A Machine Learning Approach for Predicting Post-stroke Aphasia Recovery: A Pilot Study

Yiwen Gu[1*], Murtadha Bahrani [1*], Anne Billot[2], Sha Lai[1], Emily J. Braun[2], Maria Varkanitsa[2], Julia Bighetto[1], Brenda Rapp[4], Todd B. Parrish[5], David Caplan[6], Cynthia K. Thompson[3], Swathi Kiran[2], and Margrit Betke[1]

[1] Department of Computer Science, [2] Sargent College of Health and Rehabilitation Sciences
Boston University
[3]Department of Communication Sciences and Disorders, Department of Neurology and Mesulam Cognitive Neurology and Alzheimer's Disease Center, Northwestern University, [4]Department of Cognitive Science, Johns Hopkins University, [5] Department of Radiology, Feinberg School of Medicine, Northwestern University, [6]Massachusetts General Hospital, Department of Neurology, Harvard Medical School

ABSTRACT

The potential recovery of post-stroke aphasia is highly variable and the rehabilitation outcomes are difficult to predict. This interdisciplinary collaboration builds on data collected as part of a large set of behavioral and brain variables in patients with post-stroke aphasia, charting the course of recovery associated with therapy across language domains and examining the basis of neuroplasticity. In this pilot study, we created and tested a predictive framework based on a subset of the data collected and developed machine-learning algorithms that take as input a complex set of brain and behavioral features to classify and predict the participants' responsiveness to therapy. We developed Random Forest models that enabled us to rank the importance of these features. We then compared the contributions of different feature sets and discussed their physiological implications. Our preliminary results suggest the potential of our framework, and, thus, this study takes an important first step towards predicting individualized rehabilitation outcomes.

CCS CONCEPTS

• **Computing methodologies**;

KEYWORDS

Stroke, Aphasia, Recovery, Machine Learning

1 INTRODUCTION

Stroke is a leading cause of severe long-term disability. It affects 800,000 people in the United States every year and its incidence is rising with the aging population. One third of stroke survivors lose language abilities due to brain damage, commonly defined as aphasia. Over two million people are living with post-stroke aphasia in the United States and it is considered one of the most debilitating chronic conditions [1]. In everyday life, it results in difficulty in speaking, comprehending, writing and/or reading, which leads to reduced social participation [2]. Although most patients improve with speech and language therapy, the recovery trajectory can be difficult to predict at the individual level. Multiple factors impact responsiveness to treatment and the evolution of aphasia over time, such as stroke severity, degree of initial language impairment and demographic information [3, 4]. In most cases, the prescription of therapy is largely based on the behavioral presentation and does not take into account the neural profile. While people with aphasia are asking for more information on what to expect from rehabilitation [5], clinicians have limited resources to provide a personalized prognosis and adapt rehabilitation to individual needs.

In the last decades, thanks to new techniques in neuroimaging, researchers have started to investigate how brain structure and function, along with demographics and baseline aphasia severity, could help inform rehabilitation and language outcomes in post-stroke aphasia. Lesion size and lesion location have been some of the most investigated prognostic neurological factors in this domain and the majority of studies agree on their essential role in predicting language scores and recovery over time [3, 4]. Until recently, most prediction analyses focused on brain-behavior relationships using simple correlations or regressions between damaged areas of interest and cognitive scores, as well as mass-univariate voxel-based lesion symptom mapping [6–14]. The major limitation of these methods comes from the high dimensionality and collinearity of the neural information due to the vasculature pattern [15, 16] along with small sample sizes. Machine learning algorithms are therefore an appropriate data-led approach to overcome these challenges and assist in prediction of language abilities and recovery [17, 18].

2 RELATED WORK AND CONTRIBUTION

In the last years, studies demonstrated that multivariate data-driven models could predict language profile of individuals with aphasia above chance based on neuroimaging and patient-related information. In neuroimaging studies, multivariate pattern analysis (MVPA) aims at analyzing patterns that span multiple brain regions. Saur et al. [19] showed that a multivariate pattern classification approach could best predict a composite language score six months post-stroke by adding language functional magnetic resonance imaging (fMRI) data to baseline language profile and age. In the meantime, as standard statistical methods demonstrated that lesion size and lesion location were associated with aphasia outcomes and recovery, Price et al. [20] introduced a data-led system to improve predictions

[*] The two authors contributed equally.

PETRA '20, June 30-July 3, 2020, Corfu, Greece
© 2020 Association for Computing Machinery.
ACM ISBN 978-1-4503-7773-7/20/06…$15.00
https://doi.org/10.1145/3389189.3389204

with a large database of structural MRI scans and behavioral data from individuals who suffered a stroke. This group then demonstrated that lesion information predicted speech production skills with the best accuracy (i.e., 0.59) when percentage of damage from 35 anatomically parcellated brain regions, overall lesion volume and time post-stroke were included in a Gaussian Process model regression. Two studies used a similar brain parcellation, either anatomical [21] or functional [22], to predict aphasia subtypes by training multivariate regression or support vector machine models based on proportion of damage. Furthermore, Halai et al. [22] also obtained prediction scores above chance for a variety of language scores. Other studies have used machine learning techniques to evaluate the role of connectivity data in multimodal prediction models. In these studies, diffusion-weighted imaging and/or resting state fMRI data were used to compare the prediction accuracy of models including lesion information with models including structural connectivity weights between regions [23], structural and functional graph theory metrics [24] or binary connectivity disruption measures [25]. These studies demonstrated that both lesion and connectivity data could predict language profiles but yielded inconsistent results on the superiority of a certain type of variable over another.

Another way to investigate the prediction power of lesion data is by training models on raw images to overcome preprocessing challenges. In particular, Roohani et al. [26] trained a convolutional neural network on 2D stitched images in order to classify the degree of language abilities recovered (bad versus good outcomes) and obtained an accuracy of 0.79. In addition to language profiles and aphasia subtypes, studies also presented effective models to predict change in language performance over time between initial and follow-up assessments [19, 27]. However, none of these studies included information about evidence-based rehabilitation services received by the patients between the different time points of data collection. Yet, being able to predict how a patient will respond to treatment is essential in a rehabilitation setting. This project aims to develop machine-learning algorithms that take as input a complex set of brain and behavioral markers in order to classify and predict the responsiveness to treatment. Thus, this project takes an important first step towards predicting individualized rehabilitation outcomes.

As part of a collaborative work between three universities involved in the Center for Neurobiology of Langauge Recovery, a large set of behavioral and neuroimaging data from individuals with aphasia who received evidence-based language treatment over a three-month period were collected. In this pilot, we, at Boston University, developed novel machine learning algorithms (Fig 1) to classify and predict responders versus nonresponders to aphasia treatment using multiple factors.

Besides correctly classifying and predicting the responsiveness to treatment, we are also interested in identifying factors that are important in the prediction. Feature ranking methods are used to identify relevant features in the prediction, giving insight into the data. Feature extraction methods, on the other hand, are used to create new sets of features wherein the new features are combinations of the original [28]. Principal Component Analysis is a common example of a feature extraction method. Using such methods undermines interpretability, and hence was not utilized in this pilot study. Instead, our investigation of machine learning algorithms to classify response to aphasia treatment resulted in our selection of the Random Forest model, which provides interpretability.

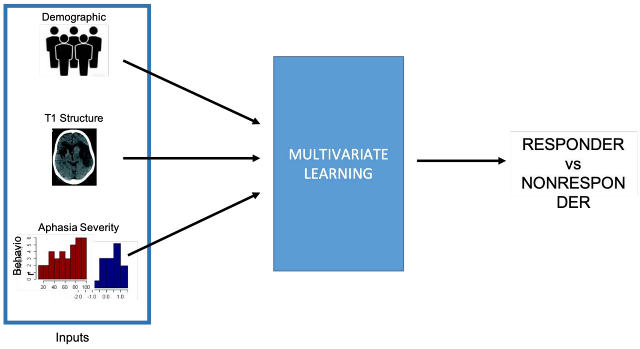

Figure 1: Framework of Multivariate Predictors

3 METHODS

3.1 Data Collection

Participants. 80 individuals were recruited from three research sites (Boston University - BU, Johns Hopkins University - JHU, and Northwestern University - NU) between 2015 and 2018, to be part of a large-scale NIDCD study within the Center for the Neurobiology of Language Recovery [1]. Inclusion criteria were a single left-hemisphere ischemic stroke, diagnosis of chronic aphasia (i.e., at least six months post stroke-onset), native language English, at least a high school education, normal or corrected-to-normal vision and hearing, and no history of neurological disorder other than a stroke. Patients with a history of multiple infarcts or of drug or alcohol abuse were excluded from this study. Exclusion also applied for patients with contraindication for MRI and motor speech disorders such as apraxia or dysarthria. Among the 80 participants, data from 65 aphasic patients were used in this study. Four were excluded due to poor structural imaging data quality and eleven were patients who did not receive treatments. Study-specific questionnaires and medical records were used to obtain demographic and neurological history information. Participants provided informed consent according to the Declaration of Helsinki, and received additional information if needed to understand the study's protocol before consenting. The study was approved by the Institutional Review Boards of all three research sites.

Behavioral data and Treatment protocol. In each site, a comprehensive battery of language tests were administered to all participants on entry and at specified test points during and following treatment. The tests included assessments of various components of language, including naming, spelling, sentence comprehension and sentence production, and the Western Aphasia Battery-Revised (WAB) [29], as the recognized primary outcome measure for aphasia rehabilitation [30]. The Aphasia Quotient (AQ) was collected from this standardized assessment as a measure of aphasia severity. After neuroimaging and behavioral data were collected, 65 patients

[1] http://cnlr.northwestern.edu/

received two-hour sessions of language therapy twice a week for 12 weeks. The remaining patients served as controls but are not included in this analysis. Therapy targeted different language components within each site: word-retrieval at BU, sentence processing at NU and spelling at JHU. For details about each therapy see [31], [32] and [33], respectively. Performance was monitored every week with a list of probes specific to each language domain trained. Treatment-related gains were determined by calculating the percent change in accuracy (i.e., average post-treatment accuracy score minus average pre-treatment accuracy score in percentages) [34]. Individuals with aphasia were classified as having a favorable response to treatment (i.e, responders) or a less satisfactory response to treatment (i.e., nonresponders) based on a cut-off at 25% for percent change in accuracy.

MRI data acquisition. Image data was gathered from four different 3 Tesla scanning systems. A Siemens TIM Trio with 32-channel head coil and a Siemens Prisma with a 64-channel head/neck coil were used at Northwestern University; a Philips Intera with a 32-channel head coil was used at Johns Hopkins University; and a Siemens TIM Trio was used at the Athinoula A. Martinos Center at Massachusetts General Hospital/Boston University. Imaging parameters were verified by the neuroimaging team to be homogeneous across the three universities. T1-weighted 3D sagittal volumes were obtained in a high resolution with an MPRAGE sequence (parameters: TI/TE/TR = 900/2.91/2300 ms, FOV=256x256 mm, voxel resolution = 1x1x1mm3, 176 sagittal slices, phase encoding direction = A/P).

Lesion mapping and lesion size (LS). Lesions were traced manually slice-by-slice by trained research assistants on T1-weighted images using MRIcron [2]. The three orthogonal views were visualized simultaneously during the tracing procedure to improve accurate identification of the damage. Spatial normalization of the T1-weighted images and lesion maps were performed with AFNI [35] and the volume of each map was calculated with in-house MATLAB scripts.

Percent spared in gray matter regions (SP). To determine the integrity of cortical regions, we first delineated spared tissue in each region of interest (ROI) of the Harvard-Oxford cortical structural atlas [36] by subtracting lesion maps from cortical ROIs. We then calculated the remaining ROIs volume and converted it into a percentage of spared tissue for each region of the atlas. Only left hemisphere regions were included and we used the MarsBaR tool in SPM [37] to process the images.

Proportion of damage in white matter regions (PD). A third type of lesion information was extracted from the overlap between lesion maps and a probabilistic atlas of white matter pathways obtained from 47 healthy controls [38] using the Tractotron tool from the BCBToolkit [39]. We obtained a proportion of damage score for each white matter tract in each patient.

3.2 Model Construction

The problem at hand is a binary classification problem where the input is a complex set of brain and behavioral markers and the output labels are responder and nonresponder. Our dataset has a high dimensionality (d = 130 features) but a limited number of data points (n = 65 samples). Two criteria were considered in choosing a classifier: performance and interpretability. The performance was assessed by the generalization error of the model. On the other hand, interpretability was assessed by the model capability to produce a measure of feature ranking. In the next sections we discuss the measures and methods used to achieve these two goals. We used the scikit-learn 0.22.1 [40] library for our model exploration.

Algorithm selection. In this study, we considered three algorithms as candidates: a random forest (RF), a support-vector machine with RBF kernel (SVM-RBF), and a gradient boosting machine (GBM). Our choice was informed by the findings of Wainer [41], who compared these three classifiers empirically against 14 other classification algorithms and found them to yield the best performance. We applied these three algorithms on the whole dataset (d = 130) and reported in Table 1 the average model performance using 5-fold cross-validation with the optimized hyperparameters settings. According to this preliminary result, no model is significantly better than the others. This observation is consistent with previous work [41]. In terms of interpretability, SVM-RBF is limited as features are transformed to higher dimensionality. As for RF and GBM, our experiment showed that RF has sharper details in the feature ranking than GBM. For these aforementioned considerations, we selected the random forest algorithm to further investigate the data.

Table 1: Comparing Performance of ML Algorithms

Algorithm	Prediction Accuracy	Standard Deviation
RF	0.738	± 0.09
SVM-RBF	0.723	± 0.10
GBM	0.723	± 0.08

Random Forest. The idea of ensemble methods is to use the results of multiple predictive models, called estimators or weak learners, to compute the final output. Random Forest is one of the most widely-used supervised learning algorithms that uses ensemble methods for prediction. In its original version, the prediction is computed through a voting mechanism, in which each estimator votes for the most popular class (for classification) or the average of all estimators is used (for regression) [42]. Previous work [41, 43] found, using the same benchmarked dataset, that Random Forest is most likely the best classification algorithm. Random Forest offers many advantages. First, it is easy to train due to parallelization in constructing trees and computing predictions. Second, it provides a built-in feature ranking measurement, called feature importance.

Moreover, RF is capable of handling high dimensional and noisy data. Furthermore, the node-splitting mechanism, as implemented in the scikit-learn library, provides feature ranking once the trees are built. We used *entropy* as the selection criterion for split points that maximize the information gain in the tree nodes. The feature

[2] http://www.mccauslandcenter.sc.edu/mricro/mricron/

Table 2: Descriptions for Feature Sets

Feature Sets	Descriptions	Dimension
Dm	Age	1
	Education level	1
	Time post-stroke onset (in months)	1
Struc	Lesion size from T1-weighted MRI	1
	Percent spared (sp) per grey matter ROI	57
	Proportion of damage (pd) in white matter regions	68
Behav	Behavioral score using WAB-AQ, indicator of the aphasia severity	1

Table 3: Hyperparameters for the Models with Largest Feature Sets

Model	M2	M4	M5
n_estimators	1,000	700	300
max_depth	20	110	60
min_samples_split	2	2	2
min_samples_leaf	2	4	2
max_features	None* (126)	None (129)	None (130)

* Numbers in parentheses were computed without a limit on the number of features

ranking ability satisfies one of our goals for this pilot study, which is to identify the most important features that are related to aphasia rehabilitation.

Model evaluation and selection. For the purpose of evaluating the Random Forest models we constructed, prediction accuracy was used to estimate model performance. To accommodate our small sample size and avoid bias in the estimated mean values of accuracy, we adopted a stratified k-fold cross-validation method, where we chose k to be 5. In each of the 5 rounds, the model under evaluation is trained over 80% of the data and tested on the remaining 20%. The overall model performance is the average of the 5-fold performance, each evaluated by the prediction accuracy on their respective test set. Using this evaluation method, we proceeded to tune and optimize the hyperparameter so as to enhance the RF model performance. However, searching for the optimal values of hyperparameters, also known as model selection, is a nontrivial process. We adopted a strategy that first randomly searches in a large range for potentials and then runs a grid search for the optimal. In both the randomized and the grid search, we applied stratified 5-fold cross-validation as mentioned in the model evaluation. The hyperparameters we optimized are:

- the number of trees in the random forest (n_estimators),
- the maximum depth of the random forest tree (max_depth),
- the minimum number of samples required to split an internal node (min_samples_split),
- the minimum number of samples required to be at a leaf node (min_samples_leaf),
- the number of features to consider when looking for the best split (max_features).

4 EXPERIMENTS AND RESULTS

We separated the 130 features according to their source into three feature sets: demographics (Dm), brain structure (Struc), and behavior (Behav). Table 2 gives a brief description of these feature sets. We applied five different combinations of these feature sets, yielding five different RF models, M1–M5. The results of the respective searches for the best hyperparameters of each of the models M2, M4, and M5 (i.e., the models with more than 100 features) are given in Table 3. The results of training and testing these five models with 5-fold cross-validation are summarized in Table 4.

We first tested each feature set separately. Our first model, M1, was trained and tested with the demographics feature set. M1 achieved an average accuracy of 0.60 (f1 score = 0.73). Among the three features in this set, our model suggests that age weighs the most (slightly over 40%) in predicting responsiveness to treatment, and time post stroke-onset (slightly below 40%) follows it closely. The patient's education level is the least important (about 20%).

Model 2 is trained and tested with the brain structure feature set, corresponding to lesion information only. M2 achieved an average accuracy of 0.70 (f1 score = 0.78). Among the three subsets in this feature set, which encompasses 126 neuroimaging variables, M2 suggests that the percentage of spared tissue in the posterior division of cingulate gyrus, in the left lateral ventricle, and in the precuneus cortex were the three most important predictors of treatment-related change (Fig 2(a)).

Our third model (M3) was trained and tested on the baseline aphasia severity score (WAB-AQ) as a predictor of responsiveness to treatment and provided an accuracy of 0.71 (f1 score = 0.79). Among the three single feature set models M1–M3, the brain structure feature set alone made the prediction whose accuracy was close enough to the prediction made by the baseline behavior score, with a higher standard deviation though.

In our next set of experiments, we concatenated feature sets to investigate the predictive performance of a more comprehensive data set, as well as the role of lesion and non-lesion information in the prediction. Model M4, which included demographics and brain structure features as inputs, yielded an accuracy of 0.69 (f1 score = 0.77). A closer look at Fig 2(b) shows that demographic information, time post stroke-onset and age, ranked within the top ten over a total of 129 features. The top three ranked features remained the same as in the M2 (Fig 2(a)).

Table 4: Random Forest Classifiers for Responders and Nonresponders

#	Feature Set	Accuracy	F1 score	Sensitivity	Selectivity
M1	Dm	0.602	0.730	0.56	0.63
M2	Struc	0.701	0.784	0.65	0.72
M3	Behav	0.707	0.785	0.58	0.74
M4	Dm, Struc	0.686	0.774	0.61	0.71
M5	Dm, Struc, Behav	**0.737**	**0.800**	**0.69**	**0.77**

Finally, when all feature sets were combined, the mean accuracy was 0.74 (f1 score = 0.80). Not surprisingly, WAB-AQ was the top predictor of behavioral treatment outcome. Further, spared tissue in the left lateral ventricle, posterior division of cingulate gyrus, left insular cortex, left precuneus cortex, and left Heschl's gyrus were the top structural factors that predicted language improvement. In this last model, age was then the 8th most important feature.

5 DISCUSSION AND FUTURE WORK

Our study proposed a preliminary method to predict treatment-related language recovery after stroke and demonstrated competitive accuracy (0.74) while indicating the most relevant predictors.

Feature importance. In this study, we replicated previous findings and confirmed the importance of age and baseline aphasia severity (WAB-AQ) in predicting language recovery. Previous work showed that older patients tend to benefit less from therapy than younger individuals, however, this may be confounded with a higher prevalence of fluent aphasia in older individuals [3, 4].

After training a model M2 on neuroimaging data only, the top features that classified responders and nonresponders correctly include the degree of spared tissue in the posterior division of the cingulate gyrus and in the precuneus cortex (Fig. 2a). These two associative regions are involved in a large range of cognitive tasks and have been identified as a pivotal node of the default-mode network [44], one of the most robust functional network of the brain, associated with self-processing, planning, and reflexive thoughts [45]. Furthermore, a study using functional MRI found that the precuneus was functionally connected to regions involved in language processing such as the left triangular part of inferior frontal gyrus (Broca's area), superior and middle temporal gyri, Heschl's gyrus, and the insula [46]. Other studies found that precuneus and posterior cingulate cortices are related to cognitive control [47] and attentional shifting [48]. Therefore, integrity of this region as part of a domain-general network may be important in benefiting the resources provided in language rehabilitation. In fact, self-regulation as well as good attentional and memory skills are important determinants of language therapy success. Two previous studies, including one from our group, found that pre-treatment non-linguistic cognitive deficits were predictive of poorer therapy outcomes [49, 50]. Furthermore, domain-general regions, such as the cingulate gyrus and precuneus cortex, are not typically damaged after a stroke, and the intactness of these regions may aid in the overall responsiveness to language treatment.

When adding demographics to the neuroimaging-only model M4, the top three remained the same, followed by time post stroke-onset and age (Fig. 2b). Ultimately, when all the variables in the same model (demographics, baseline aphasia severity, and lesion information) were included, the prediction accuracy reached the best performance. The severity of language impairments before treatment was the most important factor to predict treatment outcomes in this final model M5 (Fig. 2c). Furthermore, the top three brain structure features remain among the top five in this final model. In addition, other top features include the proportion of spared tissue in left cortical regions known to be involved in language processing, such as the insula [51], left Heschl's gyrus [52], and the posterior division of the supramarginal gyrus [53].

Model sensitivity and selectivity. Our results show that all of the 5 models have a higher selectivity than sensitivity. Recall that:

$$\text{Sensitivity} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}},$$

$$\text{Selectivity} = \frac{\text{True Negative}}{\text{True Negative} + \text{False Positive}}.$$

The fact that the models favor selectivity indicates that our classifiers have a higher probability to identify patients who are likely to be non-responders to language therapy.

Considering the limited number of data points (65) in this pilot study, our results may change in the future with additional data. In addition, our ongoing analyses will examine white matter tract integrity and resting-state fMRI connectivity between regions as additional predictors of treatment outcome.

Acknowledgements

We would like to thank all the individuals with aphasia who participated in this study and their families. We also acknowledge the work of our collaborators through the Center for the Neurobiology of Language Recovery. This work was supported by the NIH/NIDCD Clinical Research Center Grant, P50DC012283, the Hariri Institute Artificial Intelligence Research (AIR) Initiative and the Institute for Health System Innovation & Policy (IHSIP) at Boston University.

REFERENCES

[1] Jonathan MC Lam and Walter P Wodchis. The relationship of 60 disease diagnoses and 15 conditions to preference-based health-related quality of life in ontario hospital-based long-term care residents. *Medical care*, pages 380–387, 2010.

[2] Ruth JP Dalemans, Luc P De Witte, Anna JHM Beurskens, Wim JA Van Den Heuvel, and Derick T Wade. An investigation into the social participation of stroke survivors with aphasia. *Disability and Rehabilitation*, 32(20):1678–1685, 2010.

[3] Emily Plowman, Brecken Hentz, and Charles Ellis. Post-stroke aphasia prognosis: A review of patient-related and stroke-related factors. *Journal of evaluation in clinical practice*, 18(3):689–694, 2012.

[4] MM Watila and SA Balarabe. Factors predicting post-stroke aphasia recovery. *Journal of the neurological sciences*, 352(1-2):12–18, 2015.

[5] Linda Worrall, Sue Sherratt, Penny Rogers, Tami Howe, Deborah Hersh, Alison Ferguson, and Bronwyn Davidson. What people with aphasia want: Their goals according to the icf. *Aphasiology*, 25(3):309–322, 2011.

[6] Nina F Dronkers, David P Wilkins, Robert D Van Valin Jr, Brenda B Redfern, and Jeri J Jaeger. Lesion analysis of the brain areas involved in language comprehension. *Cognition*, 92(1-2):145–177, 2004.

[7] Ronald M Lazar, Allison E Speizer, Joanne R Festa, John W Krakauer, and Randolph S Marshall. Variability in language recovery after first-time stroke. *Journal of Neurology, Neurosurgery & Psychiatry*, 79(5):530–534, 2008.

[8] Julius Fridriksson, Olafur Kjartansson, Paul S Morgan, Haukur Hjaltason, Sigridur Magnusdottir, Leonardo Bonilha, and Christopher Rorden. Impaired speech repetition and left parietal lobe damage. *Journal of Neuroscience*, 30(33):11057–11061, 2010.

[9] Sarah Marchina, Lin L Zhu, Andrea Norton, Lauryn Zipse, Catherine Y Wan, and Gottfried Schlaug. Impairment of speech production predicted by lesion load of the left arcuate fasciculus. *Stroke*, 42(8):2251–2256, 2011.

[10] Julius Fridriksson, Dazhou Guo, Paul Fillmore, Audrey Holland, and Chris Rorden. Damage to the anterior arcuate fasciculus predicts non-fluent speech production in aphasia. *Brain*, 136(11):3451–3460, 2013.

[11] SH Kim and SH Jang. Prediction of aphasia outcome using diffusion tensor tractography for arcuate fasciculus in stroke. *American Journal of Neuroradiology*, 34(4):785–790, 2013.

[12] Jasmine Wang, Sarah Marchina, Andrea C Norton, Catherine Y Wan, and Gottfried Schlaug. Predicting speech fluency and naming abilities in aphasic patients. *Frontiers in human neuroscience*, 7:831, 2013.

[13] Daniel Mirman, Qi Chen, Yongsheng Zhang, Ze Wang, Olufunsho K Faseyitan, H Branch Coslett, and Myrna F Schwartz. Neural organization of spoken language revealed by lesion–symptom mapping. *Nature communications*, 6(1):1–9, 2015.

[14] Argye E Hillis, Yuan Ye Beh, Rajani Sebastian, Bonnie Breining, Donna C Tippett, Amy Wright, Sadhvi Saxena, Chris Rorden, Leonardo Bonilha, Alexandra Basilakos, et al. Predicting recovery in acute poststroke aphasia. *Annals of neurology*, 83(3):612–622, 2018.

[15] Yee-Haur Mah, Masud Husain, Geraint Rees, and Parashkev Nachev. Human brain lesion-deficit inference remapped. *Brain*, 137(9):2522–2531, 2014.

[16] Yongsheng Zhang, Daniel Y Kimberg, H Branch Coslett, Myrna F Schwartz, and Ze Wang. Multivariate lesion-symptom mapping using support vector regression. *Human brain mapping*, 35(12):5861–5876, 2014.

[17] Cathy J Price, Thomas M Hope, and Mohamed L Seghier. Ten problems and solutions when predicting individual outcome from lesion site after stroke. *Neuroimage*, 145:200–208, 2017.

[18] Stephen M Wilson and William D Hula. Multivariate approaches to understanding aphasia and its neural substrates. *Current neurology and neuroscience reports*, 19(8):53, 2019.

[19] Dorothee Saur, Olaf Ronneberger, Dorothee Kümmerer, Irina Mader, Cornelius Weiller, and Stefan Klöppel. Early functional magnetic resonance imaging activations predict language outcome after stroke. *Brain*, 133(4):1252–1264, 2010.

[20] Cathy J Price, Mohamed L Seghier, and Alex P Leff. Predicting language outcome and recovery after stroke: the ploras system. *Nature Reviews Neurology*, 6(4):202, 2010.

[21] Grigori Yourganov, Kimberly G Smith, Julius Fridriksson, and Chris Rorden. Predicting aphasia type from brain damage measured with structural mri. *Cortex*, 73:203–215, 2015.

[22] Ajay D Halai, Anna M Woollams, and Matthew A Lambon Ralph. Predicting the pattern and severity of chronic post-stroke language deficits from functionally-partitioned structural lesions. *NeuroImage: Clinical*, 19:1–13, 2018.

[23] Grigori Yourganov, Julius Fridriksson, Chris Rorden, Ezequiel Gleichgerrcht, and Leonardo Bonilha. Multivariate connectome-based symptom mapping in post-stroke patients: networks supporting language and speech. *Journal of Neuroscience*, 36(25):6668–6679, 2016.

[24] Dorian Pustina, Harry Branch Coslett, Lyle Ungar, Olufunsho K Faseyitan, John D Medaglia, Brian Avants, and Myrna F Schwartz. Enhanced estimations of post-stroke aphasia severity using stacked multimodal predictions. *Human brain mapping*, 38(11):5603–5615, 2017.

[25] Thomas MH Hope, Alex P Leff, and Cathy J Price. Predicting language outcomes after stroke: Is structural disconnection a useful predictor? *NeuroImage: Clinical*, 19:22–29, 2018.

[26] Yusuf H Roohani, Noor Sajid, Pranava Madhyastha, Cathy J Price, and Thomas MH Hope. Predicting language recovery after stroke with convolutional networks on stitched mri. *arXiv preprint arXiv:1811.10520*, 2018.

[27] Thomas MH Hope, Mohamed L Seghier, Alex P Leff, and Cathy J Price. Predicting outcome and recovery after stroke with lesions extracted from mri images. *NeuroImage: clinical*, 2:424–433, 2013.

[28] Ethem Alpaydin. *Introduction to machine learning*. MIT press, 2020.

[29] Andrew Kertesz. *WAB-R: Western aphasia battery-revised*. PsychCorp, 2007.

[30] Sarah J Wallace, Linda Worrall, Tanya Rose, Guylaine Le Dorze, Caterina Breitenstein, Katerina Hilari, Edna Babbitt, Arpita Bose, Marian Brady, Leora R Cherney, et al. A core outcome set for aphasia treatment research: The roma consensus statement. *International journal of stroke*, 14(2):180–185, 2019.

[31] Natalie Gilmore, Erin L Meier, Jeffrey P Johnson, and Swathi Kiran. Typicality-based semantic treatment for anomia results in multiple levels of generalisation. *Neuropsychological rehabilitation*, pages 1–27, 2018.

[32] Elena Barbieri, Jennifer Mack, Brianne Chiappetta, Eduardo Europa, and Cynthia K Thompson. Recovery of offline and online sentence processing in aphasia: Language and domain-general network neuroplasticity. *Cortex*, 120:394–418, 2019.

[33] Robert W Wiley and Brenda Rapp. Statistical analysis in small-n designs: using linear mixed-effects modeling for evaluating intervention effectiveness. *Aphasiology*, 33(1):1–30, 2019.

[34] Jeffrey P Johnson, Erin L Meier, Yue Pan, and Swathi Kiran. Treatment-related changes in neural activation vary according to treatment response and extent of spared tissue in patients with chronic aphasia. *Cortex*, 121:147–168, 2019.

[35] Robert W Cox. AFNI: Software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and Biomedical research*, 29(3):162–173, 1996.

[36] Rahul S Desikan, Florent Ségonne, Bruce Fischl, Brian T Quinn, Bradford C Dickerson, Deborah Blacker, Randy L Buckner, Anders M Dale, R Paul Maguire, Bradley T Hyman, et al. An automated labeling system for subdividing the human cerebral cortex on mri scans into gyral based regions of interest. *Neuroimage*, 31(3):968–980, 2006.

[37] Matthew Brett, Jean-Luc Anton, Romain Valabregue, Jean-Baptiste Poline, et al. Region of interest analysis using an spm toolbox. In *8th international conference on functional mapping of the human brain*, volume 16, page 497. Sendai, Japan, 2002.

[38] K Rojkova, E Volle, M Urbanski, F Humbert, F Dell'Acqua, and M Thiebaut De Schotten. Atlasing the frontal lobe connections and their variability due to age and education: a spherical deconvolution tractography study. *Brain Structure and Function*, 221(3):1751–1766, 2016.

[39] Chris Foulon, Leonardo Cerliani, Serge Kinkingnehun, Richard Levy, Charlotte Rosso, Marika Urbanski, Emmanuelle Volle, and Michel Thiebaut de Schotten. Advanced lesion symptom mapping analyses and implementation as bcbtoolkit. *GigaScience*, 7(3):giy004, 2018.

[40] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[41] Jacques Wainer. Comparison of 14 different families of classification algorithms on 115 binary datasets. *arXiv preprint arXiv:1606.00930*, 2016.

[42] Leo Breiman. Random forests. *Machine Learning*, 45(32):5–32, 2001.

[43] Manuel Fernández-Delgado, Eva Cernadas, Senén Barro, and Dinani Amorim. Do we need hundreds of classifiers to solve real world classification problems? *The journal of machine learning research*, 15(1):3133–3181, 2014.

[44] Peter Fransson and Guillaume Marrelec. The precuneus/posterior cingulate cortex plays a pivotal role in the default mode network: Evidence from a partial correlation network analysis. *Neuroimage*, 42(3):1178–1184, 2008.

[45] Marcus E Raichle. The brain's default mode network. *Annual review of neuroscience*, 38:433–447, 2015.

[46] Sheng Zhang and R Li Chiang-shan. Functional connectivity mapping of the human precuneus by resting state fmri. *Neuroimage*, 59(4):3548–3562, 2012.

[47] Robert Leech, Salwa Kamourieh, Christian F Beckmann, and David J Sharp. Fractionating the default mode network: distinct contributions of the ventral and dorsal posterior cingulate cortex to cognitive control. *Journal of Neuroscience*, 31(9):3217–3224, 2011.

[48] Danilo Bzdok, Adrian Heeger, Robert Langner, Angela R Laird, Peter T Fox, Nicola Palomero-Gallagher, Brent A Vogt, Karl Zilles, and Simon B Eickhoff. Subspecialization in the human posterior medial cortex. *Neuroimage*, 106:55–71, 2015.

[49] Hanane El Hachioui, Evy G Visch-Brink, Hester F Lingsma, Mieke WME van de Sandt-Koenderman, Diederik WJ Dippel, Peter J Koudstaal, and Huub AM Middelkoop. Nonlinguistic cognitive impairment in poststroke aphasia: a prospective study. *Neurorehabilitation and Neural Repair*, 28(3):273–281, 2014.

[50] Natalie Gilmore, Erin L Meier, Jeffrey P Johnson, and Swathi Kiran. Nonlinguistic cognitive factors predict treatment-induced recovery in chronic poststroke aphasia. *Archives of physical medicine and rehabilitation*, 100(7):1251–1258, 2019.

[51] Anna Oh, Emma G Duerden, and Elizabeth W Pang. The role of the insula in speech and language processing. *Brain and language*, 135:96–103, 2014.

[52] Roozbeh Behroozmand, Hiroyuki Oya, Kirill V Nourski, Hiroto Kawasaki, Charles R Larson, John F Brugge, Matthew A Howard, and Jeremy DW Greenlee. Neural correlates of vocal production and motor control in human heschl's gyrus. *Journal of Neuroscience*, 36(7):2302–2315, 2016.

[53] David W Gow Jr. The cortical organization of lexical knowledge: a dual lexicon model of spoken language processing. *Brain and language*, 121(3):273–288, 2012.

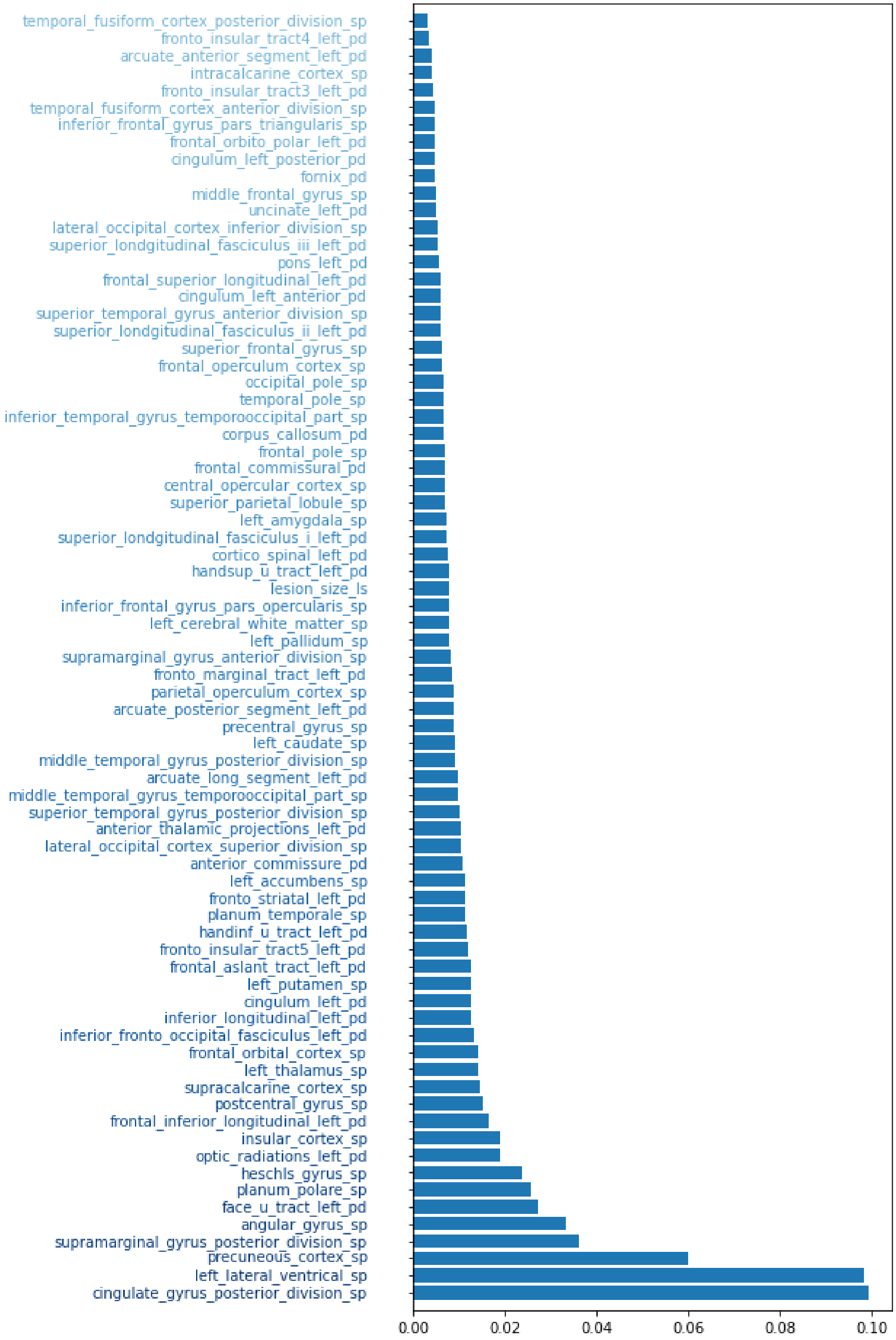

Figure 2: (a) Ranking of the brain structure features (M2). The longer the bar the more important the feature is. Sp is short for percentage spared, pd is for proportion of damage in white matter regions.

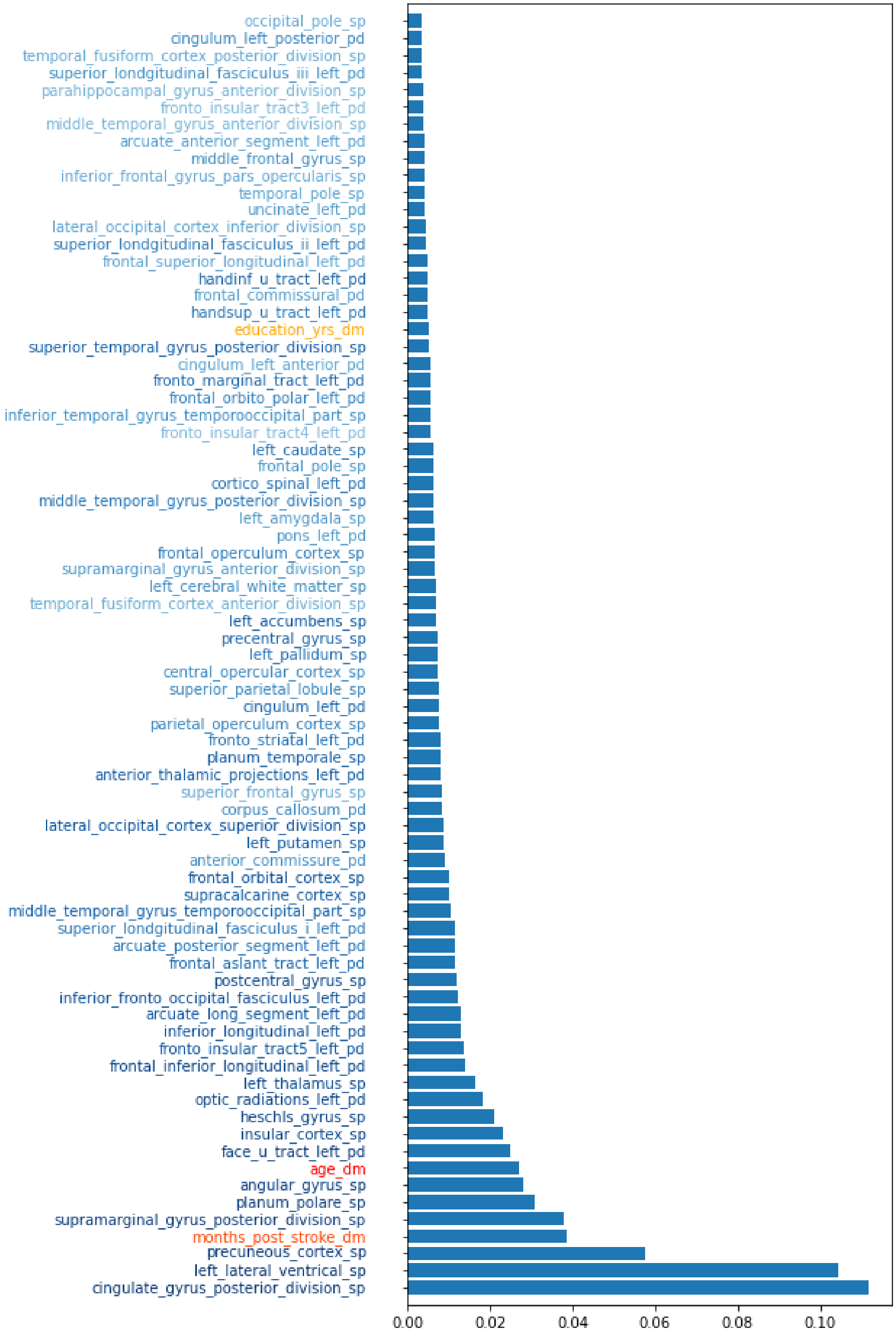

Figure 2: (b) Ranking of the features in M4: Brain structure (blue) and demographics features (red). Structure features use the same color scheme as in (a).

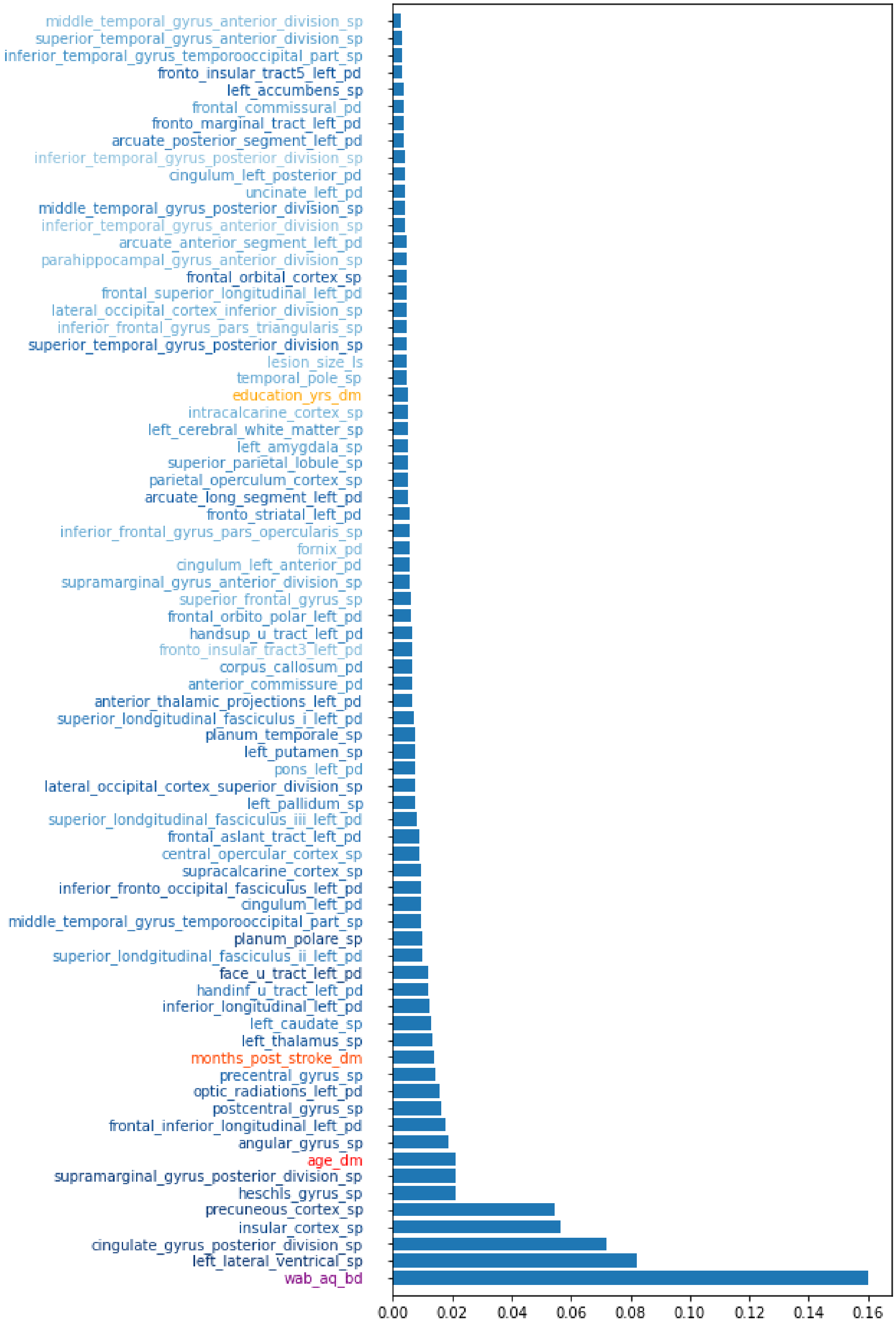

Figure 2: (c) Ranking of the features in M5: Brain structure (blue), demographics (red) and behavior features (purple). Color scheme is the same as in (a) and (b).