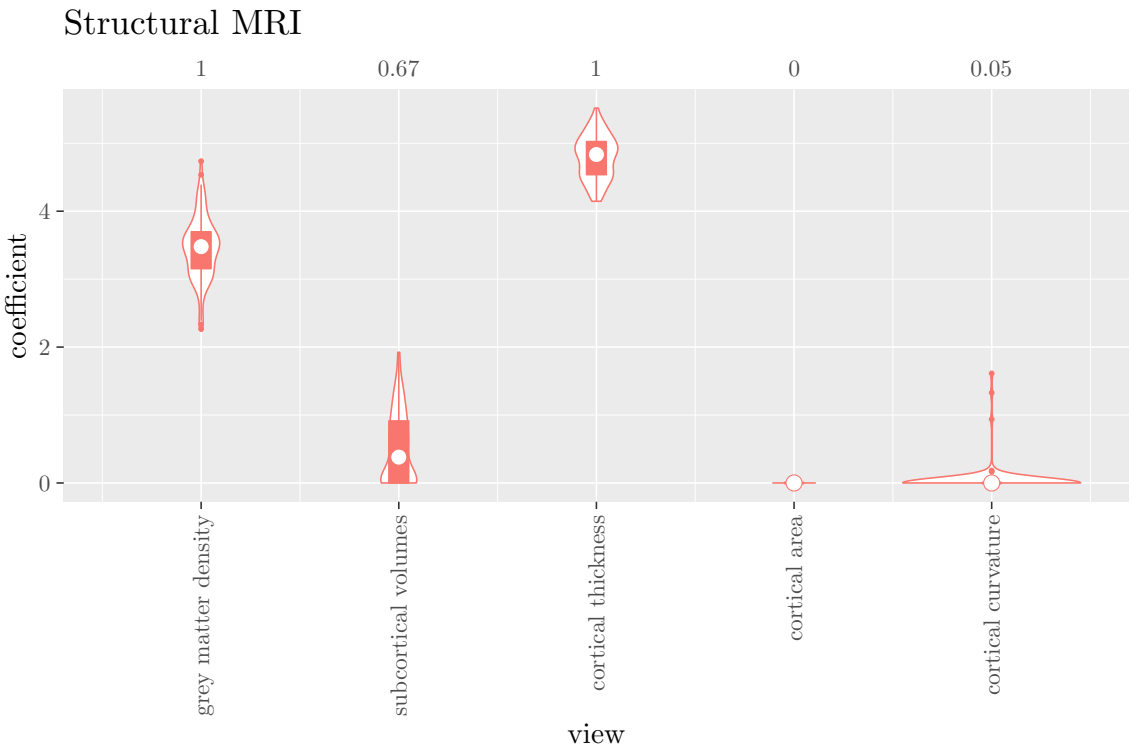

Figure 4: Box-and-violin plots of the StaPLR intermediate-level regression coefficients for the structural scan type. The numbers at the top of the graph denote the proportion of times the coefficient was nonzero.

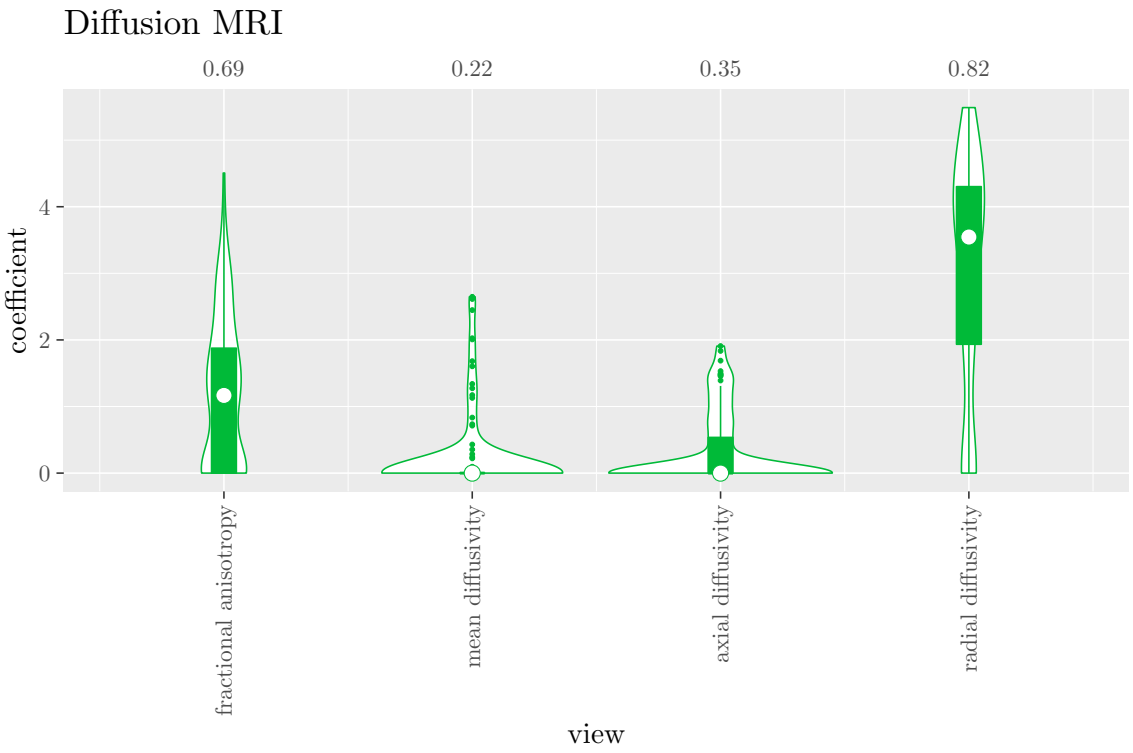

Figure 5: Box-and-violin plots of the StaPLR intermediate-level regression coefficients for the diffusion scan type. The numbers at the top of the graph denote the proportion of times the coefficient was nonzero.

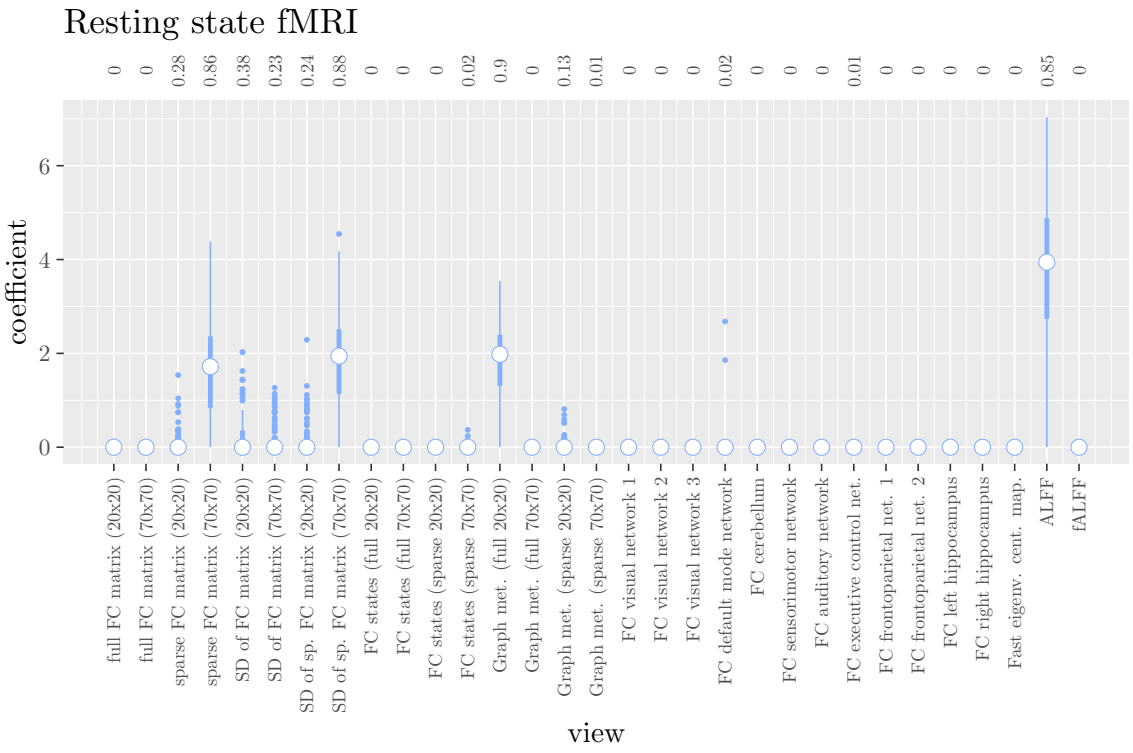

Figure 6: Box plots of the StaPLR intermediate-level regression coefficients for the functional scan type. The numbers at the top of the graph denote the proportion of times the coefficient was nonzero.

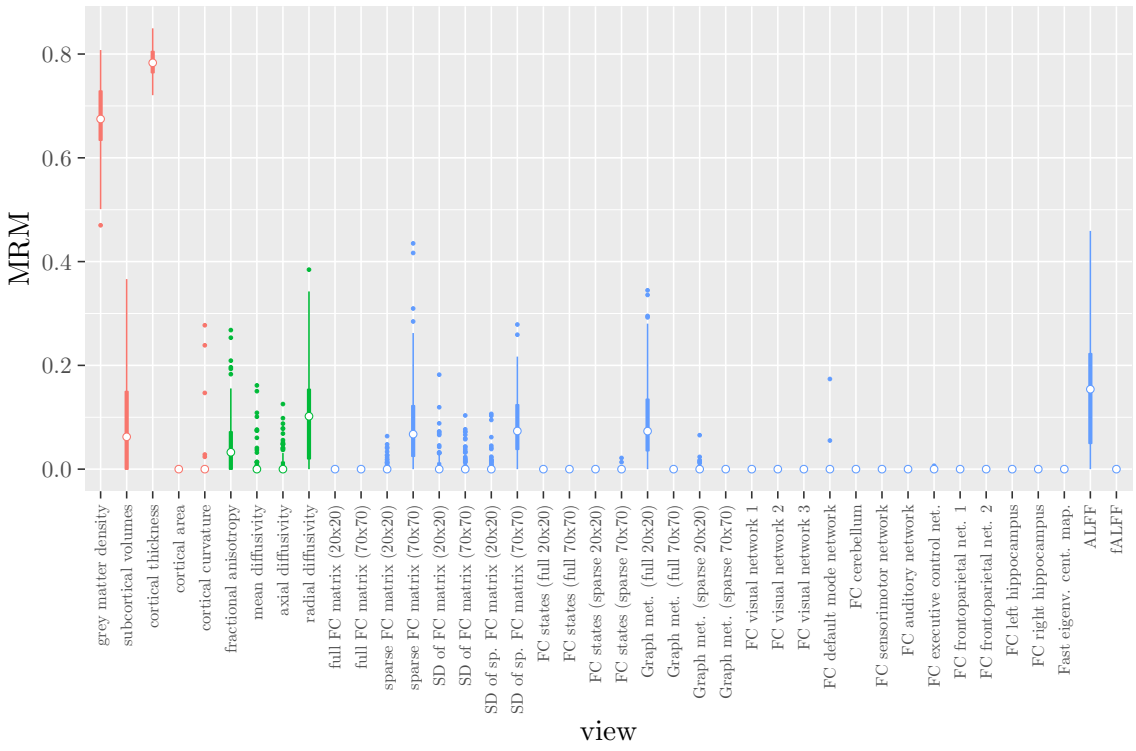

Figure 7: Influence of each of the MRI measures on the predictions of the stacked classifier, as quantified by the minority report measure. MRM values range from 0 to 1, with 0 indicating no influence and 1 indicating maximum possible influence. The colors of the bars refer to the different scan types (red = structural MRI, green = diffusion MRI, blue = resting state fMRI).

4 Discussion

We have extended the StaPLR algorithm to adapt to a hierarchical multi-view data structure, and have presented an application of this extension to a multi-view MRI data set in the context of Alzheimer's disease classification. The presented application can serve as an example of a more general class of applications within neuroimaging and biomedical science. In our specific application to AD classification, the classifier produced by StaPLR was more accurate than the one produced by elastic net regression. We have shown how in StaPLR the relative importance of MRI measures derived from the same scan type can easily be compared using their regression coefficient. Additionally, we have introduced the minority report measure, which allows for comparing the importance of MRI measures derived from different scan types.

In addition to comparing the relative importance of the different MRI measures using the regression coefficients or the MRM, we may want to make a binary decision: is this measure required for prediction of the outcome or not? This is of course more difficult, since although some measures were selected 100% or 0% of the time, for many measures the situation is not so clear-cut. One approach would be to say that for an MRI measure to be important, it should have been selected at least 50% of the time (i.e. its median coefficient should be nonzero). In this case we would select three structural measures (cortical thickness, grey matter density and subcortical volumes), two diffusion measures (fractional anisotropy and radial diffusivity), and four functional measures (ALFF, the graph metrics as computed from the full 20x20 FC matrix, the sparse 70x70 FC matrix, and the SDs associated with the latter), for a total of nine selected MRI measures. Note that this is considerably sparser than the elastic net, for which the selected features were on average spread out over 24 MRI measures. It is also interesting to see that the observed selection probabilities for the different MRI measures are not generally in the neighborhood of 50%. Instead, all measures were included either at least 67% of the time, or less than 38% of the time, providing a clear separation into a "frequently selected" and "infrequently selected" group of MRI measures.

Both StaPLR and the elastic net appeared to agree on the structural MRI measures of grey matter density and cortical thickness being the most important for classification, which is in line with earlier research identifying measures of grey matter atrophy as important bio-markers for Alzheimer's disease [47, 48]. The largest difference between the two methods was seen in terms of fMRI measures, of which StaPLR selected 4, and the elastic net 17. In particular, the elastic net appeared to include more features from the larger feature sets (Figure 1), such as the feature sets containing the voxel-wise functional connectivity with individual RSNs. In contrast, StaPLR includes MRI measures which contain summarizing information about RSNs (i.e. graph metrics, the sparse 70x70 FC matrix, and the dynamics of the sparse 70x70 FC matrix). The importance of the 70x70 FC matrix and its dynamics are in line with the results of a previous study which used only the resting state fMRI scans for AD classification [27]. The results of our StaPLR analysis suggest that although the structural scan type is dominant in the classification of Alzheimer's disease, diffusion MRI and resting-state fMRI can both provide useful,

if smaller, contributions to the classification. These results are broadly in line with a previous study investigating the relevance of a smaller subset of structural, diffusion and resting-state functional MRI measures [9].

Although StaPLR selected all MRI scan types, it did make a smaller selection of required MRI measures. In practice, such a selection may translate to less time spend on the computation of different feature sets from MRI scans. A drawback of StaPLR, which it shares with all penalized regression methods, is that for any single run of the algorithm the selection is binary: a view is either selected or not. As discussed above, the actual set of selected views may vary from run to run. In this article, we have quantified this variability by showing the distribution of results over all repeated cross-validation folds. Other re-sampling methods, such as the bootstrap, could also be used to gain more insight in the stability of the results. However, compared with subsampling, bootstrapping may increase the likelihood of noise variables being selected [49]. In addition, re-sampling methods are typically computationally expensive. In the future, we therefore aim to introduce a form of uncertainty quantification, such as confidence intervals, that can be computed from only a single run of the StaPLR algorithm.

Naturally, the results of the StaPLR algorithm depend on the specified multi-view structure. In our analysis, features were nested in *MRI measures*, which were in turn nested in *scan types*. The multi-view structure was specified this way because its matches the data collection process, and to theoretically allow for the largest reduction in costs by selecting views. The largest reduction in time (for both researcher and patient) and monetary cost can be obtained by omitting one or more of the scan types. Following that, the largest reduction in time and costs can be obtained by not computing certain MRI measures. However, one could specify a different multi-view structure to match a different research question. For example, if the primary interest is in identifying which *brain areas* are the most important for AD classification, one could treat each brain area as a separate view. This may, of course, lead to different results. For example, in our analysis the structural feature set consisting of volumetric measurements of seven different subcortical structures was found to play a minor role in AD classification compared with grey matter density or cortical thickness. Decoupling the volumes of the different subcortical structures and treating each brain area as a separate view allows each structure to obtain its own weight. In such an analysis, one might see an increased importance of certain structures traditionally associated with AD, such as the hippocampus. However, such an analysis is outside the scope of this article. The application shown in this article serves as an example of how StaPLR can be applied to hierarchical multi-view data. It should be noted that the method can be further extended to a more complex structure, such as a hierarchical structure with more levels, or a structure with a mixed number of levels. The latter may be of particular importance when the data is collected from entirely different domains. For example, the hierarchical multi-view structure for MRI data may be quite different from that of genetic data, other biomarkers, or clinical variables. Such a difference can easily be handled by the StaPLR algorithm, paving the way for applications to larger multi-source data sets such as those obtained through the UK Biobank initiative.

5 Conclusion

We have extended the StaPLR algorithm to hierarchical multi-view MRI data, and applied it to Alzheimer's disease classification. We have shown that StaPLR produces a stacked classifier that allows researchers to see which scan types, and which MRI measures derived from those scan types, play the most important role in classification. In addition, the stacked classifier showed an increase in classification accuracy when compared with logistic elastic net regression.

6 CRediT author contribution statement

Wouter van Loon: Conceptualization, Methodology, Software, Formal analysis, Investigation, Writing - original draft, Writing - review & editing, Visualization.
Frank de Vos: Investigation, Data curation, Writing - review & editing.
Marjolein Fokkema: Writing - review & editing.
Botond Szabo: Writing - review & editing.
Marisa Koini: Data curation.
Reinhold Schmidt: Data curation.
Mark de Rooij: Conceptualization, Writing - Review & Editing, Supervision.

7 Declaration of interest

Declarations of interest: none.

A Supplementary materials

These supplementary materials describe the process of obtaining the features used in "Analyzing hierarchical multi-view MRI data with StaPLR: An application to Alzheimer's disease classification". This appendix is largely a reiteration of previous work, in particular that of de Vos et al. [16, 27] and Schouten et al. [9, 28]. The relevant information from these publications is collected here for the reader's convenience.

B Participants

The data were collected at the Medical University of Graz in Austria, and consisted of 76 clinically diagnosed probable AD patients and 173 cognitively normal elderly. The AD patients were part of the prospective registry on dementia (PRODEM) [7]. The inclusion criteria for PRODEM are: Dementia diagnosis according to DSM-IV criteria [50], AD diagnosis according to the NINCDS-ADRDA criteria [51], non-institutionalisation or need for 24-h care, and the availability of a caregiver who agrees to provide information on the patients' and his or her own condition. Patients were excluded if co-morbidities were likely to preclude successful completion of the study. Informed consent was obtained

from all patients and their caregivers. We only included patients for which anatomical MRI, diffusion MRI and rs-fMRI were available. The controls were scanned at the same scanning site, over the same period, with the same scanning protocol as the AD patients as a part of the Austrian Stroke Prevention Study (ASPS). The ASPS is a community-based cohort study on the effects of vascular risk factors on brain structure and function in elderly participants without a history or signs of stroke and dementia on the inhabitants of Graz, Austria [25, 26]. Informed consent was obtained from all participants.

C MRI acquisition

Each participant was scanned on a Siemens Magnetom TrioTim 3 T MRI scanner. Anatomical T1-weighted images were acquired with TR = 1900 ms, TE = 2.19 ms, flip angle = 9, 179 slices, with an isotropic voxel size of 1 mm.

Diffusion images were acquired along 12 non-collinear directions, scanning each direction 4 times with TR = 6700 ms, TE = 95 ms, 50 axial slices, voxel size = $2.0 \times 2.0 \times 2.5$ mm.

Resting-state fMRI series of 150 volumes were obtained with TR = 3000 ms, TE = 30 ms, flip angle = 90, 40 axial slices, with an isotropic voxel size of 3 mm. Participants were instructed to lie still with their eyes closed, and to stay awake.

D MRI preprocessing

The MRI data of all subjects were preprocessed using the FMRIB Software Library (FSL version 5.0) [52, 53]. For the anatomical MRI scans, we applied brain extraction and bias field correction. For the diffusion MRI scans, we applied brain extraction and eddy current correction. For the rs-fMRI data, this included brain extraction, motion correction, a temporal high pass filter with a cutoff point of 100 seconds, 3 mm FWHM spatial smoothing, and non-linear registration to standard MNI152 space. Additionally, we used ICA-AROMA to automatically identify and remove noise components from the fMRI time course [54]. ICA-AROMA adequately removes motion related noise from fMRI data, without the need for removing volumes with excessive motion [55].

E Feature extraction

E.1 Structural MRI

The process of extracting the features corresponding to cortical thickness, area, curvature, grey matter density, and subcortical volumes, is identical to the process described in de Vos et al. [16]. For completeness, we also describe it below.

In order to calculate cortical thickness, cortical area and cortical curvature, the raw (not preprocessed) T1-weighted images were processed using Freesurfer 5.3.0 [56, 57]. First, this entails intensity normalization and brain extraction [16]. Using the resulting

image the boundary between grey and white matter was located, and a triangular mesh was constructed around the white matter surface [16]. The grey matter surface was created by deforming the mesh outward so that it closely followed the boundary between grey matter and cerebral spinal fluid [16]. Cortical thickness was calculated as the distance between the white matter and grey matter surface for each vertex [16]. The image was then registered to the Freesurfer common template using the image's cortical folding pattern [16]. The neocortex was parcellated into the 68 regions of the Desikan-Killiany atlas [36]. The thickness of each parcellation unit was calculated as the mean thickness of all the vertices within that parcellation [16]. Thus, 68 cortical thickness features are obtained per subject. The cortical surface area was calculated by summing the areas of the grey matter mesh triangles for each parcellation, yielding 68 cortical area features per subject [16]. To obtain the cortical curvature features, the mean of the curvature values in the two principal directions of the of the surface was calculated [16]. The curvature of a vertex in these directions was calculated as the inverse of the length of the radius of osculating circles in these directions [58]. The curvature values of the vertices were averaged for each of the parcellations, yielding 68 cortical curvature features per subject [16].

Grey matter density was calculated using FSL VBM (version 5.0.7) [53, 59]. The brain-extracted images were first segmented into grey matter, white matter, and CSF [16]. A study-specific grey matter template was created in two steps. First, the grey matter images were affine-registered to the ICBM-152 grey matter template and the resulting images were averaged to create a first-pass template [16]. Then, the grey matter images were nonlinearly registered to the first-pass temlate and the resulting images were averaged to obtain a final template at $2 \times 2 \times 2\text{mm}^3$ resolution in standard space [16]. The grey matter images were then registered to the final template and smoothed with a Gaussian kernel with a full width at half maximum of 3 mm [16]. The voxel wise values were then averaged within the 48 regions of the probabilistic Harvard-Oxford cortical atlas [16]. The 48 grey matter density features were obtained by calculating the weighted averages of the regions, with voxels contribution to the average of a region based on their probability of being part of that region [16].

The volumes of the subcortical structures were calculated using the FMRIB's Integrated Registration and Segmentation Tool (FIRST) in FSL [60]. The whole-head images were affine registered to the nonlinear MNI-152 template [16]. In a second stage, initialized by the result of the first stage, a subcortical mask was used to achieve a more accurate and robust affine registration [16]. The shapes of the subcortical structures were modeled by deformable meshes and the boundary voxels were classified as being part of the subcortical structure using structural segmentation [61]. The cortical volumes were then corrected for intracranial volume as obtained by FSL, yielding 14 subcortical volume features per subject, corresponding to the thalamus, caudate, putamen, pallidum, hippocampus, amygdala and accumbens of both hemispheres [16].

E.2 Diffusion-weighted MRI

The diffusion MRI scans were used to calculate fractional anisotropy (FA), mean diffusivity (MD), axial diffusivity (DA), and radial diffusivity (DR). First, DTIFIT in FSL [52, 53] was used to fit a diffusion tensor model at each voxel to calculate voxel-wise FA, MD, DA and DR images for each subject. Then subjects' FA, MD, DA and DR images were projected onto the FMRIB58_FA mean FA image using tract based spatial statistics (TBSS) [62]. Finally, weighted averages of the FA, MD, DA and DR values were calculated within the 20 regions of the probabilistic JHU white-matter tractography atlas [37], yielding 20 features for FA as well as MD, DA and DR [9].

E.3 Resting state fMRI

The resting state fMRI feature sets used in this article have already been described in detail in de Vos et al. [27]; the following is a reiteration of the most relevant sections.

Resting state networks (RSNs) were obtained using temporal concatenation independent component analysis (ICA) in FSL MELODIC [63]. The functional data of all participants was registered to standard space and concatenated along the time dimension. ICA was performed on the concatenated dataset, once with 20 and once with 70 components [27]. The resulting ICA component weight maps were registered back to subject space, weighted by subject specific grey matter density maps, and multiplied with the functional data, resulting in mean time course for each component [27]. These time course were then used to calculate functional connectivity matrices using both full and sparse partial correlations [27]. The partial correlation matrices were calculated using the graphical lasso [64] implemented in MATLAB [65], with $\lambda = 100$ [27]. The resulting two 20×20 matrices contain 190 unique elements, and the 70×70 matrices contain 2415 unique elements to be used as candidate features in the classification. Any features with zero variance were removed.

The dynamics of the FC matrices were calculated using a sliding window approach with a window size of 33s [27]. The windows were shifted one volume at a time, leading to 140 windows [27]. The previously described four FC matrices were calculated within each window, and the standard deviation of the FC matrices of all windows was obtained [27].

The sliding window FC matrices were clustered using k-means clustering (with k=5 and Manhattan distance) to obtain 5 "FC states" [27]. The number of sliding window matrices that were assigned to each of the five FC states was then calculated for each participant [27].

Graph metrics were calculated using the Brain Connectivity Toolbox [66] in MATLAB [65] for each of the four FC matrices. Connection strength, weighted betweenness centrality and weighted clustering coefficients were calculated for every node, and weighted characteristic path length and weighted transitivity for the entire network [27]. Additionally, several graph metrics were calculated on binarized versions of the FC matrices: connection degree, betweenness centrality and clustering coefficient for every node, and characteristic path length and transitivity for the entire network [27]

Whole brain FC with 10 RSNs was calculated using dual regression in FSL [67], using the RSN templates of Smith et al. [68]. The voxel-wise whole brain FC results for each of the 10 RSNs were used as 10 distinct feature sets. Voxel-wise whole brain FC maps were additionally obtained for the left and right hippocampus [27].

Eigenvector centrality maps were calculated for each participant using fastECM [69, 70]. The amplitude of low frequency fluctuations (ALFF) [71, 72] and fractional ALFF [73] were calculated for each participant using REST [74]. The voxels' ALFF and fALFF values were divided by the mean ALFF/fALFF within a subjects whole brain [27, 73].

References

[1] Y. Li, F.-X. Wu, and A. Ngom, "A review on machine learning principles for multi-view biological data integration," *Briefings in Bioinformatics*, vol. 19, no. 2, pp. 325–340, 2018.

[2] M. Fratello, G. Caiazzo, F. Trojsi, A. Russo, G. Tedeschi, R. Tagliaferri, and F. Esposito, "Multi-view ensemble classification of brain connectivity images for neurodegeneration type discrimination," *Neuroinformatics*, vol. 15, no. 2, pp. 199–213, 2017.

[3] C. Sudlow, J. Gallacher, N. Allen, V. Beral, P. Burton, J. Danesh, P. Downey, P. Elliott, J. Green, M. Landray *et al.*, "UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age," *PLoS Medicine*, vol. 12, no. 3, p. e1001779, 2015.

[4] T. J. Littlejohns, J. Holliday, L. M. Gibson, S. Garratt, N. Oesingmann, F. Alfaro-Almagro, J. D. Bell, C. Boultwood, R. Collins, M. C. Conroy *et al.*, "The UK biobank imaging enhancement of 100,000 participants: rationale, data collection, management and future directions," *Nature Communications*, vol. 11, no. 1, pp. 1–12, 2020.

[5] S. G. Mueller, M. W. Weiner, L. J. Thal, R. C. Petersen, C. Jack, W. Jagust, J. Q. Trojanowski, A. W. Toga, and L. Beckett, "The Alzheimer's disease neuroimaging initiative," *Neuroimaging Clinics of North America*, vol. 15, no. 4, p. 869, 2005.

[6] K. Krysinska, P. S. Sachdev, J. Breitner, M. Kivipelto, W. Kukull, and H. Brodaty, "Dementia registries around the globe and their applications: A systematic review," *Alzheimer's & Dementia*, vol. 13, no. 9, pp. 1031–1047, 2017.

[7] S. Seiler, H. Schmidt, A. Lechner, T. Benke, G. Sanin, G. Ransmayr, R. Lehner, P. Dal-Bianco, P. Santer, P. Linortner *et al.*, "Driving cessation and dementia: results of the prospective registry on dementia in Austria (PRODEM)," *PLoS ONE*, vol. 7, no. 12, p. e52710, 2012.

[8] F. Liem, G. Varoquaux, J. Kynast, F. Beyer, S. K. Masouleh, J. M. Huntenburg, L. Lampe, M. Rahim, A. Abraham, R. C. Craddock, S. Riedel-Heller, T. Luck,

M. Loeffler, M. L. Schroeter, A. V. Witte, A. Villringer, and D. S. Margulies, "Predicting brain-age from multimodal imaging data captures cognitive impairment," *NeuroImage*, vol. 148, pp. 179–188, 2017.

[9] T. Schouten, M. Koini, F. De Vos, S. Seiler, J. van der Grond, A. Lechner, A. Hafkemeijer, C. Möller, R. Schmidt, M. de Rooij, and S. Rombouts, "Combining anatomical, diffusion, and resting state functional magnetic resonance imaging for individual classification of mild and moderate Alzheimer's disease," *NeuroImage: Clinical*, vol. 11, pp. 46–51, 2016.

[10] M. Rahim, B. Thirion, C. Comtat, and G. Varoquaux, "Transmodal learning of functional networks for Alzheimer's disease prediction," *IEEE Journal of Selected Topics in Signal Processing*, vol. 10, no. 7, pp. 1204–1213, 2016.

[11] R. Li, A. Hapfelmeier, J. Schmidt, R. Perneczky, A. Drzezga, A. Kurz, and S. Kramer, "A case study of stacked multi-view learning in dementia research," in *13th Conference on Artificial Intelligence in Medicine*, 2011, pp. 60–69.

[12] J. Zhao, X. Xie, X. Xu, and S. Sun, "Multi-view learning overview: recent progress and new challenges," *Information Fusion*, vol. 38, pp. 43–54, 2017.

[13] S. Sun, L. Mao, Z. Dong, and L. Wu, *Multiview Machine Learning*. Springer, 2019.

[14] E. Garcia-Ceja, C. E. Galván-Tejada, and R. Brena, "Multi-view stacking for activity recognition with sound and accelerometer data," *Information Fusion*, vol. 40, pp. 45–56, 2018.

[15] W. van Loon, M. Fokkema, B. Szabo, and M. de Rooij, "Stacked penalized logistic regression for selecting views in multi-view learning," *Information Fusion*, vol. 61, pp. 113–123, 2020.

[16] F. de Vos, T. Schouten, A. Hafkemeijer, E. Dopper, J. van Swieten, M. de Rooij, J. van der Grond, and S. Rombouts, "Combining multiple anatomical MRI measures improves Alzheimer's disease classification," *Human Brain Mapping*, vol. 37, pp. 1920–1929, 2016.

[17] R. Salvador, E. Canales-Rodríguez, A. Guerrero-Pedraza, S. Sarró, D. Tordesillas-Gutiérrez, T. Maristany, B. Crespo-Facorro, P. McKenna, and E. Pomarol-Clotet, "Multimodal integration of brain images for MRI-based diagnosis in schizophrenia," *Frontiers in Neuroscience*, vol. 13, no. 1203, pp. 1–9, 2019.

[18] M. Guggenmos, K. Schmack, I. M. Veer, T. Lett, M. Sekutowicz, M. Sebold, M. Garbusow, C. Sommer, H.-U. Wittchen, U. S. Zimmermann, M. N. Smolka, H. Walter, A. Heinz, and P. Sterzer, "A multimodal neuroimaging classifier for alcohol dependence," *Scientific Reports*, vol. 10, no. 298, pp. 1–12, 2020.

[19] D. A. Engemann, O. Kozynets, D. Sabbagh, G. Lemaître, G. Varoquaux, F. Liem, and A. Gramfort, "Combining magnetoencephalography with magnetic resonance imaging enhances learning of surrogate-biomarkers," *eLife*, vol. 9, no. e54055, pp. 1–32, 2020.

[20] L. Ali, Z. He, W. Cao, H. T. Rauf, Y. Imrana, and M. B. B. Heyat, "MMDD-ensemble: A multimodal data driven ensemble approach for Parkinson's disease detection," *Frontiers in Neuroscience*, vol. 15, no. 754058, pp. 1–11.

[21] A. E. Hoerl and R. W. Kennard, "Ridge regression: biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, no. 1, pp. 55–67, 1970.

[22] S. Le Cessie and J. C. Van Houwelingen, "Ridge estimators in logistic regression," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 41, no. 1, pp. 191–201, 1992.

[23] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B*, vol. 58, no. 1, pp. 267–288, 1996.

[24] J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent," *Journal of Statistical Software*, vol. 33, no. 1, pp. 1–22, 2010. [Online]. Available: http://www.jstatsoft.org/v33/i01/

[25] R. Schmidt, H. Lechner, F. Fazekas, K. Niederkorn, B. Reinhart, P. Grieshofer, S. Horner, H. Offenbacher, M. Koch, B. Eber *et al.*, "Assessment of cerebrovascular risk profiles in healthy persons: definition of research goals and the Austrian stroke prevention study (ASPS)," *Neuroepidemiology*, vol. 13, no. 6, pp. 308–313, 1994.

[26] P. Freudenberger, K. Petrovic, A. Sen, A. M. Töglhofer, A. Fixa, E. Hofer, S. Perl, R. Zweiker, S. Seshadri, R. Schmidt *et al.*, "Fitness and cognition in the elderly: the Austrian stroke prevention study," *Neurology*, vol. 86, no. 5, pp. 418–424, 2016.

[27] F. de Vos, M. Koini, T. Schouten, S. Seiler, J. van der Grond, A. Lechner, R. Schmidt, M. de Rooij, and S. Rombouts, "A comprehensive analysis of resting state fMRI measures to classify individual patients with Alzheimer's disease," *NeuroImage*, vol. 167, pp. 62–72, 2017.

[28] T. M. Schouten, M. Koini, F. de Vos, S. Seiler, M. de Rooij, A. Lechner, R. Schmidt, M. van den Heuvel, J. van der Grond, and S. A. Rombouts, "Individual classification of Alzheimer's disease with diffusion magnetic resonance imaging," *NeuroImage*, vol. 152, pp. 476–481, 2017.

[29] W. van Loon, M. Fokkema, B. Szabo, and M. de Rooij, "View selection in multi-view stacking: choosing the meta-learner," *arXiv preprint arXiv:2010.16271*, 2020.

[30] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B*, vol. 67, no. 2, pp. 301–320, 2005.

[31] P. T. Trzepacz, P. Yu, J. Sun, K. Schuh, M. Case, M. M. Witte, H. Hochstetler, and A. Hake, "Comparison of neuroimaging modalities for the prediction of conversion from mild cognitive impairment to Alzheimer's dementia," *Neurobiology of Aging*, vol. 35, no. 1, pp. 143–151, 2014.

[32] S. J. Teipel, J. Kurth, B. Krause, M. J. Grothe, ADNI *et al.*, "The relative importance of imaging markers for the prediction of Alzheimer's disease dementia in mild cognitive impairment — beyond classical regression," *NeuroImage: Clinical*, vol. 8, pp. 583–593, 2015.

[33] F. D. Bowman, D. F. Drake, and D. E. Huddleston, "Multimodal imaging signatures of Parkinson's disease," *Frontiers in Neuroscience*, vol. 10, no. 131, pp. 1–11, 2016.

[34] T. M. Nir, J. E. Villalon-Reina, B. A. Gutman, D. Moyer, N. Jahanshad, M. Dehghani, C. R. Jack, M. W. Weiner, P. M. Thompson, ADNI *et al.*, "Alzheimer's disease classification with novel microstructural metrics from diffusion-weighted MRI," in *Computational Diffusion MRI*, 2016, pp. 41–54.

[35] *Harvard-Oxford Cortical Atlas*. Harvard Center for Morphometric Analysis. [Online]. Available: https://cma.mgh.harvard.edu/

[36] R. S. Desikan, F. Ségonne, B. Fischl, B. T. Quinn, B. C. Dickerson, D. Blacker, R. L. Buckner, A. M. Dale, R. P. Maguire, B. T. Hyman *et al.*, "An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest," *NeuroImage*, vol. 31, no. 3, pp. 968–980, 2006.

[37] K. Hua, J. Zhang, S. Wakana, H. Jiang, X. Li, D. S. Reich, P. A. Calabresi, J. J. Pekar, P. C. van Zijl, and S. Mori, "Tract probability maps in stereotaxic spaces: analyses of white matter anatomy and tract-specific quantification," *NeuroImage*, vol. 39, no. 1, pp. 336–347, 2008.

[38] S. Varma and R. Simon, "Bias in error estimation when using cross-validation for model selection," *BMC Bioinformatics*, vol. 7, no. 91, 2006.

[39] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2017. [Online]. Available: https://www.R-project.org/

[40] W. van Loon, *R package 'multiview' - Methods for high-dimensional multi-view learning (v0.3.1)*, Feb. 2021. [Online]. Available: https://doi.org/10.5281/zenodo.4630669

[41] W. van Loon, *Code repository accompanying "Analyzing hierarchical multi-view MRI data with StaPLR: An application to Alzheimer's disease classification"*, Jul. 2021. [Online]. Available: https://doi.org/10.5281/zenodo.5105730

[42] *Oxford Languages*, retrieved May 2021. [Online]. Available: https://languages.oup.com/

[43] P. K. Dick, *The Minority Report.* Kensington Publishing, 2002, reprint of 1952 original.

[44] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[45] A. Fisher, C. Rudin, and F. Dominici, "All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously." *Journal of Machine Learning Research*, vol. 20, no. 177, pp. 1–81, 2019.

[46] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proceedings of the 31st international conference on neural information processing systems*, 2017, pp. 4768–4777.

[47] J. P. Lerch, J. C. Pruessner, A. Zijdenbos, H. Hampel, S. J. Teipel, and A. C. Evans, "Focal decline of cortical thickness in Alzheimer's disease identified by computational neuroanatomy," *Cerebral cortex*, vol. 15, no. 7, pp. 995–1001, 2005.

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

[49] R. De Bin, S. Janitza, W. Sauerbrei, and A.-L. Boulesteix, "Subsampling versus bootstrapping in resampling-based model selection for multivariable regression," *Biometrics*, vol. 72, no. 1, pp. 272–280, 2016.

[50] *Diagnostic and statistical manual of mental disorders*, 4th ed. American Psychiatric Association, Washington, DC, 2000.

[51] G. M. McKhann, D. S. Knopman, H. Chertkow, B. T. Hyman, C. R. Jack Jr, C. H. Kawas, W. E. Klunk, W. J. Koroshetz, J. J. Manly, R. Mayeux *et al.*, "The diagnosis of dementia due to alzheimer's disease: recommendations from the national institute on aging-alzheimer's association workgroups on diagnostic guidelines for alzheimer's disease," *Alzheimer's & dementia*, vol. 7, no. 3, pp. 263–269, 2011.

[52] M. Jenkinson, C. F. Beckmann, T. E. Behrens, M. W. Woolrich, and S. M. Smith, "Fsl," *NeuroImage*, vol. 62, no. 2, pp. 782–790, 2012.

[53] S. M. Smith, M. Jenkinson, M. W. Woolrich, C. F. Beckmann, T. E. Behrens, H. Johansen-Berg, P. R. Bannister, M. De Luca, I. Drobnjak, D. E. Flitney *et al.*, "Advances in functional and structural mr image analysis and implementation as fsl," *NeuroImage*, vol. 23, pp. S208–S219, 2004.

[54] R. H. Pruim, M. Mennes, D. van Rooij, A. Llera, J. K. Buitelaar, and C. F. Beckmann, "Ica-aroma: A robust ica-based strategy for removing motion artifacts from fmri data," *NeuroImage*, vol. 112, pp. 267–277, 2015.

[55] L. Parkes, B. Fulcher, M. Yücel, and A. Fornito, "An evaluation of the efficacy, reliability, and sensitivity of motion correction strategies for resting-state functional mri," *NeuroImage*, vol. 171, pp. 415–436, 2018.

[56] A. M. Dale, B. Fischl, and M. I. Sereno, "Cortical surface-based analysis: I. segmentation and surface reconstruction," *NeuroImage*, vol. 9, no. 2, pp. 179–194, 1999.

[57] B. Fischl, M. I. Sereno, and A. M. Dale, "Cortical surface-based analysis: Ii: inflation, flattening, and a surface-based coordinate system," *NeuroImage*, vol. 9, no. 2, pp. 195–207, 1999.

[58] L. Ronan, R. Pienaar, G. Williams, E. Bullmore, T. J. Crow, N. Roberts, P. B. Jones, J. Suckling, and P. C. Fletcher, "Intrinsic curvature: a marker of millimeter-scale tangential cortico-cortical connectivity?" *International journal of neural systems*, vol. 21, no. 05, pp. 351–366, 2011.

[59] J. Ashburner and K. J. Friston, "Voxel-based morphometry—the methods," *NeuroImage*, vol. 11, no. 6, pp. 805–821, 2000.

[60] B. Patenaude, S. M. Smith, D. N. Kennedy, and M. Jenkinson, "A bayesian model of shape and appearance for subcortical brain segmentation," *NeuroImage*, vol. 56, no. 3, pp. 907–922, 2011.

[61] Y. Zhang, M. Brady, and S. Smith, "Segmentation of brain mr images through a hidden markov random field model and the expectation-maximization algorithm," *IEEE transactions on medical imaging*, vol. 20, no. 1, pp. 45–57, 2001.

[62] S. M. Smith, M. Jenkinson, H. Johansen-Berg, D. Rueckert, T. E. Nichols, C. E. Mackay, K. E. Watkins, O. Ciccarelli, M. Z. Cader, P. M. Matthews *et al.*, "Tract-based spatial statistics: voxelwise analysis of multi-subject diffusion data," *NeuroImage*, vol. 31, no. 4, pp. 1487–1505, 2006.

[63] C. F. Beckmann and S. M. Smith, "Probabilistic independent component analysis for functional magnetic resonance imaging," *IEEE transactions on medical imaging*, vol. 23, no. 2, pp. 137–152, 2004.

[64] J. Friedman, T. Hastie, and R. Tibshirani, "Sparse inverse covariance estimation with the graphical lasso," *Biostatistics*, vol. 9, no. 3, pp. 432–441, 2008.

[65] T. M. Inc., *MATLAB and Statistics Toolbox Release, 2013a*, 2013.

[66] M. Rubinov and O. Sporns, "Complex network measures of brain connectivity: uses and interpretations," *NeuroImage*, vol. 52, no. 3, pp. 1059–1069, 2010.

[67] N. Filippini, B. MacIntosh, M. Hough, G. Goodwin, G. Frisoni, K. Ebmeier, S. Smith, P. Matthews, C. Beckmann, and C. Mackay, "Distinct patterns of brain activity in young carriers of the apoe e4 allele," *NeuroImage*, vol. 47, pp. S139–S139, 2009.

[68] S. M. Smith, K. L. Miller, S. Moeller, J. Xu, E. J. Auerbach, M. W. Woolrich, C. F. Beckmann, M. Jenkinson, J. Andersson, M. F. Glasser *et al.*, "Temporally-independent functional modes of spontaneous brain activity," *Proceedings of the National Academy of Sciences*, vol. 109, no. 8, pp. 3131–3136, 2012.

[69] A. M. Wink, J. C. de Munck, Y. D. van der Werf, O. A. van den Heuvel, and F. Barkhof, "Fast eigenvector centrality mapping of voxel-wise connectivity in functional magnetic resonance imaging: implementation, validation, and interpretation," *Brain connectivity*, vol. 2, no. 5, pp. 265–274, 2012.

[70] M. A. Binnewijzend, S. M. Adriaanse, W. M. Van der Flier, C. E. Teunissen, J. C. de Munck, C. J. Stam, P. Scheltens, B. N. van Berckel, F. Barkhof, and A. M. Wink, "Brain network alterations in alzheimer's disease measured by eigenvector centrality in fmri are related to cognition and csf biomarkers," *Human brain mapping*, vol. 35, no. 5, pp. 2383–2393, 2014.

[71] Z. Yu-Feng, H. Yong, Z. Chao-Zhe, C. Qing-Jiu, S. Man-Qiu, L. Meng, T. Li-Xia, J. Tian-Zi, and W. Yu-Feng, "Altered baseline brain activity in children with adhd revealed by resting-state functional mri," *Brain and Development*, vol. 29, no. 2, pp. 83–91, 2007. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0387760406001549

[72] B. B. Biswal, M. Mennes, X.-N. Zuo, S. Gohel, C. Kelly, S. M. Smith, C. F. Beckmann, J. S. Adelstein, R. L. Buckner, S. Colcombe *et al.*, "Toward discovery science of human brain function," *Proceedings of the National Academy of Sciences*, vol. 107, no. 10, pp. 4734–4739, 2010.

[73] Q.-H. Zou, C.-Z. Zhu, Y. Yang, X.-N. Zuo, X.-Y. Long, Q.-J. Cao, Y.-F. Wang, and Y.-F. Zang, "An improved approach to detection of amplitude of low-frequency fluctuation (alff) for resting-state fmri: fractional alff," *Journal of neuroscience methods*, vol. 172, no. 1, pp. 137–141, 2008.

[74] X.-W. Song, Z.-Y. Dong, X.-Y. Long, S.-F. Li, X.-N. Zuo, C.-Z. Zhu, Y. He, C.-G. Yan, and Y.-F. Zang, "Rest: a toolkit for resting-state functional magnetic resonance imaging data processing," *PloS one*, vol. 6, no. 9, p. e25031, 2011.