Assessment of Machine Learning Pipelines for Prediction of Behavioral Deficits from Brain Disconnectomes

Marco Zorzi[1,3(✉)], Michele De Filippo De Grazia[3], Elvio Blini[1], and Alberto Testolin[1,2]

[1] Department of General Psychology, University of Padova, 35141 Padova, Italy
{marco.zorzi,alberto.testolin}@unipd.it
[2] Department of Information Engineering, University of Padova, 35141 Padova, Italy
[3] IRCCS San Camillo Hospital, Venice-Lido 30126, Italy

Abstract. Recent studies have shown that brain lesions following stroke can be probabilistically mapped onto disconnections of white matter tracts, and that the resulting "disconnectome" is predictive of the patient's behavioral deficits. Disconnectome maps are sparse, high-dimensional 3D matrices that require unsupervised dimensionality reduction followed by supervised learning for prediction of the associated behavioral data. However, the optimal machine learning pipeline for disconnectome data still needs to be identified. We examined four dimensionality reduction methods at varying levels of compression and used the extracted features as input for cross-validated regularized regression to predict the associated language and motor deficits. Features extracted by Principal Component Analysis and Non-Negative Matrix Factorization were found to be the best predictors, followed by Independent Component Analysis and Dictionary Learning. Optimizing the number of extracted features improved predictive accuracy and greatly reduced model complexity. Moreover, the choice of dimensionality reduction technique was found to optimally combine with a specific type of regularized regression (ridge *vs.* LASSO). Overall, our findings represent an important step towards an optimal pipeline that yields high prediction accuracy with a small number of features, which can also improve model interpretability.

Keywords: Stroke · Structural connectome · Disconnections · Machine learning · Feature extraction · Dimensionality reduction · Predictive modeling

1 Introduction

Stroke is a major cause of serious disability for adults and it can affect multiple behavioral domains, from motor control to language and cognition [1]. A classic approach in cognitive neuroscience is to establish which brain lesion is associated to a specific behavioral deficit [2]. The reverse inference is more challenging because lesion information is used to predict the behavioural performance of new (i.e., held out) patients despite the considerable individual variability of lesion-behavior relationships [3]. Moreover,

M. Mahmud et al. (Eds.): BI 2021, LNAI 12960, pp. 1–12, 2021.
https://doi.org/10.1007/978-3-030-86993-9_20

white-matter lesions cause structural disconnections that produce widespread dysfunction of brain networks [4] and can be better predictors of behavioral deficits than lesion site [1, 5, 6].

Assessing damage to the structural connectome requires complex neuroimaging methods (i.e., diffusion tensor imaging) that are difficult to implement in clinical practice. However, Foulon et al. [7] recently proposed an indirect method for estimating structural disconnection from clinical structural Magnetic Resonance Imaging (MRI) scans. Using the connectome of healthy individuals as reference atlas, the method estimates the probability that a lesion at any given location (voxel) causes disconnection of white matter tracts. Therefore, a disconnectome map indicates, for each voxel in a standard brain template, the probability of structural disconnection. Salvalaggio et al. [6] showed that disconnectome maps can be used to predict behavioral deficits. Their machine learning pipeline was adapted from previous work on structural lesions [1, 8] and was not optimized for disconnectome data. In short, Principal Component Analysis (PCA) was used for dimensionality reduction and the components cumulatively explaining 95% of the variance were retained as features for prediction. The latter was based on cross-validated ridge regression using a behavioral score as outcome variable.

Unsupervised dimensionality reduction is a necessary step for neuroimaging data, which typically have a much greater number of features than observations [9]. A variety of techniques can be used to extract a limited number of features that can compactly describe the data distribution. Prediction from a compact set of brain-related features can then be carried out using regularized regression methods such as ridge regression [4] or Least Absolute Shrinkage and Selection Operator (LASSO) [10]. Regularized regression includes a penalty term that pushes the estimated coefficients of irrelevant features toward zero, limiting the risk of multicollinearity and overfitting [11, 12]. Moreover, regularized methods often also improve model interpretability [13, 14], making them particularly suitable for the analysis of neuroimaging data [15].

In the present study, we systematically investigated the effect of both feature extraction techniques and regularized regression on predictive accuracy, in order to identify the most effective machine learning pipeline for disconnectome data. Indeed, different methods can show considerable variability in performance depending on the type of neuroimaging data and task considered [9, 15, 16]. We recently investigated this issue in the context of resting-state functional connectivity data [10]. Here we extended our approach to disconnectome maps of stroke patients, which were used to predict behavioral scores in the language and motor domains. Disconnectome maps were first processed using different dimensionality reduction methods: Principal Component Analysis, Independent Component Analysis, Non-Negative Matrix Factorization and Dictionary Learning. The extracted features were then used as predictors for cross-validated regularized regression. We assessed the predictive performance while systematically varying the number of extracted features for each dimensionality reduction method as well as a function of the type of regularization method used for supervised learning. Finally, we examined the quality of the brain maps that display disconnectome voxels that are most predictive of behavioral performance in each domain.

In summary, structural disconnectomes [7] represent a relatively new type of neuroimaging data that has not been systematically approached with machine learning techniques. With respect to the state-of-the-art [6], the main contributions of the present work

include: i) the consideration of a broad range of dimensionalty reduction techniques; ii) the evaluation of different types of regularized regression; iii) the inclusion of feature selection to reduce the number of predictors; iv) the assessment of predictive accuracy across multiple measures that also consider model complexity.

2 Materials and Methods

2.1 Participants and Data Acquisition

Behavioral and MRI data were obtained from a previously published study [1], in which 132 symptomatic stroke patients underwent MRI scanning and behavioral testing 1–2 weeks after the stroke occurred. The data for each patient consisted of a 3D image of the lesion, reconstructed from the original MRI image (see [1] for details) and registered to a common coordinate space provided by the Montreal Neurological Institute (MNI space, 2 mm isovoxel resolution) using affine and diffeomorphic deformations. Structural disconnections for 131 patients were computed in [6] from the lesion image with the BCB-toolkit [7], using 176 healthy controls from the "Human Connectome Project" 7T diffusion-weighted imaging dataset as a reference to track fibers passing through each lesioned voxel. In the resulting disconnectome map, the value in each voxel indicates the probability of disconnection from 0 to 1 (see Fig. 1). Note that each map is a sparse high-dimensional 3D matrix, with size $91 \times 109 \times 91$.

Behavioral assessment spanned several cognitive domains [1]. In the present work we focus on language and motor scores, which are available for $n = 116$ and $n = 108$ patients, respectively. For each domain we used an overall "factor score" [1, 6], which

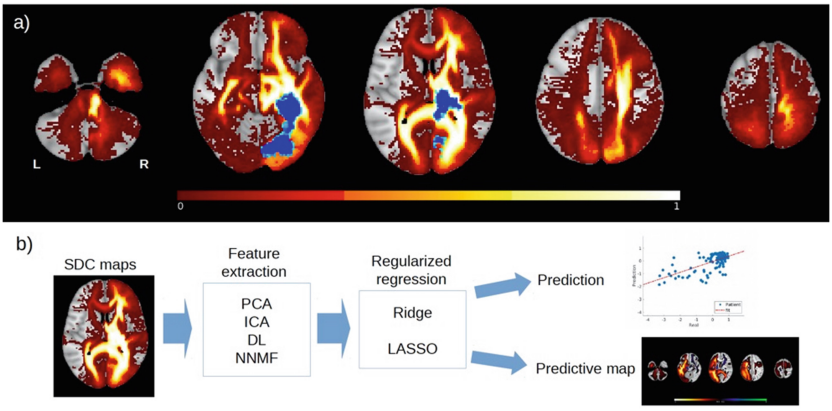

Fig. 1. a) The 3D disconnectome map of a sample stroke patient is displayed here using 5 axial slices. A localized right hemisphere lesion involving the thalamus and the lateral occipital cortex (overlayed on the map in blue color) produces more widespread white matter disconnections (with probability indexed by the red-yellow scale) that include posterior thalamic radiation, superior corona radiata, and extend to the left hemisphere through the splenial part of the corpus callosum. b) Machine learning pipelines assessed in our study. At each processing stage, the performance of different methods was systematically compared to establish the optimal combination of feature extraction and regularized regression techniques. (Color figure online)

captures the shared variance of several sub-tests. For example, the language factor score is the first principal component accounting for 77.3% of the variance across a variety of language tasks. The motor factor score expresses contralesional motor performance (e.g., right limb for left hemisphere damage and vice versa). Each factor score was normalized to represent impaired performance with negative values.

2.2 Unsupervised Feature Extraction

Unsupervised feature extraction was performed using the entire dataset ($n = 131$ and $p = 902{,}629$), to take advantage of all patients' data regardless of the availability of specific neuropsychological scores. All feature extraction methods used in the present work are linear, which means that they aim to find a weight matrix W that can transform the original $n \times p$ data matrix X into a new set of k features, with $k < p$ and usually $k < n$, such that:

$$F = XW \tag{1}$$

where F is the feature space. Since choosing the value of k is nontrivial, we systematically varied k from 5 to $n-1$, with step size = 5, where n is the number of patients entered in the regularized regression. A cross-validation procedure was then used to select the optimal value of k (see Sect. 2.4 below). To compare the compression ability of the different feature extraction methods, the reconstruction error was calculated as the mean squared error (MSE) between the original disconnection maps X and the reconstructed maps X_R, for each value of k. The original maps can be reconstructed by simply backprojecting the feature set into the original input space using the transposed weight matrix:

$$X_R = FW^T \tag{2}$$

Principal Component Analysis (PCA). PCA Linearly Transforms the Input Data into a Smaller Set of Uncorrelated Features Called Principal Components, Sorted by the Explained Data Variance [17]. The Input Data is First Centered, so that It Has Zero-Mean. The Eigenvalues and Eigenvectors of the $p \times p$ Covariance Matrix $X^T X$ Are then Computed Using Matrix Factorization via Singular Value Decomposition:

$$X = UDW^T \tag{3}$$

where U is an $n \times n$ matrix containing the eigenvectors of XX^T, D is an $n \times p$ matrix containing the square root of the eigenvalues on the diagonal, and W is a $p \times p$ matrix containing the eigenvectors of $X^T X$. When $p > n$ there are only $n - 1$ non-zero eigenvalues, so only the first $n - 1$ columns of D and W are kept. Eigenvectors are sorted in descending order of explained variance. Hence, W contains $n - 1$ principal components, expressed as a set of p weights that can map the original variables in a new (compressed) feature space. Since PCA is a deterministic method, it was performed only once, and the first k features were then iteratively selected. The other feature extraction methods are probabilistic in nature, so the procedure was repeated for each value of k.

Independent Component Analysis (ICA). ICA Assumes that a p-dimensional Signal Vector $X_{i,*}^T$ is Generated by a Linear Combination of k Features (with $k <= p$) that Are Assumed to Be Independent and Non-gaussian [18], Leading to:

$$X_{i,*}^T = AF_{i,*}^T \tag{4}$$

where A is a $p \times k$ unmixing matrix, which maps the signal in the sources, and F is the feature vector. Hence, the features can be obtained by:

$$F_{i,*}^T = WX_{i,*}^T \tag{5}$$

where W is the inverse of the unmixing matrix A. The input data is first centered, and then further pre-processed through whitening so that a new vector with uncorrelated components and unit variance is obtained. The *FastICA* function of the scikit-learn library was used, and PCA was used for data whitening [18].

Non-negative Matrix Factorization (NNMF). NNMF is a Form of Matrix Factorization into Non-negative Factors W and H [19], Such that the Linear Combination of Each Column of W Weighted by the Columns of H Can Approximate the Original Data X:

$$X \approx WH \tag{6}$$

NNMF aims to minimize the following loss function:

$$\begin{aligned} &\|A - WH\|_F^2 \\ &\textit{subject to } W, H \geq 0 \end{aligned} \tag{7}$$

The *nnmf* MATLAB function with the "multiplicative update algorithm" was used.

Dictionary Learning (DL). The DL Algorithm, Sometimes Known as *Sparse Coding*, Jointly Solves for a $p \times k$ Dictionary W and the New Set of Features F that Best Represent the Data. To Obtain Only Few Non-zero Entrances an Additional L_1 Penalty Term is Included in the Cost Function:

$$\begin{aligned} (W, F) = \min_{(W,F)} \frac{1}{2} \left\| X - FW^T \right\|_2^2 + \lambda \|F_1\| \\ \textit{subject to } \left\| W_j \right\|_2 \leq 1, \forall j = 1, \ldots, k \end{aligned} \tag{8}$$

where λ is the L_1 penalty coefficient, controlling for the sparsity of the compressed representation [20]. The *DictionaryLearning* function of the scikit-learn library was used.

2.3 Regularized Regression

The features extracted by each unsupervised learning method were then used as regressors for the prediction of the language and motor scores. Note that only the subjects with available target scores were kept in this phase. The regressors were normalized and then

entered either a ridge regression or a LASSO regression [10, 21]. In both cases the loss function can be defined as:

$$\min_{\beta}\left(\frac{1}{2n}\sum_{i=1}^{n}\left(y_i - x_i^T\beta\right)^2 + \lambda R\right) \tag{9}$$

where n is the number of observations, y_i is the prediction target, x_i is the data observation, β represents the p regression coefficients, and R is the regularization term, weighted by the non-negative coefficient λ. In the case of ridge regression, the regularizer is defined as:

$$R(\beta) = \sum_{j=1}^{p}\beta_j^2 \tag{10}$$

while in the case of LASSO regression the regularizer is defined as:

$$R(\beta) = \sum_{j=1}^{p}\left|\beta_j\right| \tag{11}$$

The main difference is that LASSO forces the estimates of non-predictive coefficients to have exactly-zero values, whereas the ridge regularization shrinks those coefficients towards near-zero values [21]. The optimal λ was chosen among 100 values in the range $[10^{-5}, 10^{5}]$ with logarithmic step.

2.4 Cross-Validation Setup and Model Comparison

To find the optimal values for the hyper-parameters λ and k while controlling for overfitting, we employed a search procedure over a range of possible values using cross-validation (CV). The complete dataset was thus split into a *training* set and a *test* set: the training set was used for tuning the hyper-parameters, and the resulting model was then evaluated on the left-out test set. We adopted a Leave-One-Out (LOO) cross validation setup, where just one sample was circularly included in the test set. As a control analysis, we also explored hyper-parameter tuning using nested CV [22]. However, since this method led to negligible differences compared to the standard CV, for computational convenience we did not include it in the final analyses.

To compare the models generated by the different feature extraction methods, both the R^2 and the Bayesian Information Criterion (BIC) [23] were calculated (note that only the non-zero coefficients were used for BIC calculation). Finally, for each method, the optimal regression coefficients were backprojected in the original space, by means of a linear transformation through the features' weights, and restored in the 3D volume of the template brain [6]. This provides a brain map that displays the predictive voxels for a given behavioral domain. The machine learning pipeline is depicted in Fig. 1b.

3 Results

The feature extraction methods were first assessed based on their reconstruction error. For all methods, the reconstruction error decreased when increasing the number of features (Fig. 2a), and NNMF showed generally higher reconstruction error. Systematic

variation of the number of features (k) used as input for regression revealed a pattern of predictive accuracy (in terms of R^2) that was strongly influenced by the feature extraction method (Fig. 2b). PCA- and NNMF-based models were largely insensitive to k, whereas performance of the ICA- and DL-based models deteriorated for large k values. For ICA the performance loss was almost linear and markedly steep when the number of independent components exceeded the optimal k (here 10), suggesting that over-decomposition introduces noise and makes the components less informative. In contrast, the strict order of components' extraction in PCA implies that increasing k has no effect on previously extracted features. Nonetheless, at the optimal value of k, all models showed very good predictive performance with R^2 around 0.40.

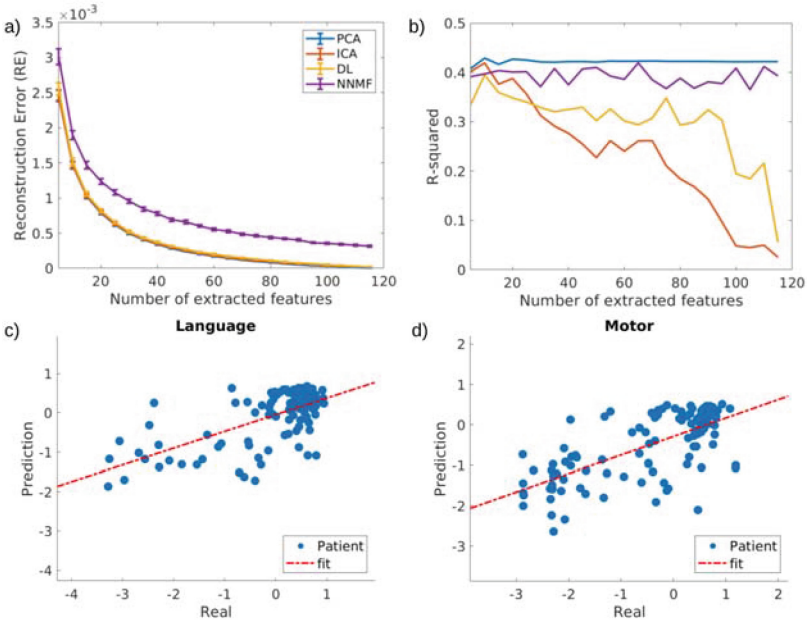

Fig. 2. a) Reconstruction error as a function of the number of extracted features, separately reported for each feature extraction method. b) R^2 values of the models in predicting language scores as a function of the number of extracted features. c) PCA + ridge model predictions with LOO CV for the language domain and d) NNMF + LASSO model predictions for the motor domain.

Performance of the selected model (i.e., optimized for k) for each feature extraction method and behavioral score (language vs. motor) is reported in Table 1 (ridge) and Table 2 (LASSO). The tables also report the optimal hyper-parameters values, as well as the number of non-zero weights in the LASSO regression model after training. Interestingly, different feature extraction methods led to slightly different optimal hyper-parameters. The ICA- and DL-based models were chosen with fewer features than the PCA- and NNMF-based models. A marked difference can be observed between NNMF and the

other methods, with k up to six times larger. Regarding the regularization coefficient, the variability was more substantial for ridge regression than for LASSO.

PCA and NNMF yielded the best predictive accuracy across types of regularization and behavioral domains in terms of R^2. NNMF combined with LASSO regression reached the best performance in the prediction of both motor and language scores, although by a tight margin with respect to PCA + ridge (see model predictions in Fig. 2c and d). When considering the BIC, on the other hand, the PCA + ridge pipeline was favored by the lower model complexity (smaller number of weights). We also report as baseline the performance of models with non-optimized (i.e., fixed) number of features (Tables 1 and 2). For PCA, we used the principal components cumulatively accounting for 95% of the variance ($k = 30$), as used in previous studies [6]. The level of reconstruction error yielded with these components was matched across feature extraction methods to select corresponding k values as baseline models ($k = 30$ for ICA, $k = 30$ for DL, $k = 65$ for NNMF). Importantly, the optimized models were in all cases superior to the baseline (fixed k) models both in terms of prediction performance and reduced models' complexity, as also shown by the large gap in terms of BIC. LASSO regression further reduced the number of predictors by setting some weights to zero. The latter seems to be particularly important for the NNMF-based models because the selected k was larger in comparison to the other methods (accordingly, NNMF + ridge was poor in terms of BIC).

Table 1. Performance of ridge regression in the prediction of language and motor scores. The selected λ and k values are also reported.

Method		Ridge (fixed k)				Ridge			
		R^2	BIC	λ	k	R^2	BIC	λ	K
Lang (n = 116)	PCA	0.42	413.4	475.1	30	0.43	316.8	298.4	10
	ICA	0.31	433.3	46.4	30	0.42	318.7	14.5	10
	DL	0.33	430.6	756.5	30	0.39	323.5	117.7	10
	NNMF	0.42	580.2	1519.9	65	0.42	580.2	1519.9	65
Motor (n = 108)	PCA	0.43	428.9	298.4	30	0.44	333.4	187.4	10
	ICA	0.32	446.9	29.2	30	0.43	334.8	9.1	10
	DL	0.22	461.7	475.1	30	0.42	312.1	11.5	5
	NNMF	0.36	603.7	954.6	65	0.42	571.2	475.1	60

The back-projected weights for the best two models in each domain are depicted in Fig. 3. The language score was predicted by a broad white matter network encompassing tracts traditionally associated with linguistic processing, including the superior longitudinal/arcuate fasciculus, predominantly in the left hemisphere. For the motor domain, the predictive map shows a more circumscribed, subcortical, and bilateral set of white matter tracts surrounding the basal ganglia (e.g., internal and external capsule) as well as, more dorsally, parts of the corona radiata.

Table 2. Performance of LASSO regression in the prediction of language and motor scores. The selected λ, and k values and the number of non-zero features (NZ) is also reported.

		LASSO (fixed k)				LASSO			
		R^2	BIC	λ	k(NZ)	R^2	BIC	λ	k(NZ)
Lang (n = 116)	PCA	0.41	358.2	0.055	30(18)	0.42	336.9	0.055	20(14)
	ICA	0.30	435.8	0.043	30(30)	0.42	319.3	0.022	10(10)
	DL	0.24	354	0.11	30(11)	0.35	316.7	0.055	10(7)
	NNMF	0.41	358	0.087	65(18)	0.43	317	0.069	25(10)
Motor (n = 108)	PCA	0.38	405	0.055	30(23)	0.43	335	0.022	10(10)
	ICA	0.24	459.6	0.003	30(30)	0.43	335.1	0.022	10(10)
	DL	0.14	397.3	0.11	30(14)	0.42	313.6	0.001	5(5)
	NNMF	0.23	502.5	0.043	65(39)	0.45	451.6	0.055	60(36)

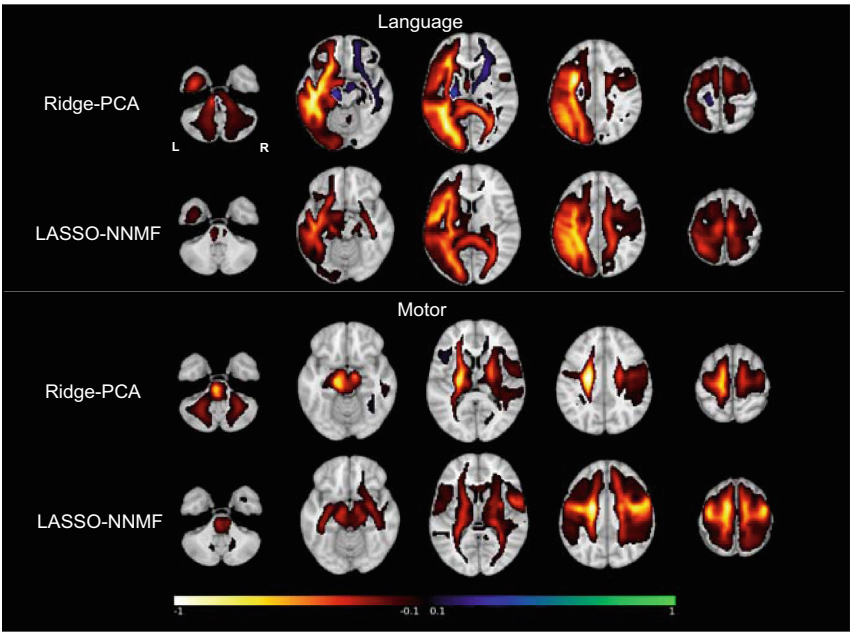

Fig. 3. Back-projected weights for the best two models for prediction of each behavioral score, highlighting the most predictive white matter tracts.

4 Discussion

In this work we systematically compared four unsupervised dimensionality reduction methods in their ability to extract relevant features from probabilistic structural disconnection maps of stroke patients at different levels of compression. We then assessed how these methods influence a regularized regression model trained on the features to predict patients' behavioral performance.

Overall, PCA and NNMF turned out to be the best methods for extracting robust predictors, followed by ICA and DL. Optimizing the number of extracted features (k) to be entered as predictors for regression was crucial for the predictive accuracy of ICA- and DL-based models, but not for PCA and NNMF. Nevertheless, when compared to non-optimized models (fixed k), we observed that the optimization of k improved the accuracy for all methods and greatly reduced model complexity, thereby leading to large gains in terms of BIC. This suggests that good predictive accuracy can be obtained with a limited number of features. A compact representation is desirable as it improves model interpretability and it might favor out-of-sample generalization. Interestingly, while LASSO regression can further reduce model complexity by setting some weights to 0, we found that it was not superior to ridge regression when the optimization of k was in place.

Finally, the type of regularizer interacted with the feature extraction method: PCA optimally combined with ridge regression, whereas NNMF optimally combined with LASSO regression. These two pipelines achieved the best performance, but PCA + ridge regression was more consistently the best approach when also considering the BIC score. However, the differences between these two pipelines (given the optimal hyperparameters) were small. Indeed, the back-projection of the most relevant features for the two best models in each domain were similar and neuroanatomically sound.

Overall, our findings represent an important step towards the definition of the optimal pipeline for disconnectome data. Compared to the previous state-of-the-art [6], the gain in terms of r-squared is marginal (e.g., 2% of variance for language scores) but this is achieved with a much more parismonius model with just one third of the number of predictors (10 vs. 29 in [6]). A potential limitation of the study is due to the relatively small sample size of the patient group, but this simply reflects the lack of large-scale stroke datasets including both neuroimaging and behavioral data. Further efforts should be spent in assembling larger-scale datasets, which would allow to deploy even more powerful predictive models, such as those based on deep learning [24]. Future work should also extend these results to the prediction of a broader range of behavioral scores, to better assess whether some feature extraction methods could be more general than others and to further compare them in terms of interpretability and neuroscientific accuracy of the predictive maps.

Acknowledgments. This work was supported by grants from the Italian Ministry of Health (RF-2013–02359306 to MZ, Ricerca Corrente to IRCCS Ospedale San Camillo) and by MIUR (Dipartimenti di Eccellenza DM 11/05/2017 n. 262 to the Department of General Psychology). We are grateful to Prof. Maurizio Corbetta for providing the stroke dataset, which was collected in a study funded by grants R01 HD061117–05 and R01 NS095741.

References

1. Corbetta, M., et al.: Common behavioral clusters and subcortical anatomy in stroke. Neuron **85**, 927–941 (2015)
2. Rorden, C., Karnath, H.O.: Using human brain lesions to infer function: a relic from a past era in the fMRI age. Nat. Rev. Neurosci. **5**, 813–819 (2004)
3. Price, C.J., Hope, T.M., Seghier, M.L.: Ten problems and solutions when predicting individual outcome from lesion site after stroke. Neuroimage **145**, 200–208 (2017)
4. Siegel, J.S., et al.: Disruptions of network connectivity predict impairment in multiple behavioral domains after stroke. Proc. Natl. Acad. Sci. USA **113**, E4367–E4376 (2016)
5. Thiebaut de Schotten, M., Foulon, C., Nachev, P.: Brain disconnections link structural connectivity with function and behaviour. Nat. Commun. **11**, 5094 (2020)
6. Salvalaggio, A., de Filippo De Grazia, M., Zorzi, M., de Schotten, M.T., Corbetta, M.: Post-stroke deficit prediction from lesion and indirect structural and functional disconnection. Brain **143**, 2173–2188 (2020)
7. Foulon, C., et al.: Advanced lesion symptom mapping analyses and implementation as BCBtoolkit. Gigascience **7**, 1–17 (2018)
8. Chauhan, S., et al.: A comparison of shallow and deep learning methods for predicting cognitive performance of stroke patients from MRI lesion images. Front. Neuroinf. **13**, 53 (2019)
9. Mwangi, B., Tian, T.S., Soares, J.C.: A review of feature reduction techniques in Neuroimaging. Neuroinformatics **12**, 229–244 (2014)
10. Calesella, F., Testolin, A., De Filippo De Grazia, M., Zorzi M.: A comparison of feature extraction methods for prediction of neuropsychological scores from functional connectivi-ty data of stroke patients. Brain Inf. **8**, 8 (2021)
11. Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. J. Mach. Learn. Res. **3**, 1157–1182 (2003)
12. Hua, J., Tembe, W.D., Dougherty, E.R.: Performance of feature-selection methods in the classification of high-dimension data. Pattern Recogn. **42**, 409–424 (2009)
13. Carroll, M.K., Cecchi, G.A., Rish, I., Garg, R., Rao, A.R.: Prediction and interpretation of distributed neural activity with sparse models. Neuroimage **44**, 112–122 (2009)
14. Teipel, S.J., Kurth, J., Krause, B., Grothe, M.J.: 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 (2015)
15. Cui, Z., Gong, G.: The effect of machine learning regression algorithms and sample size on individualized behavioral prediction with functional connectivity features. Neuroimage **178**, 622–637 (2018)
16. Jollans, L., et al.: Quantifying performance of machine learning methods for neuroimaging data. Neuroimage **199**, 351–365 (2019)
17. Jolliffe, I.T.: Principal component analysis. In: Encyclopedia of Statistics in Behavioral Science (2002)
18. Hyvärinen, A., Oja, E.: Independent component analysis: algorithms and applications. Neural Netw. **13**, 411–430 (2000)
19. Lee, D.D., Seung, H.S.: Algorithms for non-negative matrix factorization. In: Advances in Neural Information Processing Systems, pp. 556–562 (2001)
20. Mairal, J., Bach, F., Ponce, J., Sapiro, G.: Online dictionary learning for sparse coding. In: ACM International Conference Proceeding Series, pp. 689–696 (2009)
21. Tibshirani, R.: Regression shrinkage and selection via the lasso. J. Roy. Stat. Soc. Ser. B (Methodol.) **58**, 267–288 (1996)

22. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning. Springer Series in Statistics. Springer, New York (2009). https://doi.org/10.1007/978-0-387-84858-7
23. Schwarz, G.: Estimating the dimension of a model. Ann. Stat. **6**, 461–464 (1978)
24. Vieira, S., Pinaya, W.H., Mechelli, A.: Using deep learning to investigate the neuroimaging correlates of psychiatric and neurological disorders: methods and applications. Neurosci. Biobehav. Rev. **74**, 58–75 (2017)

Author Queries

Chapter 20

Query Refs.	Details Required	Author's response
AQ1	This is to inform you that corresponding author has been identified as per the information available in the Copyright form.	Testolin should be co-corresponding author (as in original manuscript)