frontiers | Frontiers in Neuroscience

ORIGINAL RESEARCH
published: 25 April 2022
doi: 10.3389/fnins.2022.830630

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*[4,5], *Marisa Koini*[6], *Reinhold Schmidt*[6] and *Mark de Rooij*[1,3]

[1] Department of Methodology and Statistics, Leiden University, Leiden, Netherlands, [2] Department of Radiology, Leiden University Medical Center, Leiden, Netherlands, [3] Leiden Institute for Brain and Cognition, Leiden, Netherlands, [4] Department of Decision Sciences, Bocconi University, Milan, Italy, [5] Bocconi Institute for Data Science and Analytics, Bocconi University, Milan, Italy, [6] Division of Neurogeriatrics, Department of Neurology, Medical University of Graz, Graz, Austria

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 derived 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

OPEN ACCESS

Edited by:
Nicha C. Dvornek,
Yale University, United States

Reviewed by:
Jingyu Liu,
Georgia State University, United States
Yi Zhao,
Indiana University, United States

****Correspondence:***
Wouter van Loon
w.s.van.loon@fsw.leidenuniv.nl

Specialty section:
This article was submitted to Brain Imaging Methods, a section of the journal Frontiers in Neuroscience

Received: *07 December 2021*
Accepted: *23 March 2022*
Published: *25 April 2022*

Citation:
van Loon W, de Vos F, Fokkema M, Szabo B, Koini M, Schmidt R and de Rooij M (2022) Analyzing Hierarchical Multi-View MRI Data With StaPLR: An Application to Alzheimer's Disease Classification. Front. Neurosci. 16:830630. doi: 10.3389/fnins.2022.830630

1. INTRODUCTION

In biomedical research, the integration of data from different sources into a single classification model is becoming increasingly common (Fratello et al., 2017; Li et al., 2018). This is fueled by the increasing availability of multi-source data, for example through the UK Biobank (Sudlow et al., 2015; Littlejohns et al., 2020), the Alzheimer's Disease Neuroimaging Initiative (ADNI) (Mueller et al., 2005), and various dementia registries around the world (Krysinska et al., 2017) such as the Prospective Registry on Dementia (PRODEM) (Seiler et al., 2012). Training a model on multiple data sources has been found to increase accuracy in the prediction of brain-age (Liem et al., 2017) and the classification of Alzheimer's disease (AD) (Li et al., 2011; Rahim et al., 2016; Schouten et al., 2016).

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* (Zhao et al., 2017; Sun et al., 2019). Of particular interest to this study is the multi-view learning framework known as *multi-view stacking* (Li et al., 2011; Garcia-Ceja et al., 2018; van Loon et al., 2020a). 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, and have shown better performance than single-view or non-stacked approaches (Li et al., 2011; de Vos et al., 2016; Rahim et al., 2016; Liem et al., 2017; Salvador et al., 2019; Engemann et al., 2020; Guggenmos et al., 2020; Ali et al., 2021). However, these methods are generally *ad-hoc* approaches tailored specifically to the data at hand, and there is little consistency between the used methods.

Although previous studies have mostly focused on improving classification accuracy, 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 (van Loon et al., 2020a). StaPLR essentially integrates the penalized logistic regression models which are already commonly used in neuroimaging classification, such as ridge regression (Hoerl and Kennard, 1970; Le Cessie and Van Houwelingen, 1992) and the lasso (Tibshirani, 1996; Friedman et al., 2010), 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 (van Loon et al., 2020a). StaPLR is, to our knowledge, the only multi-view learning method which can be extended to hierarchical multi-view structures 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 subviews. 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 propose an extension of the StaPLR method to hierarchical multi-view data. We show an application of this extension 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 (Seiler et al., 2012) and the Austrian Stroke Prevention Study (ASPS) (Schmidt et al., 1994; Freudenberger et al., 2016), have been used in previous studies (Schouten et al., 2016, 2017; de Vos et al., 2017), 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 (van Loon et al., 2020a,b). In this paper, we therefore extend 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 (Zou and Hastie, 2005), which has been used in many previous multi-view neuroimaging classification studies (Trzepacz et al., 2014; Teipel et al., 2015; Bowman et al., 2016; de Vos et al., 2016; Nir et al., 2016; Schouten et al., 2016). We also compare the proposed extension with the original StaPLR algorithm.

In addition to its advantages in view selection and computation time (van Loon et al., 2020a), 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 (Liem et al., 2017; Engemann et al., 2020). 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 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.

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 (Seiler et al., 2012). 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 (Schmidt et al., 1994; Freudenberger et al., 2016). 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. Gray matter density refers to the percentage of gray matter in a certain area of the brain. The 48 features correspond to the 48 regions of the probabilistic Harvard-Oxford cortical atlas (Center for Morphometric Analysis, 2006). Subcortical volumes describe the size of several subcortical brain structures. The 14 features correspond to the thalamus, caudate, putamen, pallidum, hippocampus, amygdala, and accumbens of both hemispheres. The neocortex was parcellated into the 68 regions of the Desikan-Killiany atlas (Desikan et al., 2006). 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 (Hua et al., 2008).

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 (de Vos et al., 2017). 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. (2016).

TABLE 1 | Overview of the scan types, MRI measures, and corresponding number of features used in this study.

(s) scan type	(v) MRI measure	number of features
(1) structural MRI	(1) gray 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[**]

*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 (Zou and Hastie, 2005). Elastic net regression employs a mixture of L_1 and L_2 penalties on the vector of regression coefficients (Zou and Hastie, 2005). 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. For α, the set of candidate values is a sequence from 0 to 1 in increments of 0.1. The set of candidate values for λ is a sequence of 100 values adaptively chosen by the software (Friedman et al., 2010). In particular, the 100 values are decreasing values on a log scale from $\lambda_{\max}$ to $\lambda_{\min}$, where $\lambda_{\max}$ is the smallest value such that the entire regression coefficient vector is zero, and $\lambda_{\min} = \epsilon\lambda_{\max}$ (Friedman et al., 2010). We set $\epsilon = 0.01$ (the default). To prevent overfitting, we assessed the models classification performance using a double (nested) cross-validation approach (Varma and Simon, 2006): 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 (Team, 2017), using the package `glmnet` 1.9–8 (Friedman et al., 2010).

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**. Thus, the data have a hierarchical multi-view structure with three levels, and we therefore propose an extension of the StaPLR algorithm to three levels. The previous version of StaPLR only allowed for a two-level structure, where features were nested within views (van Loon et al., 2020a). This means one has to choose to either use the MRI measures as views, or the scan types. Thus, one would have to either ignore part of the hierarchy, or perform two separate analyses. **Algorithm 1** presents the extension of StaPLR to 3 levels: 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 gray 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 (Wolpert, 1992). 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 hierarchical

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

Data: $\mathbf{X}^{(v)}$, $v = 1 \ldots 40$, the 40 different MRI measures as shown in **Table 1**, and 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 $(\mathbf{X}^{(v)}, y)$, $v = 1, \ldots, 40$, to obtain view-specific classifiers $\hat{f}_1, \ldots, \hat{f}_{40}$.
2. Apply 10-fold cross-validation to obtain a vector of predictions $z^{(v)}$ 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 $z^{(v)}$ which correspond to that scan type column-wise into the matrix $Z^{(s)}$.
4. Train a logistic nonnegative lasso classifier (including cross-validation for λ) on the pairs $(Z^{(s)}, y)$, $s = 1, 2, 3$, 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 $z^{(s)}_{\text{inter}}$ for each of the $\hat{f}^{(1)}_{\text{inter}}, \hat{f}^{(2)}_{\text{inter}}, \hat{f}^{(3)}_{\text{inter}}$.
6. Collect the predictions $z^{(1)}_{\text{inter}}$, $z^{(2)}_{\text{inter}}$, $z^{(3)}_{\text{inter}}$ column-wise into the matrix Z_{inter}.
7. Train a logistic nonnegative lasso classifier (including cross-validation for λ) on the pair (Z_{inter}, y) to obtain a meta-classifier $\hat{f}_{\text{meta}}$.
8. Define the final stacked classifier as:

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

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. (2020a) 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 training 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).

The model is again evaluated using double (nested) 10-fold cross-validation: In the outer validation loop, the entirety of **Algorithm 1** is applied to 90% of the data, and the remaining 10% is used only for the calculation of classification accuracy and AUC. This procedure was repeated 10 times. StaPLR was performed in R using the package `multiview` 0.3.1 (van Loon, 2021), using the default optimization settings. The scripts used for model fitting and evaluation are available in a public code repository (van Loon, 2022). For a more general discussion of the original StaPLR algorithm with only 2 levels we refer to van Loon et al. (2020a).

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 hierarchical 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* (Dick, 2002). 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 $\mathbf{X}^{[s,i]}$, 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 $\mathbf{X}^{[s,i]}, i = 1 \ldots m_s$, $s = 1 \ldots S$, the ith MRI measure of scan type s, with m_s the total number of measures corresponding to scan type s. Denote by $\hat{\beta}_0^{[s]}, s = 1 \ldots S$ the intercept of the intermediate classifier corresponding to scan type s. Denote by $\hat{\beta}_i^{[s]}, i = 1 \ldots m_s, s = 1 \ldots S$ the coefficient of the ith MRI measure of scan type s. Denote by $\hat{\omega}_0$ the intercept of the meta-classifier, and by $\hat{\omega}^{[s]}, s = 1 \ldots S$ the weight of scan type s. Then for the ith MRI measure corresponding to scan type s, we define the MRM as:

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

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

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

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 (Breiman, 2001; Fisher et al., 2019) or SHAP values (Lundberg and Lee, 2017).

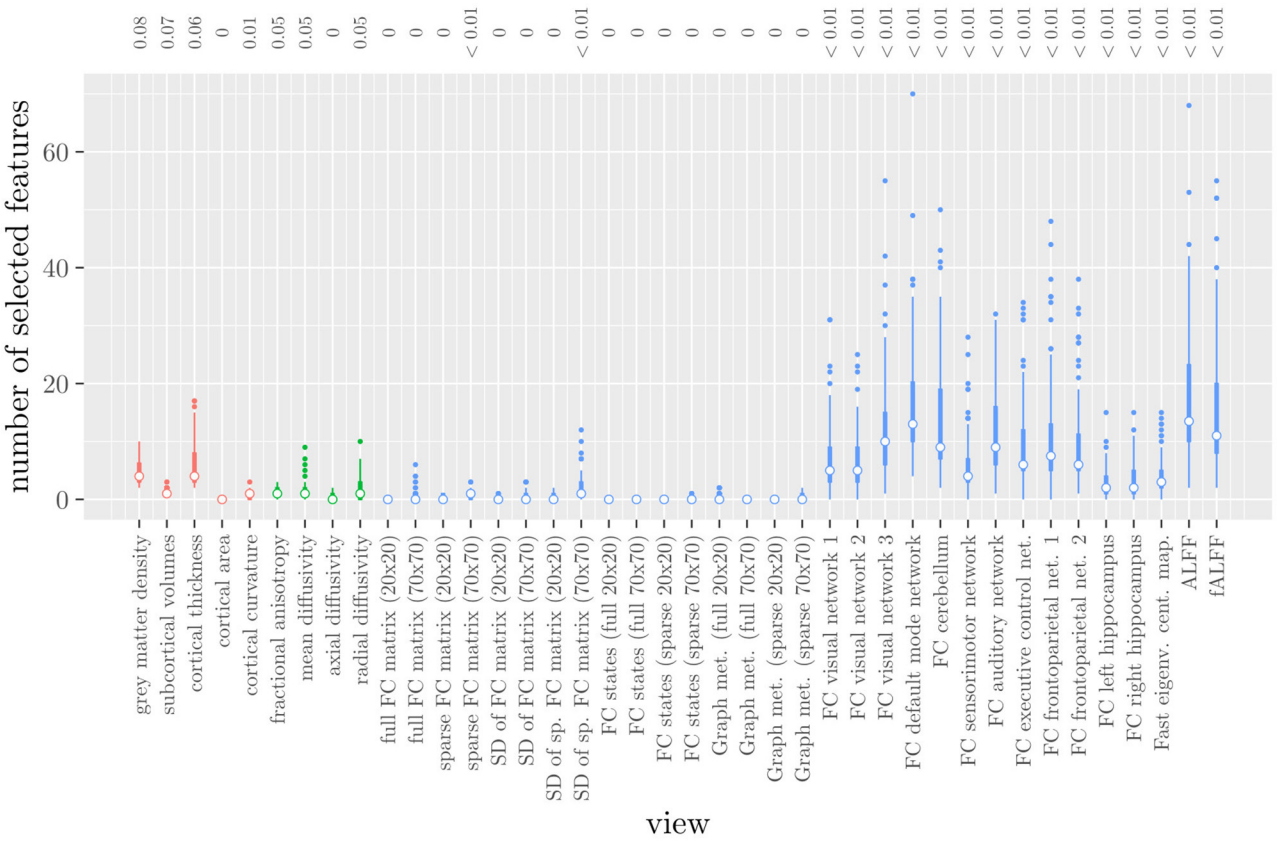

FIGURE 1 | Boxplots of the number of selected features for each MRI measure resulting from the elastic net regression, colored by scan type (red = structural MRI, green = diffusion MRI, blue = resting state fMRI). The numbers at the top of the graph denote the median proportion of nonzero coefficients.

3. RESULTS

3.1. Logistic 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**).

These results highlight several drawbacks of elastic net regression for multi-view data: 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

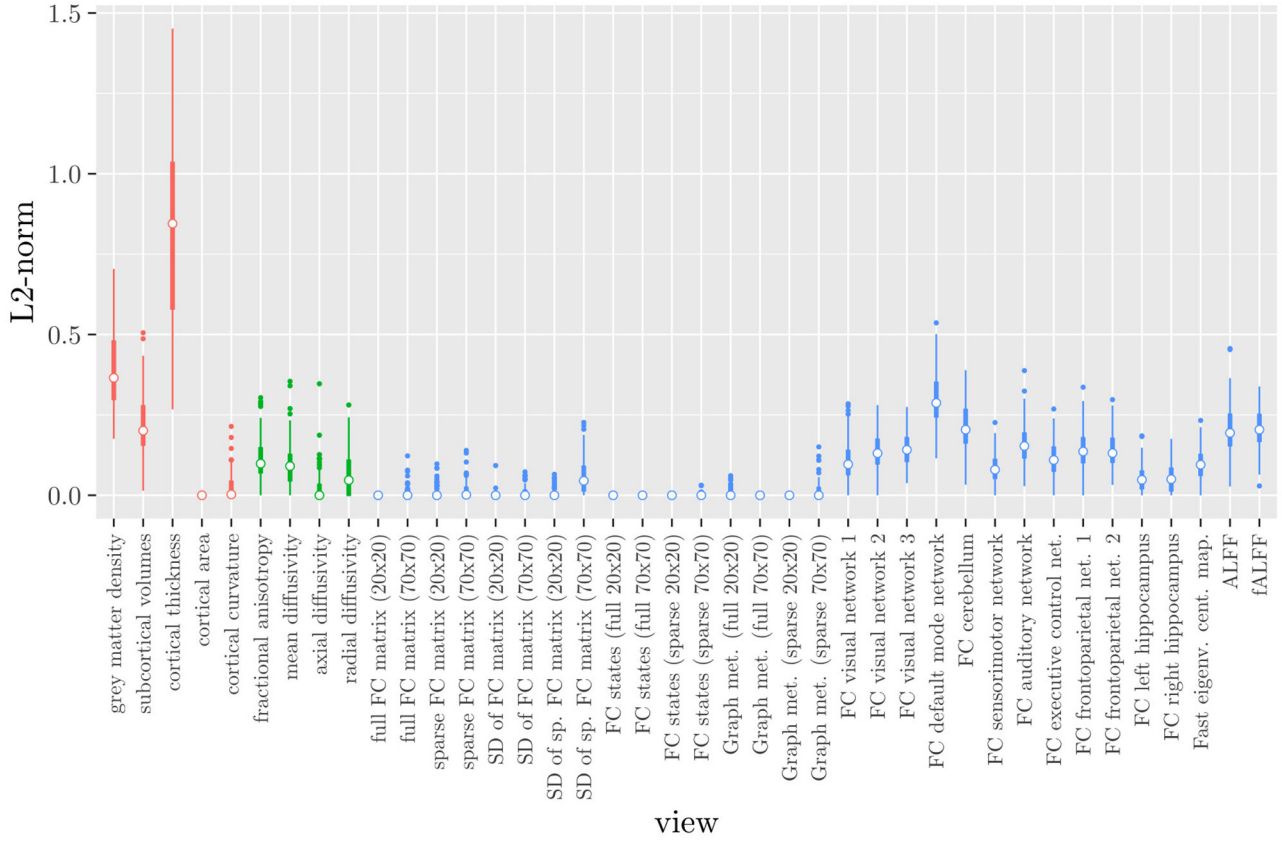

FIGURE 2 | Boxplots of the L_2-norm of the regression coefficient vector for each MRI measure resulting from the elastic net regression, colored by scan type (red = structural MRI, green = diffusion MRI, blue = resting state fMRI).

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 gray matter density and cortical thickness which have the largest L_2-norm. Although **Figures 1**, **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.

3.2. Original StaPLR Algorithm

The original StaPLR algorithm only allows for a two-level structure, with features nested within views. Thus, one has to choose between using the MRI measures as views, or the scan types. Here, we show the results of both choices.

3.2.1. MRI Measures Only

The mean AUC of the model using the MRI measures as views was 0.942 (SD = 0.006). The mean accuracy was 0.888 (SD = 0.007). The median regression coefficient for each MRI measure, as well as their distribution across the 10 × 10 fitted stacked classifiers can be observed in **Figure 3**. The resulting classifier is considerably sparser than the one obtained through elastic net regression (**Figure 2**). Note that this analysis only gives a measure of importance for each MRI measure, not for the scan types.

3.2.2. Scan Types Only

The mean AUC of the model using the scan types as views was 0.942 (SD = 0.008). The mean accuracy was 0.897 (SD = 0.006). 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 4**. Structural MRI obtains the highest coefficient, followed by diffusion MRI. Resting state fMRI is never selected. Note that this analysis does not perform selection of MRI measures, only of scan types. Thus, it is not possible to select a subset of relevant MRI measures for any scan type; the complete scan type has to be included or excluded from the model.

3.3. Hierarchical StaPLR

The mean AUC of the hierarchical 3-level StaPLR was 0.942 (SD = 0.006), which was higher than that of the elastic net (0.922, SD = 0.008), and identical to that of the original StaPLR algorithm applied to either the MRI measures or the scan types. The mean

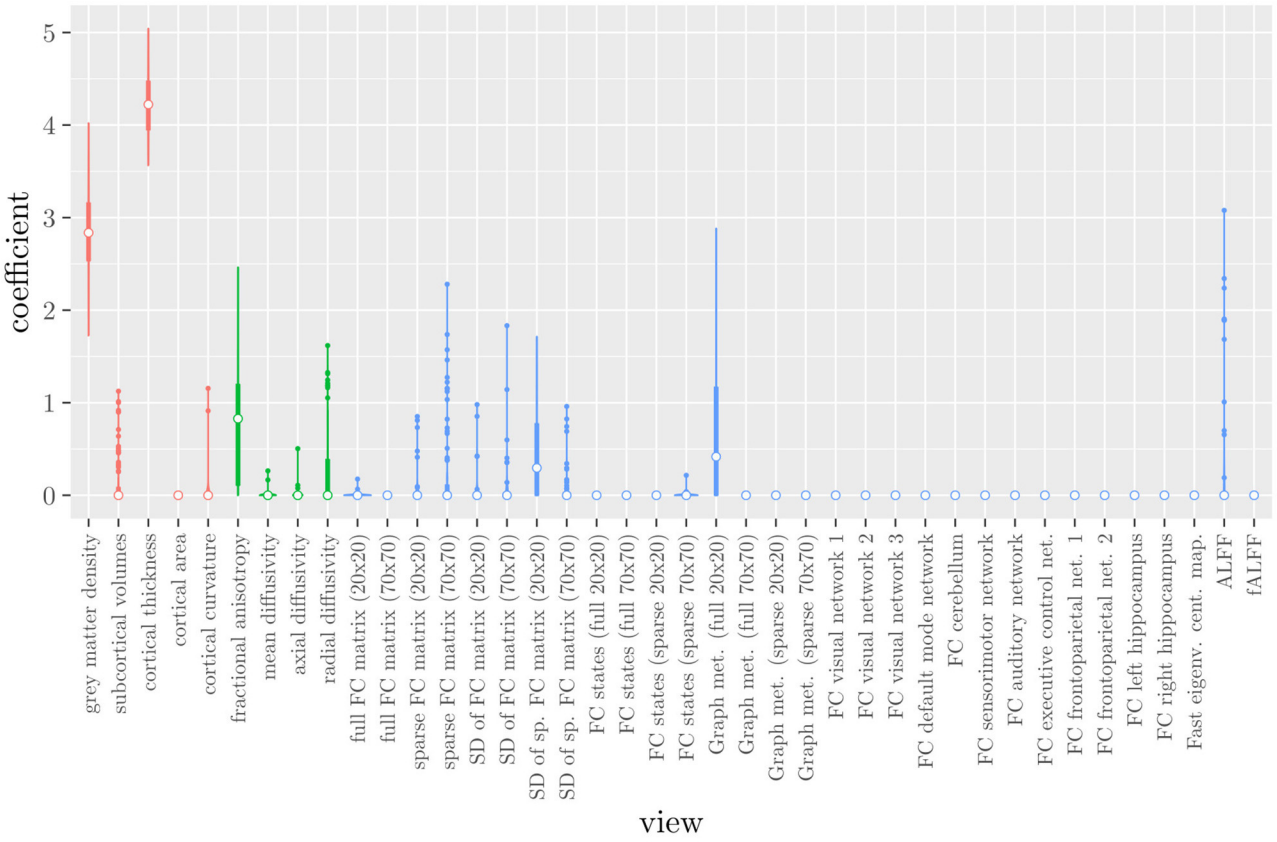

FIGURE 3 | Boxplots of the regression coefficients of StaPLR applied with only 2 levels, using the MRI measures as views, colored by scan type (red = structural MRI, green = diffusion MRI, blue = resting state fMRI).

accuracy was 0.893 (SD = 0.008), which was also higher than that of the elastic net (0.848, SD = 0.012), and comparable to that of the original StaPLR algorithm. 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 5**. 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 5**, 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 6** shows that within the structural MRI scan type, the measures of cortical thickness and gray 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 7**, **8** 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 6**–**8**, 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 gray 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 5**). 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 9**, the

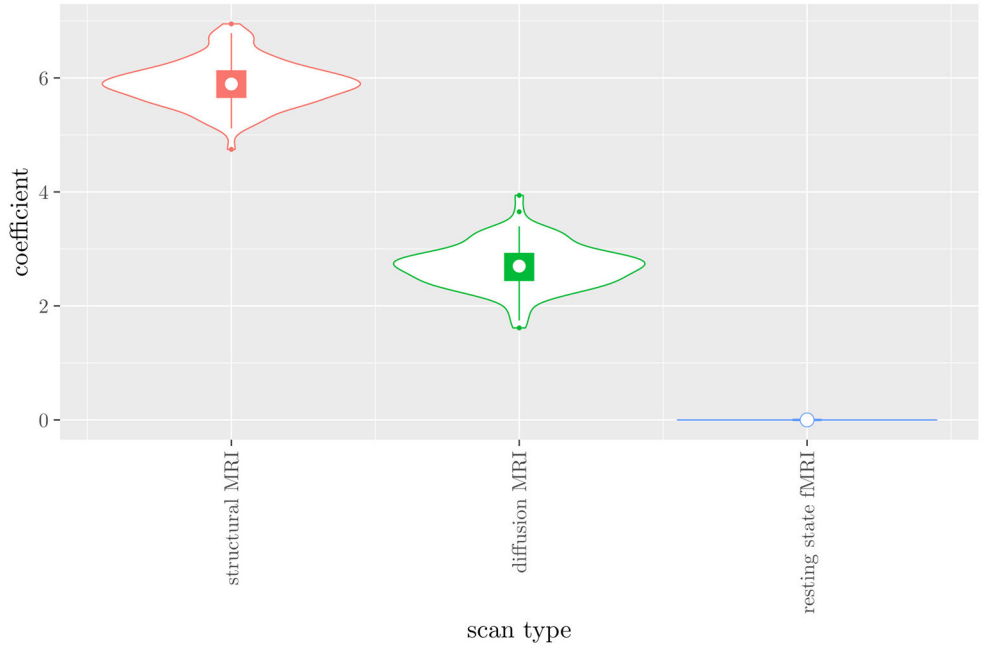

FIGURE 4 | Box-and-violin plots of the regression coefficients of StaPLR applied with only 2 levels, using the scan types as views.

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 hierarchical StaPLR with the results of elastic net regression (**Figures 1**, **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 gray matter density, and 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." Within the scan types which have a smaller contribution, i.e., diffusion-weighted MRI and resting state fMRI, we can see 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.

4. DISCUSSION

We have extended the StaPLR algorithm to adapt to a hierarchical multi-view data structure. That is, the extended StaPLR algorithm allows for the analysis of datasets containing large numbers of features which are nested within lower-level views, which are in turn nested within higher-level views. We applied this extension to a multi-view MRI data set in the context of Alzheimer's disease classification, where features are nested within MRI measures, which are in turn nested within scan types. The presented application can serve as an example of a more general class of applications within neuroimaging and the biomedical sciences. 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.

If we compare the results of the hierarchical extension of StaPLR with the original algorithm applied to either MRI measures or scan types, we see three different resulting classifiers which have nearly identical classification performance. Naturally, the results of the StaPLR algorithm depend on the specified multi-view structure, and specifying a different structure will lead to a different model. Since MRI data generally contain high amounts of collinearity it is not all that surprising that different models may have similar classification performance. However, it is still natural to ask which of these three models

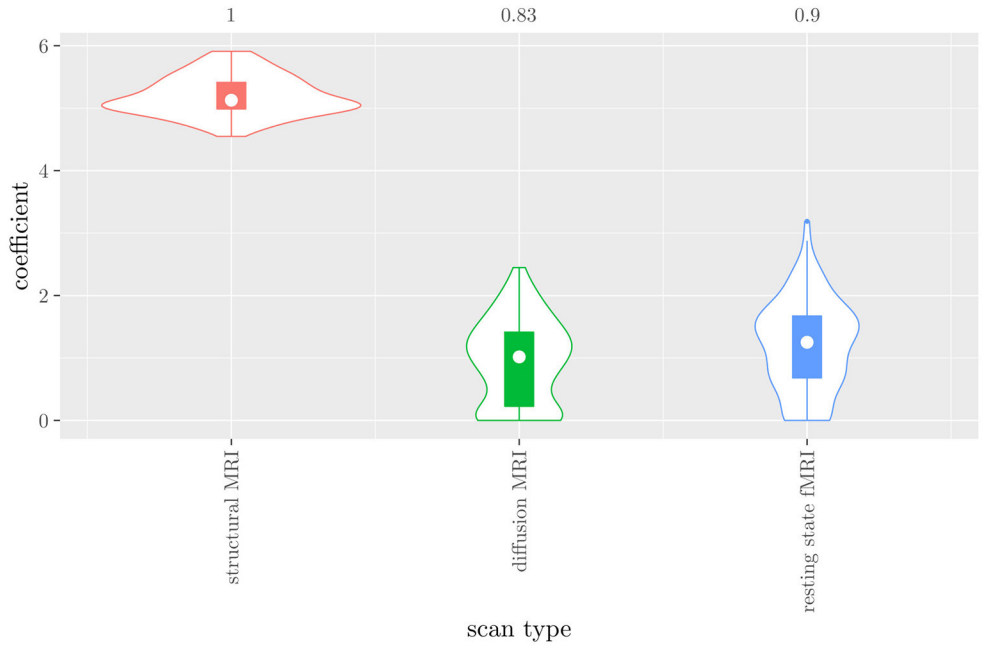

FIGURE 5 | Box-and-violin plots of the hierarchical StaPLR meta-level regression coefficients for each scan type. The numbers at the top of the graph denote the proportion of times the coefficient was nonzero.

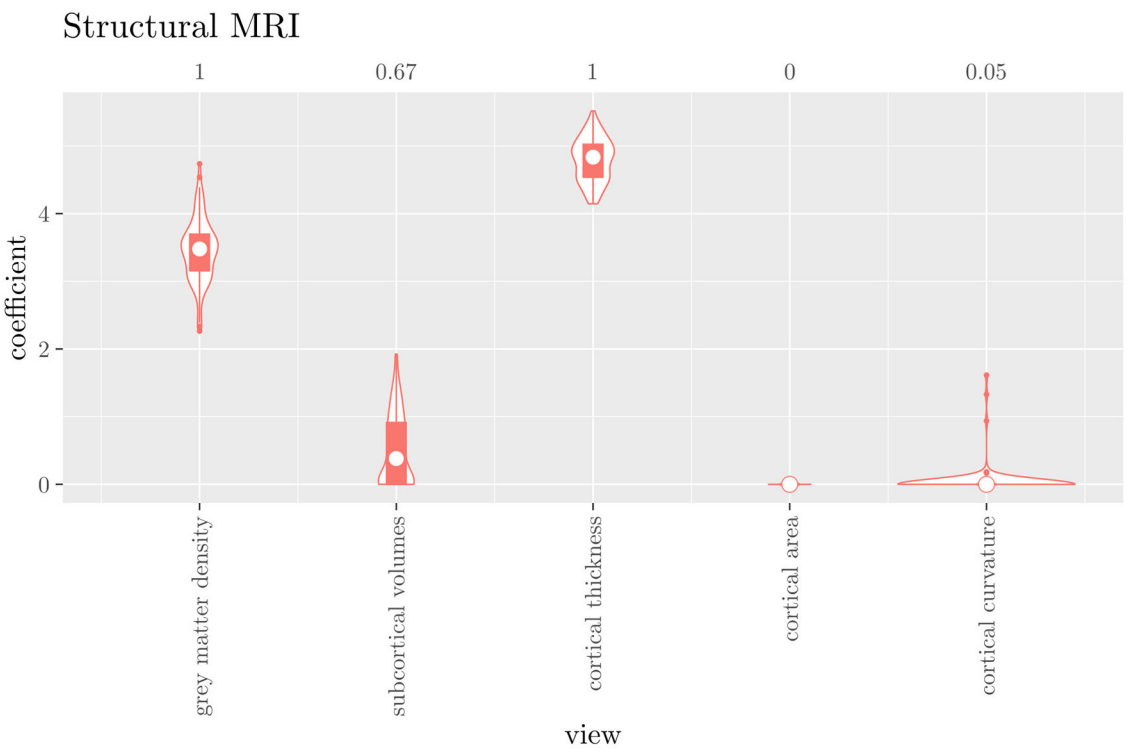

FIGURE 6 | Box-and-violin plots of the hierarchical StaPLR intermediate-level regression coefficients for the structural scan type. The numbers at the top of the graph denote the proportion of times the coefficient was nonzero.

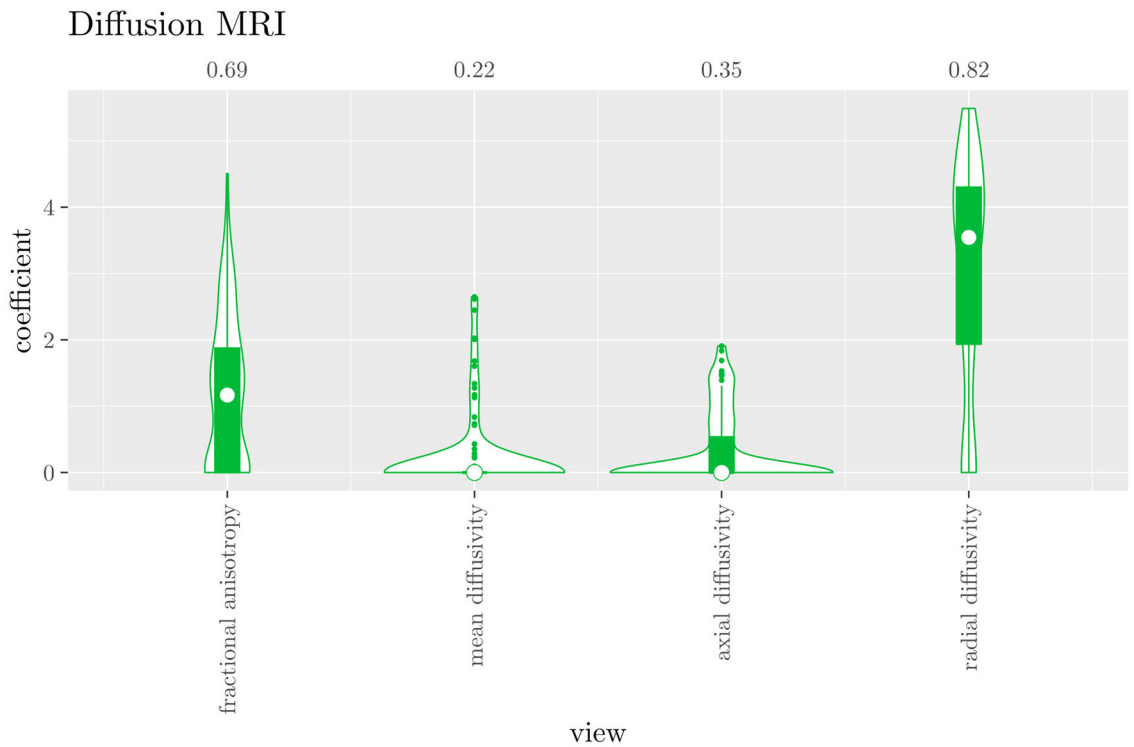

FIGURE 7 | Box-and-violin plots of the hierarchical StaPLR intermediate-level regression coefficients for the diffusion scan type. The numbers at the top of the graph denote the proportion of times the coefficient was nonzero.

is the "best." Since all three models have nearly identical classification performance, and statistical models are generally "wrong" in the sense that they can only provide a simplification or approximation of reality, the most important question is probably: which approach is the most useful? We argue that the hierarchical StaPLR model is more useful for several reasons. First, hierarchical StaPLR is the only approach that accurately matches the study design because it most closely follows how the data are collected and processed: data are collected through different MRI scans, and then researches derive different MRI measures from these scan types. Second, unlike the original StaPLR algorithm, hierarchical StaPLR allows easy computation of measures of importance for both MRI measures and scan types. If applying the original StaPLR algorithm, two separate analyses are required, which considerably increases computational cost. In addition, the two different analyses may lead to different conclusions, as was the case in our data set (compare **Figure 3** with **4**). If the MRI measures are used as views, some resting state fMRI measures are generally included, but if the scan types are used instead, the fMRI scan type is always discarded. This discrepancy is probably caused by the fact that when the scan types are used as views, the algorithm is forced to select or discard all features within this scan type. Since the fMRI scan type has a very high number of features and likely a low signal-to-noise ratio, the addition of a large amount of noise to the model is not worth the inclusion of the scan type by the two-level StaPLR algorithm. However, if a selection of MRI measures is made first, such as in the hierarchical StaPLR model, then signal present in the fMRI scan type can still be picked up by the algorithm.

Given the results of the hierarchical StaPLR algorithm, we can compare the relative importance of the different MRI measures using the regression coefficients or the MRM. However, we may additionally 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, gray 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 70 x 70 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

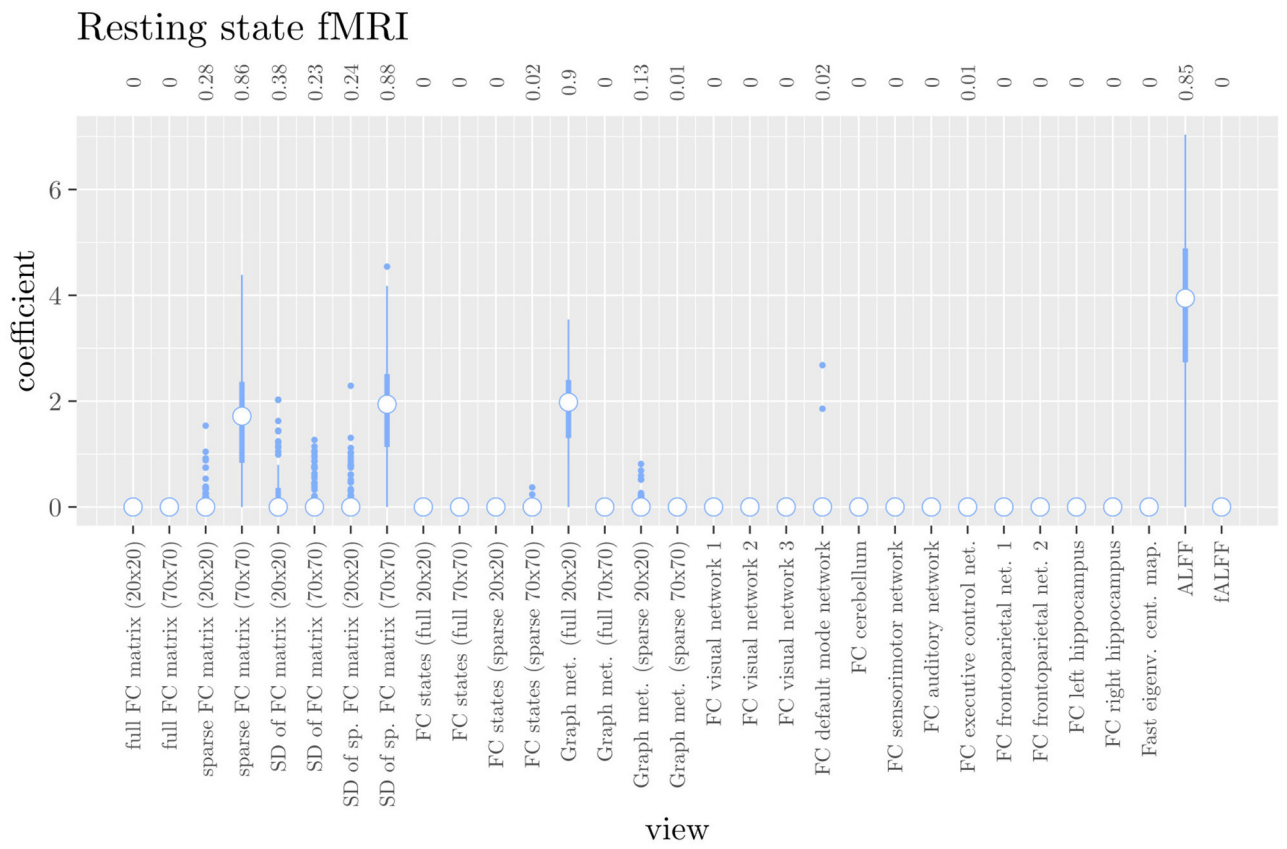

FIGURE 8 | Box plots of the hierarchical StaPLR intermediate-level regression coefficients for the functional scan type. The numbers at the top of the graph denote the proportion of times the coefficient was nonzero.

a "frequently selected" and "infrequently selected" group of MRI measures.

Of course, binary decisions regarding which MRI measures or scan types are required for prediction are further complicated by the fact that we have shown different two-level StaPLR models with comparable classification performance. However, it is important to note that these models correspond to different research questions. For example, when using only the scan types as views, the research question pertains to the relevance of *complete* scan types (i.e., including any noise), rather than to the most informative subset of MRI measures derived from those scan types, as in the hierarchical StaPLR model. Naturally, a different research question may lead to a different answer.

The results of hierarchical StaPLR model are in line with earlier research. Hierarchical StaPLR, the original StaPLR algorithm using MRI measures as views, and the elastic net all appeared to agree on the structural MRI measures of gray matter density and cortical thickness being the most important for classification, which is in line with earlier research identifying measures of gray matter atrophy as important bio-markers for Alzheimer's disease (Lerch et al., 2005; Frisoni et al., 2010). The largest difference between the methods was seen in terms of fMRI measures, of which hierarchical StaPLR selected 4, StaPLR using only the MRI measures selected 2, StaPLR using only the scan types selected none, and the elastic net selected 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, hierarchical 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 (de Vos et al., 2017). The results of the hierarchical 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 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 (Schouten et al., 2016).

Although hierarchical 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.

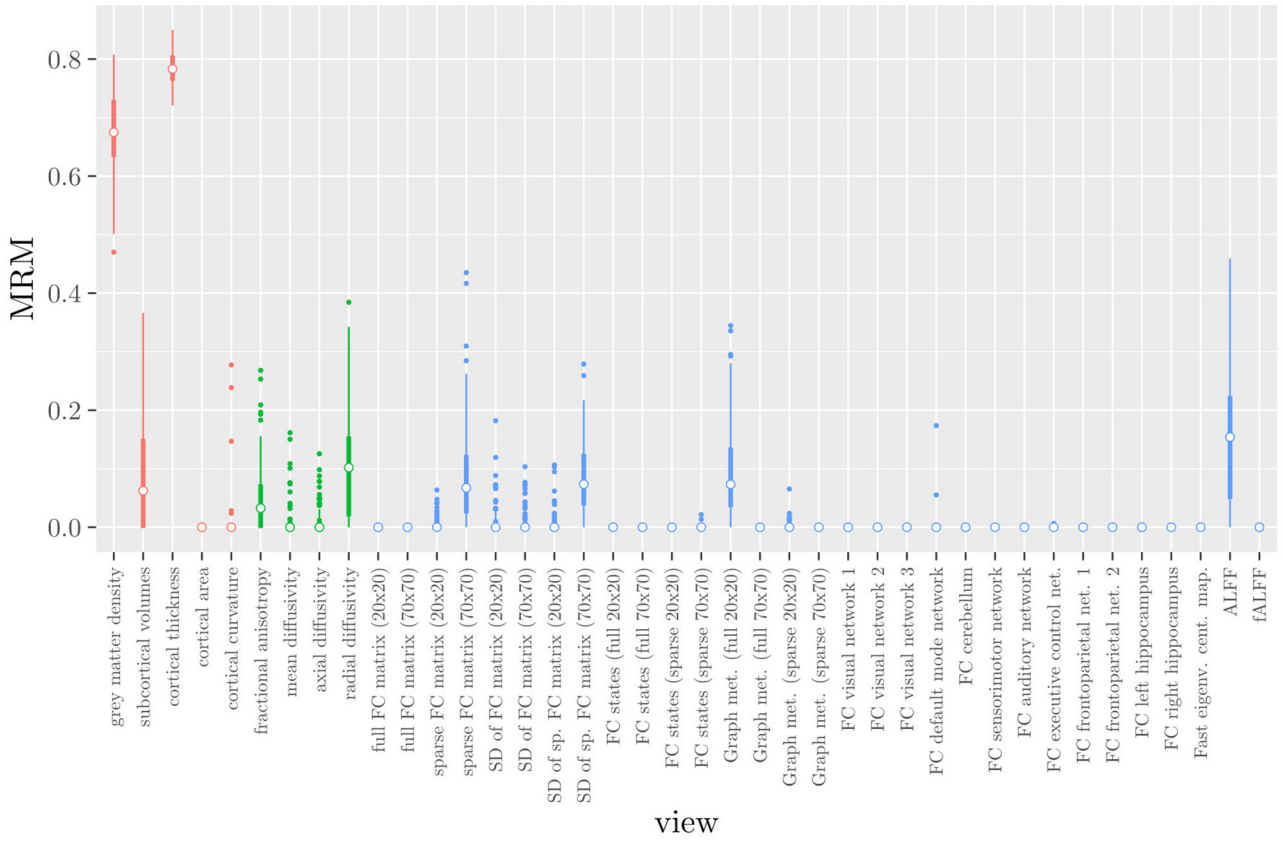

FIGURE 9 | Influence of each of the MRI measures on the predictions of the hierarchical stacked classifier, as quantified by the minority report measure. MRM values range from 0 to 1, with 0 indicating no influence and 1 indicating maximum possible influence. The colors of the bars refer to the different scan types (red = structural MRI, green = diffusion MRI, blue = resting state fMRI).

Furthermore, our results indicate that StaPLR can adequately deal with imbalanced numbers of features within each view. Whereas standard elastic net tends to select many features from very large views (e.g., the functional connectivity views; **Figure 1**), StaPLR does not show this preference for large views (most functional connectivity views obtained a weight of 0 as shown in **Figure 8**). 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 (De Bin et al., 2016). 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.

As mentioned before, 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 it matches the study design. However, one could specify a different multi-view structure to match a different research question. In fact, we have done so when applying StaPLR with only two levels. Another example of a different research question would occur if the primary interest is in identifying which *brain areas* are the most important for AD classification. In this case, one could treat each brain area as a separate view. This may, of course, again lead to different results. For example, the feature set that consists of the volumes of the subcortical structures was found to play only a minor role in AD classification in the hierarchical StaPLR model, whereas this feature set also contains the volumes of the left and right hippocampus that are considered to be AD hallmarks. Decoupling the volumes of the different subcortical structures and treating each brain area as a separate view would allow 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.

DATA AVAILABILITY STATEMENT

The data analyzed in this study is subject to the following licenses/restrictions: a formal data sharing agreement is mandatory. Requests to access these datasets should be directed to RS, reinhold.schmidt@medunigraz.at.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Ethics Committee of the Medical University of Graz (ASPS, PRODEM), and Ethics Committees of 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 (PRODEM). The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

WL: conceptualization, methodology, software, formal analysis, investigation, writing—original draft, writing—review and editing, and visualization. FV: investigation, data curation, and writing—review and editing. MF and BS: writing—review and editing. MK and RS: data curation. MR: conceptualization, writing—review and editing, and supervision.

FUNDING

This research was funded by Leiden University. BS received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 101041064).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnins.2022.830630/full#supplementary-material

REFERENCES

Ali, L., He, Z., Cao, W., Rauf, H. T., Imrana, Y., and Heyat, M. B. B. (2021). MMDD-ensemble: a multimodal data driven ensemble approach for Parkinson's disease detection. *Front. Neurosci.* 15, 1–11. doi: 10.3389/fnins.2021.754058

Bowman, F. D., Drake, D. F., and Huddleston, D. E. (2016). Multimodal imaging signatures of Parkinson's disease. *Front. Neurosci.* 10, 1–11. doi: 10.3389/fnins.2016.00131

Breiman, L. (2001). Random forests. *Mach. Learn.* 45, 5–32. doi: 10.1023/A:1010933404324

Center for Morphometric Analysis (CMA) at Massachusetts General Hospital (MGH), Harvard Medical School. (2006). *Harvard-Oxford Cortical Atlas*. Boston, MA. Available online at: https://cma.mgh.harvard.edu

De Bin, R., Janitza, S., Sauerbrei, W., and Boulesteix, A.-L. (2016). Subsampling versus bootstrapping in resampling-based model selection for multivariable regression. *Biometrics* 72, 272–280. doi: 10.1111/biom.12381

de Vos, F., Koini, M., Schouten, T., Seiler, S., van der Grond, J., Lechner, A., et al. (2017). A comprehensive analysis of resting state fMRI measures to classify individual patients with Alzheimer's disease. *NeuroImage* 167, 62–72. doi: 10.1016/j.neuroimage.2017.11.025

de Vos, F., Schouten, T., Hafkemeijer, A., Dopper, E., van Swieten, J., de Rooij, M., et al. (2016). Combining multiple anatomical MRI measures improves Alzheimer's disease classification. *Hum. Brain Map.* 37, 1920–1929. doi: 10.1002/hbm.23147

Desikan, R. S., Ségonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., et al. (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *NeuroImage* 31, 968–980. doi: 10.1016/j.neuroimage.2006.01.021

Dick, P. K. (2002). *The Minority Report*. New York, NY: Kensington Publishing.

Engemann, D. A., Kozynets, O., Sabbagh, D., Lemaître, G., Varoquaux, G., Liem, F., et al. (2020). Combining magnetoencephalography with magnetic resonance imaging enhances learning of surrogate-biomarkers. *eLife* 9, 1–32. doi: 10.7554/eLife.54055

Fisher, A., Rudin, C., and Dominici, F. (2019). All models are wrong, but many are useful: learning a variable's importance by studying an entire class of prediction models simultaneously. *J. Mach. Learn. Res.* 20, 1–81. Available online at: https://www.jmlr.org/papers/volume20/18-760/18-760.pdf

Fratello, M., Caiazzo, G., Trojsi, F., Russo, A., Tedeschi, G., Tagliaferri, R., et al. (2017). Multi-view ensemble classification of brain connectivity images for neurodegeneration type discrimination. *Neuroinformatics* 15, 199–213. doi: 10.1007/s12021-017-9324-2

Freudenberger, P., Petrovic, K., Sen, A., Töglhofer, A. M., Fixa, A., Hofer, E., et al. (2016). Fitness and cognition in the elderly: the Austrian stroke prevention study. *Neurology* 86, 418–424. doi: 10.1212/WNL.0000000000002329

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* 33, 1–22. doi: 10.18637/jss.v033.i01

Frisoni, G. B., Fox, N. C., Jack, C. R., Scheltens, P., and Thompson, P. M. (2010). The clinical use of structural MRI in Alzheimer disease. *Nat. Rev. Neurol.* 6, 67–77. doi: 10.1038/nrneurol.2009.215

Garcia-Ceja, E., Galván-Tejada, C. E., and Brena, R. (2018). Multi-view stacking for activity recognition with sound and accelerometer data. *Inf. Fusion* 40, 45–56. doi: 10.1016/j.inffus.2017.06.004

Guggenmos, M., Schmack, K., Veer, I. M., Lett, T., Sekutowicz, M., Sebold, M., et al. (2020). A multimodal neuroimaging classifier for alcohol dependence. *Sci. Rep.* 10, 1–12. doi: 10.1038/s41598-019-56923-9

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* 12, 55–67.

Hua, K., Zhang, J., Wakana, S., Jiang, H., Li, X., Reich, D. S., et al. (2008). Tract probability maps in stereotaxic spaces: analyses of white matter anatomy and tract-specific quantification. *NeuroImage* 39, 336–347. doi: 10.1016/j.neuroimage.2007.07.053

Krysinska, K., Sachdev, P. S., Breitner, J., Kivipelto, M., Kukull, W., and Brodaty, H. (2017). Dementia registries around the globe and their applications: a systematic review. *Alzheimer's Dementia* 13, 1031–1047. doi: 10.1016/j.jalz.2017.04.005

Le Cessie, S., and Van Houwelingen, J. C. (1992). Ridge estimators in logistic regression. *J. R. Stat. Soc. Series C (Appl. Stat.)* 41, 191–201.

Lerch, J. P., Pruessner, J. C., Zijdenbos, A., Hampel, H., Teipel, S. J., and Evans, A. C. (2005). Focal decline of cortical thickness in Alzheimer's disease identified by computational neuroanatomy. *Cereb. Cortex* 15, 995–1001. doi: 10.1093/cercor/bhh200

Li, R., Hapfelmeier, A., Schmidt, J., Perneczky, R., Drzezga, A., Kurz, A., et al. (2011). "A case study of stacked multi-view learning in dementia research," in *13th Conference on Artificial Intelligence in Medicine* (Bled), 60–69.

Li, Y., Wu, F.-X., and Ngom, A. (2018). A review on machine learning principles for multi-view biological data integration. *Briefings Bioinf.* 19, 325–340. doi: 10.1093/bib/bbw113

Liem, F., Varoquaux, G., Kynast, J., Beyer, F., Masouleh, S. K., Huntenburg, J. M., et al. (2017). Predicting brain-age from multimodal imaging data captures cognitive impairment. *NeuroImage* 148, 179–188. doi: 10.1016/j.neuroimage.2016.11.005

Littlejohns, T. J., Holliday, J., Gibson, L. M., Garratt, S., Oesingmann, N., Alfaro-Almagro, F., et al. (2020). The UK biobank imaging enhancement of 100,000 participants: rationale, data collection, management and future directions. *Nat. Commun.* 11, 1–12. doi: 10.1038/s41467-020-15948-9

Lundberg, S. M., and Lee, S.-I. (2017). "A unified approach to interpreting model predictions," in *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Long Beach, CA), 4768–4777.

Mueller, S. G., Weiner, M. W., Thal, L. J., Petersen, R. C., Jack, C., Jagust, W., et al. (2005). The Alzheimer's disease neuroimaging initiative. *Neuroimag. Clin. North America* 15, 869. doi: 10.1016/j.nic.2005.09.008

Nir, T. M., Villalon-Reina, J. E., Gutman, B. A., Moyer, D., Jahanshad, N., Dehghani, M., et al. (2016). "Alzheimer's disease classification with novel microstructural metrics from diffusion-weighted MRI," in *Computational Diffusion MRI*, Cham: Springer. 41–54.

Rahim, M., Thirion, B., Comtat, C., and Varoquaux, G. (2016). Transmodal learning of functional networks for Alzheimer's disease prediction. *IEEE J. Sel. Top. Signal Process.* 10, 1204–1213. doi: 10.1109/JSTSP.2016.2600400

Salvador, R., Canales-Rodríguez, E., Guerrero-Pedraza, A., Sarró, S., Tordesillas-Gutiérrez, D., Maristany, T., Crespo-Facorro, B., et al. (2019). Multimodal integration of brain images for MRI-based diagnosis in schizophrenia. *Front. Neurosci.* 13, 1–9. doi: 10.3389/fnins.2019.01203

Schmidt, R., Lechner, H., Fazekas, F., Niederkorn, K., Reinhart, B., Grieshofer, P., et al. (1994). Assessment of cerebrovascular risk profiles in healthy persons: definition of research goals and the Austrian stroke prevention study (ASPS). *Neuroepidemiology* 13, 308–313.

Schouten, T., Koini, M., De Vos, F., Seiler, S., van der Grond, J., Lechner, A., et al. (2016). Combining anatomical, diffusion, and resting state functional magnetic resonance imaging for individual classification of mild and moderate Alzheimer's disease. *NeuroImage Clin.* 11, 46–51. doi: 10.1016/j.nicl.2016.01.002

Schouten, T. M., Koini, M., de Vos, F., Seiler, S., de Rooij, M., Lechner, A., et al. (2017). Individual classification of Alzheimer's disease with diffusion magnetic resonance imaging. *NeuroImage* 152, 476–481. doi: 10.1016/j.neuroimage.2017.03.025

Seiler, S., Schmidt, H., Lechner, A., Benke, T., Sanin, G., Ransmayr, G., et al. (2012). Driving cessation and dementia: results of the prospective registry on dementia in Austria (PRODEM). *PLoS ONE* 7, e52710. doi: 10.1371/journal.pone.0052710

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., et al. (2015). UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* 12, e1001779. doi: 10.1371/journal.pmed.1001779

Sun, S., Mao, L., Dong, Z., and Wu, L. (2019). *Multiview Machine Learning*. Singapore: Springer.

Team, R. C. (2017). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Teipel, S. J., Kurth, J., Krause, B., Grothe, M. J., ADNI, et al. (2015). The relative importance of imaging markers for the prediction of Alzheimer's disease dementia in mild cognitive impairment – beyond classical regression. *NeuroImage Clin.* 8, 583–593. doi: 10.1016/j.nicl.2015.05.006

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Series B* 58, 267–288.

Trzepacz, P. T., Yu, P., Sun, J., Schuh, K., Case, M., Witte, M. M., et al. (2014). Comparison of neuroimaging modalities for the prediction of conversion from mild cognitive impairment to Alzheimer's dementia. *Neurobiol. Aging* 35, 143–151. doi: 10.1016/j.neurobiolaging.2013.06.018

van Loon, W. (2021). R package 'multiview' - Methods for high-dimensional multi-view learning (v0.3.1). *Zenodo*. doi: 10.5281/zenodo.4630669

van Loon, W. (2022). Code repository accompanying "Analyzing hierarchical multi-view MRI data with StaPLR: An application to Alzheimer's disease classification". *Zenodo*. doi: 10.5281/zenodo.5105729

van Loon, W., Fokkema, M., Szabo, B., and de Rooij, M. (2020a). Stacked penalized logistic regression for selecting views in multi-view learning. *Inf. Fusion* 61, 113–123. doi: 10.1016/j.inffus.2020.03.007

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

Varma, S., and Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. *BMC Bioinf.* 7, 91. doi: 10.1186/1471-2105-7-91

Wolpert, D. H. (1992). Stacked generalization. *Neural Netw.* 5, 241–259.

Zhao, J., Xie, X., Xu, X., and Sun, S. (2017). Multi-view learning overview: recent progress and new challenges. *Inf. Fusion* 38, 43–54. doi: 10.1016/j.inffus.2017.02.007

Zou, H., and Hastie, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Series B* 67, 301–320. doi: 10.1111/j.1467-9868.2005.00503.x

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.