Fusion Approaches to Predict Post-stroke Aphasia Severity from Multimodal Neuroimaging Data

Saurav Chennuri, Sha Lai, Anne Billot, Maria Varkanitsa, Emily J. Braun, Swathi Kiran, Archana Venkataraman, Janusz Konrad, Prakash Ishwar, and Margrit Betke
Boston University
{saurav07,lais823,abillot,mvarkan,ejbraun,kirans,archanav,jkonrad,pi,betke}@bu.edu

Abstract

This paper explores feature selection and fusion methods for predicting the clinical outcome of post-stroke aphasia from medical imaging data. Utilizing a multimodal neuroimaging dataset derived from 55 individuals with chronic aphasia resulting from left-hemisphere lesions following a stroke, two distinct approaches, namely Early Fusion and Late Fusion, were developed using Support Vector Regression or Random Forest regression models for prognosticating patients' functional communication skills measured by Western Aphasia Battery (WAB) test scores. A supervised learning method is proposed to reduce the number of features derived from each imaging modality. The fusion approaches were then applied to find combinations of these reduced feature sets that yield the most accurate WAB predictions. The same nested training/validation/test sets were used for the feature selection and fusion methods. Experiments showed that the best model based on the correlation metric is a Late Fusion RF model (r=0.63), while the best model based on the RMSE is an Early Fusion SVR model (RMSE=16.72). Experiments also revealed several feature set combinations that yielded more accurate predictions than both single-modality feature sets and feature sets that combine all modalities, justifying both fusion and reduction of features derived from multimodal neuroimaging data. It was also found that the percentage of tissue in gray matter regions of the brain, spared by the stroke as identified on structural Magnetic Resonance Imaging, is the single feature set that appeared in all highest ranked feature set combinations of both fusion approaches.

1. Introduction

Aphasia, a language disorder characterized by impairment in language comprehension and speech production, affects approximately one-third of stroke survivors [15]. Predicting the evolution of aphasia is important for clinicians to choose treatment and for patients to engage in their recovery process. Analyzing multimodal neuroimaging data of individuals with aphasia, previous work has focused on the binary problem of predicting whether patients will respond or not respond to treatment, measured by changes in patients' Western Aphasia Battery (WAB) test scores [2, 9, 19]. Other works have proposed methods to predict WAB scores from multimodal neuroimaging data but with features selected using the entirety of the dataset [18, 23]. In this work, we develop approaches for predicting WAB scores by selecting features only using data from the training folds and not test folds and thus avoiding potentially biasing the predictors.

The creation of predictive models for neuroimaging data faces challenges stemming from an imbalance between the large number of features and the limited number of subjects available for analysis. This imbalance raises concerns about overfitting, wherein the predictive models may become overly tailored to the specific dataset, potentially reducing their generalization capabilities. To address this, researchers commonly employ feature reduction strategies, but commonly this has been done using the entirety of the dataset. We show how a nested cross-validation procedure can be used to avoid selecting features using the entire dataset but rather choosing them within each training fold of a supervised learning scheme that also uses the same training/test data splits as the validation experiments of the modality fusion methods.

The input features to the proposed prediction methods have been processed from resting-state functional Magnetic Resonance Imaging (fMRI), Diffusion Tensor Imaging (DTI), and structural MRI. These neuroimaging modalities offer the opportunity to capture complementary information about brain structure, connectivity, and function. By leveraging the strengths of each modality, studies that combine multimodal neuroimaging data have the potential to provide a more comprehensive and nuanced understanding of the relationship between brain damage intact brain tissue functionality, and language function in individuals with aphasia [18].

In multimodal medical data analysis, early and late fusion approaches have been used, which are also called feature and probability fusion approaches [12, 14]. Our work proposes two fusion approaches to combine aphasia input data. In the early fusion approach, the reduced feature sets derived from each imaging modality are stacked into a single vector that is then interpreted by a prediction model, either a Support Vector Regression (SVR) or a Random Forest (RF) regression model (two widely used classical machine learning models), which predicts the patient's WAB test score. The late fusion approach has two phases. In the first phase, each reduced feature set is interpreted by its own prediction model, again, either an SVR or RF. In the second phase, the prediction values obtained by these models are then stacked and passed into another SVR or RF that serves as an interpreter of these predictions and yields the patient's predicted WAB test score.

Our experiments show that there is no clear winner among the two fusion approaches. However, more importantly, when applied to an exhaustive number of multimodal feature combinations, these approaches produce prediction models that can be used to evaluate which feature combinations are most useful as predictors of aphasia severity. They also show the importance of feature reduction.

In summary, this paper contributes to the existing literature by

- Developing two fusion strategies to predict aphasia severity of patients from features derived multimodal neuroimaging data,
- Proposing a principled supervised feature selection method,
- Determining predictive importance of different multimodal combinations of feature sets.

2. Background

Aphasia and WAB test scores. Among the various consequences of stroke, Aphasia stands out as one of the most devastating conditions. Clinicians face considerable challenges in accurately assessing the severity of Aphasia and tailoring appropriate interventions to meet the unique needs of each patient. Our study focuses on predicting the severity of Aphasia in individuals with left-hemisphere lesions, who are known to be at higher risk for language impairments.

To address the challenges in aphasia assessment, the Western Aphasia Battery (WAB) test has emerged as a widely used clinical tool to measure the severity of aphasia by evaluating various linguistic skills, such as speech, fluency, auditory comprehension, reading and writing, and well as some non-linguistic skills, such as drawing. High scores correlate with good functional communication skills in stroke patients with aphasia [1]. In this study, we use patient scores from the revised WAB test [16], which we simply call WAB scores.

Previous Neuroimaging Work. The investigation of brain damage manifestation in aphasia through the analysis of multimodal neuroimaging data within the same study has been the focus of prior research efforts, incorporating various structural and functional measures for diverse objectives, including the identification of neurological biases of language and the examination of language recovery [2, 8, 20, 25, 27]. However, our current understanding of the factors distinguishing treatment responders from non-responders in aphasia is limited [6]. Therefore, further exploration of this topic is crucial to advance our comprehension of the intricate relationships between brain damage, neural functionality, and language impairments in individuals with aphasia. Additionally, investigating the influence and interaction of lesion size and location with functionally intact brain regions holds promise for the development of personalized rehabilitation strategies in aphasia, addressing the individual variability in treatment response [5, 17].

The combination of multiple neuroimaging modalities within the context of post-stroke aphasia and its association with aphasia severity remains an ongoing area of investigation. Efforts have been made to examine the patterns of brain damage and structurally and functionally intact cortical and subcortical regions that correspond to aphasia severity [10, 13, 23, 30]. All these works are based on the rationale that using multi-modal complex higher-order functions like language are represented in widely distributed cortical networks [21, 26, 28].

Previous Machine Learning Work. As mentioned above, previous work proposed machine learning models that address the binary problem of distinguishing patients that will respond or not respond to treatment [2, 9, 19]. Gu et al. [9] suggested the use of a Random Forest model as the binary prediction model due to the interpretability of the results. The feature set included only structural MRI data and patient demographics. Their work pointed to the importance of the features that measured the fraction of spared brain issue in gray matter regions. Lai et al. [19] proposed the use of Support Vector Machine and Random Forest models for the binary prediction problem, additionally using functional MRI data. Billot et al. [2] developed the most comprehensive study on binary prediction of aphasia response, including multimodal data from structural and functional MRI, as well as diffusion Tensor imaging (DTI). Their results show the importance of aphasia severity, demographics, measures of anatomic integrity and resting-state functional connectivity as predictive feature combinations.

Two previous studies addressed the regression problem of predicting the WAB test score of aphasia patients [18, 23]. The study by Pustina et al. [23] proposed a method, called STAMP for "stacked multimodal predictions" that

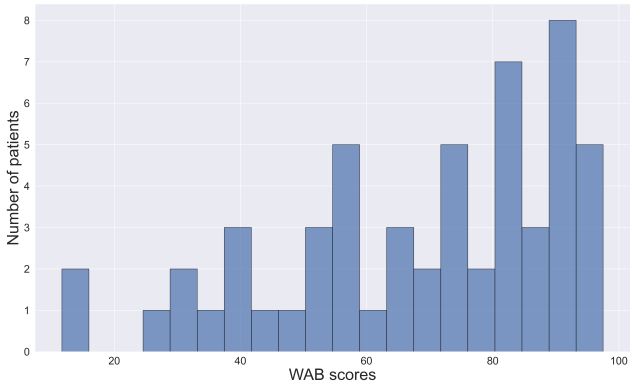

Figure 1. Histogram of the WAB scores of the 55 patients in our study (mean = 68.5, median = 74.3, std. dev. = 22.7).

uses features derived from structural and functional MRIs and DTI, including graph theoretic features. The method involved a two-phase approach, where features were eliminated recursively in the first phase by modality-specific Random Forests. The resulting modality-specific predictions were interpreted by another Random Forest in the second phase. Kristinsson et al. [18] proposed the use of SVR prediction models directly on the stacked features derived from structural and functional MRI and cerebral blood flow. They showed that a feature-reduced multimodal prediction model yielded the most accurate prediction (up to r=0.67) compared to models that used the neuroimaging modalities separately or all features combined. The results of our study, showing the benefits of multimodal prediction, align with the literature. In addition, our work avoids selecting feature combinations using the entire dataset, which we will henceforth call *Oracle Selection*, an issue that both Pustina et al. [23] and Kristinsson et al. [18] mention, and proposes a more principled supervised selection method instead.

3. Imaging Data and Feature Extraction

This study utilizes a dataset collected from 55 patients recruited at Boston University, Johns Hopkins University, and Northwestern University. These patients demonstrate varying levels of aphasia severity, as depicted in Figure 1. We processed the patient structural MRI scans to compute lesion size, white matter and grey matter remaining percentages in each brain region of the left hemisphere, given their strong association with language-related functions. From fMRI scans, we obtained resting-state measurements in the form of 50×50 connectivity matrices from which we extracted a subset of 625 values as in [2] and also computed graph theoretic measures (betweenness, degree, efficiency, transitivity) . We used the Brain Connectivity Toolbox [24], see also Bullmore and Sporns (2009) [4]. DTI data produced diffusion-based fractional anisotropy (FA) values.

Table 1. Neuroimaging dataset collected from 55 patients

Data	**Feature Set**	**Symbol**	**#**
Dem.	Age, education, time poststroke	DM	3
DTI	Fractional anisotropy	FA	12
fMRI	Bidirectional correlations	RS	625
	Betweenness	Bet	50
	Degree	Deg	50
	Efficiency	Eff	50
	Transitivity	Trans	50
MRI	Percent of spared gray matter	PSG	69
MRI	Percent of spared white matter	PSW	36

We also used demographic information of the patients (age, post-stroke onset of aphasia, education) as a feature set. For convenience, we assigned shorthand notations to each modality, with their dimensions presented in Table 1. The nine feature sets have a total number of 945 dimensions.

4. Supervised Feature Selection

We propose a Supervised Feature Selection algorithm (see pseudocode of Algorithms 1 and 2 below) that determines how many and which features will be used by a particular fusion method to predict WAB scores. The algorithm computes this per modality, typically reducing the size of a feature set significantly. Feature reduction is particularly important for the resting state data (as noted previously by others [19, 23]), which here contain 625 features per patient. Per modality, the algorithm is called four times, computing separately, for each fusion approach (early or late fusion) and each prediction model (SVR or RF), the optimal feature set to be used in predicting WAB scores. The optimality criterion will be described below.

Our design of the Supervised Feature Selection algorithm was inspired by nested cross-validation procedures that split a dataset into three disjoint sets, *[[train:validate]:test]*, train a machine learning model on the *train* data, validate its hyperparameters on the *validate* data, and then test its performance on the *test* data. Reserving the data in a particular *test* set D_{test} for later (lines 7, 8 of Algorithm 2 for early/late fusion, different splits of *[train:validate]* data are evaluated for feature selection within Algorithm 1 (lines 7, 8, 9) The Supervised Feature Selection algorithm trains and tunes the hyperparameters of an SVR or an RF in this manner. These hyperparameters include the number k of features to be kept for a given modality.

Specifically, for each *train* data set D_{train}, Algorithm 1 searches through possible numbers k of features (line 4). Within each training sub-fold D'_{train} (among 10 sub-folds), Algorithm 1 then computes the empirical cross-correlation

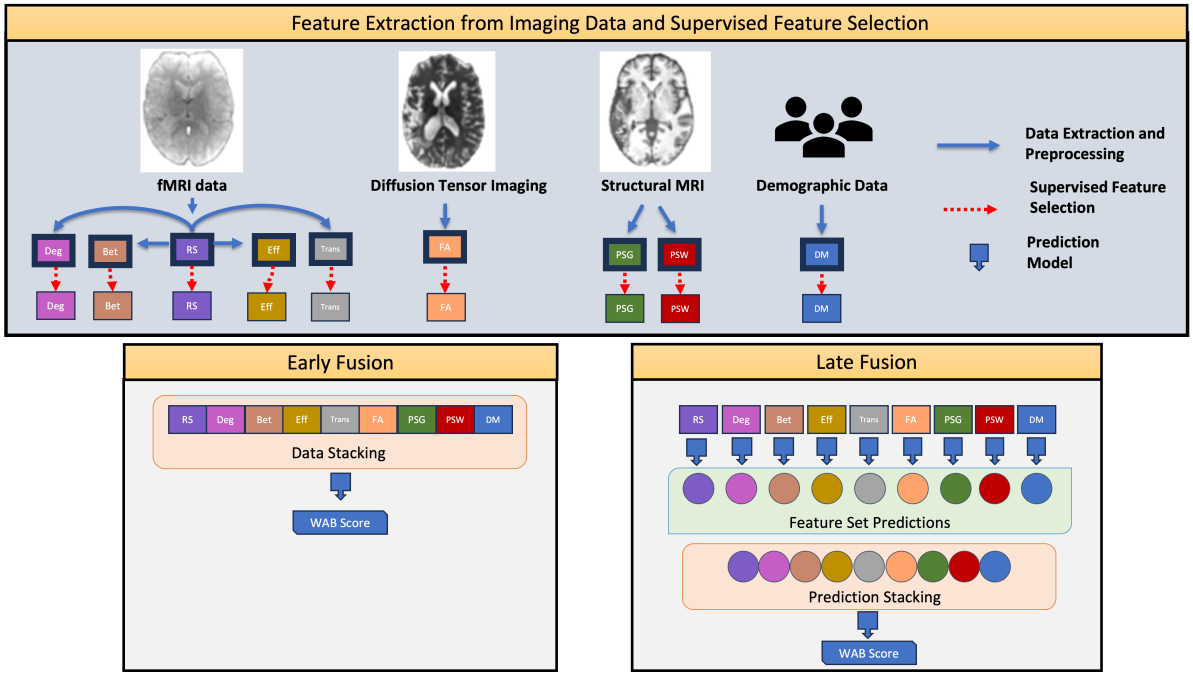

Figure 2. Overview of method. Imaging data are processed into feature sets that are reduced using the Supervised Feature Selection Algorithm. Selected feature sets are used by Late and Early Fusion approaches to predict aphasia recovery outcome using RF or SVM prediction models.

coefficient between the feature value (for each feature s in the feature-set S) and the ground-truth WAB score across all patients in D'_{train}. It then sorts features in decreasing order of the the absolute value of their cross-correlation coefficient (line 10). Then for the k under consideration, features S_k which occur among the top k features most often across all 10 training sub-folds (see line 18) are chosen to train M (SVR or RF) to predict WAB scores using the entire training fold D_{train} (see line 19) and tune its hyperparameters using the same 10 sub-folds created previously in line 3. This yields a model M_k together with its average 10-fold cross-validated predictive performance (RMSE) V_k (lines 20, 21). The value of k with the best V_k (smallest RMSE) is the optimum number of features k^* for D_{train} and S_{k^*} is the optimum set of features for D_{train} returned by Algorithm 1. Note that this optimality is specific to D_{train}, and so these optimal features can be different for each test set and modality. The tests sets and features selected in this manner are used to evaluate the fusion methods we propose next.

5. Multimodal Fusion Approaches for WAB Score Prediction

In this study, we explored two distinct stacking-based multimodal approaches, Early and Late Fusion, to predict WAB test scores using different combinations of modalities (see Fig. 2).

The Early Fusion (EF) approach takes the reduced feature sets, computed by the Supervised Feature Selection algorithm for each neuroimaging modality, and stacks them into a single vector that serves as the input to a WAB score prediction model, i.e., a SVR or a RF.

The Late Fusion (LF) approach uses "prediction stacking," a strategy used in machine learning where preliminary predictions (see line 22 in Algorithm 1) from different data sources are combined to make final predictions [3, 29]. Specifically, a separate prediction model (also either a SVR or a RF) for each imaging modality is used to predict a WAB score based on the reduced feature set for that modality. The resulting set of WAB scores are then stacked into a score vector that serves as the input to a single prediction model that then predicts a WAB score for the combined input.

The visualization of the fusion approaches in Fig. 2 shows the combined use of every feature modality. Details of training and evaluating both fusion approaches are explained via pseudocode in Algorithm 2. We also evaluate the fusion approaches for combinations of feature sets that only include a subset of modalities. Given that there are 9 modalities, there are $2^9 - 1 = 511$ feature set combinations, which we all tested for each of the two fusion approaches and two types of prediction models.

Algorithm 1 Supervised Feature Selection (S, D_{train}, M)

Input: Feature set S, training data fold D_{train}, regression model M
Output: Optimum set of features and prediction set
Split: D_{train} into 10 disjoint equal-size subsets randomly
for $k \leftarrow 1$ to $|S|$ **do**
  Initialize: for each s in S, **set** $C_s \leftarrow 0$
  for $j \leftarrow 1$ to 10 **do**
    Denote: the jth subset of D_{train} as $D^{(j)}_{\text{train}}$
    Designate: $D_{\text{val}} \leftarrow D^{(j)}_{\text{train}}$
    Designate: $D'_{\text{train}} \leftarrow D_{\text{train}} \setminus D^{(j)}_{\text{train}}$
    Rank: features by the magnitude of their corre-
      -lation with WAB scores within D'_{train}
    for each $s \in S$ **do**
      if s appears among top-k features **then**
        increase C_s by 1
      end if
    end for
  end for
  Designate: $S_k \leftarrow \{s | C_s \in \text{top } k \text{ of } C_1, \ldots, C_{|S|}\}$
  Train & Tune: M with S_k using D_{train} and the same 10-fold cross-validation as in line 3
  Denote: the trained and tuned model as M_k
  Denote: the performance of M_k on D_{train} as V_k
  Denote: the predictions of M_k on D_{train} as P_k
end for
Denote: $k^* \leftarrow \arg\min_k V_k$
Return: S_{k^*}, P_{k^*}

Algorithm 2 Early/Late Fusion (D, M)

Input: Dataset D, model M
Output: Predictions list P
Split: D into 11 equal-size subsets randomly
Initialize: P as an empty list of predictions
for $i \leftarrow 1$ to 11 **do**
  Denote: the ith subset of D as $D^{(i)}$
  Designate: $D_{\text{test}} \leftarrow D^{(i)}$
  Designate: $D_{\text{train}} \leftarrow D \setminus D^{(i)}$
  for each feature set S **do**
    $(S_{k^*}, P_{k^*}) =$ Algorithm 1(S, D_{train}, M)
  end for
  Denote: $T \leftarrow \{S_{k^*}\}$ family of all reduced feature sets
  Denote: $P_T \leftarrow \{P_{k^*}\}$ family of all prediction-sets of reduced feature sets.
  if Early Fusion **then**
    for all feature-set combinations $t \subseteq T$ **do**
      Train & Tune: M with t on D_{train} and the same 10-fold cross-validation as in line 3 of Algo. 1
      Append: predictions on D_{test} to P
    end for
  end if
  if Late Fusion **then**
    for all prediction-set combinations $p \subseteq P_T$ **do**
      Train & Tune: M with p on D_{train} and the same 10-fold cross-validation as in line 3 of Algo. 1
      Append: predictions on D_{test} to P
    end for
  end if
end for

6. Cross Validation Experiments

We utilized the Support Vector Regressor [7] and the Random Forest Regressor [11] as our prediction models. For the SVR, we tuned 5 hyperparameters, their testing parameter choices are listed in **Table** 2. The RF underwent tuning for three hyperparameters: the maximum depth of trees, the maximum number of features considered during the splitting process, and the maximum number of trees used for predictions. Details of the hyperparameter choices for the Random Forest can be found in **Table** 3. All experiments were conducted using the sci-kit-learn library [22].

To evaluate the predictive performance, we measured the Root Mean Square Error (RMSE) of the trained prediction models in predicting WAB scores. The RMSE offers an indication of the proximity between the predictions and the actual values. We also employed the normalized cross-correlation between the predicted WAB scores and the ground truth scores as a performance metric. This metric provides insights into the linear association between the predicted and actual scores.

We used the same data splits into nested *[[train:validate]:test]* sets as we used for the Supervised Feature Selection algorithm, described in Section 4. With these data splits, we conducted a nested cross-validation procedure to train and tune the SVR and RF models used in the fusion approaches. Given the data of 55 patients, we used a data split of [45:5:5]. This means that

Table 2. Hyperparameters explored for Support Vector Regressor

Hyperparameter	**Values**
kernel	Linear, RBF
C	1, 10, 50
gamma	10^{-3},10^{-2},10^{-1},1
tolerance	10^{-6},10^{-5},10^{-4},10^{-3},10^{-2},10^{-1},1
epsilon	10^{-6},10^{-5},10^{-4},10^{-3},10^{-2},10^{-1},1

Table 3. Hyperparameters explored for Random Forest Regressor

Hyperparameter	**Values**
Max. Depth	2, 5, 7, 10, 15, 20, 25, 30
Max. Features	sqrt, $\log_2$, 0.1, 0.2, 0.3
No. of Estimators	100, 150, 250, 300

the "outer loop" involved 11 tests of 5 patients (11-fold cross-validation), and the "inner loop" 10 validations of trained models (10-fold cross-validation).

In addition to conducting four cross-validation experiments for each of the four fusion/prediction model pairs, where the features are selected by the Supervised Selection Algorithm, we also conducted the experiments with Oracle Selection of features. With Oracle Selection, the values of a particular feature for *all* patients are stacked into a 55-dimensional vector and correlated with the ground-truth WAB scores of the patients. The correlation values are sorted to yield the ranking of the features. The optimal number k per modality is chosen as in the Supervised Selection Algorithm.

7. Results and Discussion

The average performance results of our cross-validation experiments are given in Table 4. We list the average RMSE and r for the the four fusion/prediction model pairs across the 11 test folds. We show the results when the features were selected by the Supervised Selection Algorithm and the Oracle Selection method. Both fusion approaches and prediction model types yield similar performance values. With the Supervised Selection Algorithm, for both RF and SVR, the Early Fusion Approach has lower RMSEs than the Late Fusion Approach but has higher correlation values for RF. This ordering is reversed for the Oracle Selection Method.

The performance with Oracle Selection is significantly higher than with the Supervised Selection Algorithm. This is noteworthy. It shows that having access to all data for feature selection gives the machine learning model an advantage. We warn that these higher levels might be misleading. The models are not likely to uphold such high levels for unseen data, as our experiments with the Supervised Selection Algorithm shows.

The Supervised Selection Algorithm when used with an SVR model reduced the number of features from those in Table 1) to the following mean values per modality: DM 2, FA 6, PSG 18, PSW 12, RS 75, Bet 42, Deg 13, Eff 7, Tran 12. When used with a RF, the features were on average reduced to: DM 2, FA 5, PSG 7, PSW 31, RS 116, Bet 33, Deg 37, Eff 20, Tran 14.

We further study the performance of our methods with respect to feature combinations. We list the fold-mean and standard deviation values (across folds) of RMSE and r for the top ranked feature combinations (ranked by mean-fold-RMSE), all features, and single features in Table 5 for Early fusion with RF. The table shows that top ranked feature combinations have similar RMSE and r values. Using all modalities is not advantageous (only rank 50), and prediction based on single feature sets performs poorly. For the other three approaches, we show the top three ranked feature combinations (ranked by mean-fold RMSE), as well as the best combination based on r, and the results for combining all features in Tables 6, 7, and 8.

Table 4. Average prediction performance on test sets in nested cross-validation experiment.

	Early Fusion		**Late Fusion**	
	RF	**SVR**	**RF**	**SVR**
	Supervised Feature Selection			
Mean RMSE	17.41	16.72	17.45	16.72
Mean r	0.48	0.64	0.58	0.58
	Oracle Selection			
Mean RMSE	15.19	14.4	13.78	13.36
Mean r	0.80	0.77	0.80	0.81

Feature importance plots: To gauge the importance of each feature set, we measure the number of times each feature set appears within top-ranked combinations based on mean-fold RMSE metric for rank values ranging from 1 through 511, as shown in Figure 3. The left subplots illustrate the cumulative frequency across all modality combinations, while the right graphs specifically focus on the top-20 modality combinations to give a better idea of the best-performing feature set combinations. The discussion below is based on Figure 3.

PSG is an important predictor: In the plots for Late Fusion with SVR and RF, PSG visibly stands out as the most included feature set among all the feature set combinations. This is also observed in the Early Fusion approach with RF. In the case of Early Fusion with SVR, Percent Spared White Matter (PSW), Demographics information (DM) and Percent Spared Gray Matter(PSG) frequently occur among the most included feature sets based on different numbers of top feature set combinations.

Next most important predictor: For the Late Fusion approach, RF and SVR have different best predictors beyond PSG. RF gives more importance to Bet as can be seen from the LF/RF plot, while SVR gives more importance to FA and Eff. When we focus only on the top-20 feature set combinations for Late Fusion, RF gives almost the same importance to PSG and Bet in the first 3 or 4 combinations, but Bet becomes prominent thereafter and other feature sets have varying levels of importance. In the case of SVR, FA occurs almost as frequently as PSG until the top 18 combinations beyond which Eff takes over as the next most included feature set.

For the Early Fusion approach, RF includes Bet as the next most included feature set as can be seen from the plot for top-20 feature set combinations. SVR in Early Fusion stands apart from LF-RF, LF-SVR, and EF-SVR as it gives almost equal importance to PSW, DM, PSG, and FA as can be seen from the top-20 feature set combinations. Based on the above observations, beyond PSG, the next most included predictor varies based on the fusion approach and

Table 5. **Early Fusion with RF as the prediction model.** Lowest mean RMSE and highest r in bold. Results for all features in italics.

Rank	Feature Set Combination	Mean-fold RMSE	Std. Dev. RMSE	Mean-fold r	Std. Dev. r	Remarks
1	**DM PSG**	**17.41**	5.95	0.48	0.46	
2	PSG PSW bet	17.50	5.50	0.53	0.35	
3	DM PSG PSW bet	17.50	5.74	0.51	0.38	Top 5 RMSE
4	DM PSG bet	17.56	5.89	0.53	0.35	
5	FA PSG bet	17.57	5.52	0.54	0.32	
23	**DM FA PSG PSW RS deg**	18.35	5.12	**0.58**	0.24	Top r
50	*DM FA PSG PSW RS bet deg eff trans*	18.63	6.13	0.53	0.27	All modalities
24	PSG	18.36	6.04	0.49	0.47	
322	FA	20.85	8.37	0.50	0.35	
334	PSW	20.97	5.95	0.52	0.33	
339	bet	21.07	6.84	0.39	0.40	
426	RS	22.19	8.61	0.11	0.56	Individual
502	trans	23.14	7.01	0.28	0.49	
506	DM	23.45	6.99	-0.12	0.31	
508	eff	23.50	6.70	0.29	0.20	
510	deg	23.68	6.32	0.21	0.49	

Table 6. **Early Fusion with SVR as the prediction model.** Lowest mean RMSE and highest r in bold. Results for all features in italics.

Rank	Feature Set Combination	Mean-fold RMSE	Std. Dev. RMSE	Mean-fold r	Std. Dev. r	Remarks
1	**DM FA PSG PSW eff**	**16.72**	5.14	0.59	0.34	
2	DM PSG PSW trans	16.96	6.31	0.59	0.37	Top 3 RMSE
3	DM FA PSG PSW eff trans	16.97	5.10	0.51	0.43	
8	**DM PSG PSW**	17.20	6.30	**0.60**	0.37	Top r
224	*DM FA PSG PSW RS bet deg eff trans*	22.33	10.22	0.00	0.00	All modalities

Table 7. **Late Fusion with RF as the prediction model.** Lowest mean RMSE and highest r in bold. Results for all features in italics.

Rank	Feature Set Combination	Mean-fold RMSE	Std RMSE	Mean-fold r	Std. Dev. r	Remarks
1	**PSG RS bet deg**	**17.45**	7.39	0.50	0.43	
2	PSG bet trans	17.47	7.28	0.52	0.38	Top 3 RMSE
3	PSG bet deg	17.53	8.09	0.50	0.43	
72	**PSW bet deg trans**	20.00	8.20	**0.63**	0.28	Top r
285	*DM FA PSG PSW RS bet deg eff trans*	21.16	7.20	0.40	0.42	All modalities

Table 8. **Late Fusion with SVR as the prediction model.** Lowest mean RMSE and highest r in bold. Results for all features in italics.

Rank	Feature Set Combination	Mean-fold RMSE	Std RMSE	Mean-fold r	Std. Dev. r	Remarks
1	**FA PSG eff trans**	**16.85**	7.85	0.20	0.48	
2	FA PSG eff	16.98	8.36	0.22	0.47	Top 3 RMSE
3	DM FA PSG trans	17.15	7.37	0.21	0.37	
116	**PSW**	20.29	8.14	**0.55**	0.35	Top r
132	*DM FA PSG PSW RS bet deg eff trans*	20.51	11.53	0.14	0.36	All modalities

choice of the prediction model. But there are some consistencies among certain predictors like Bet, DM and FA which are given more importance by both RF and SVR and they are almost similarly included as PSG among the top-20

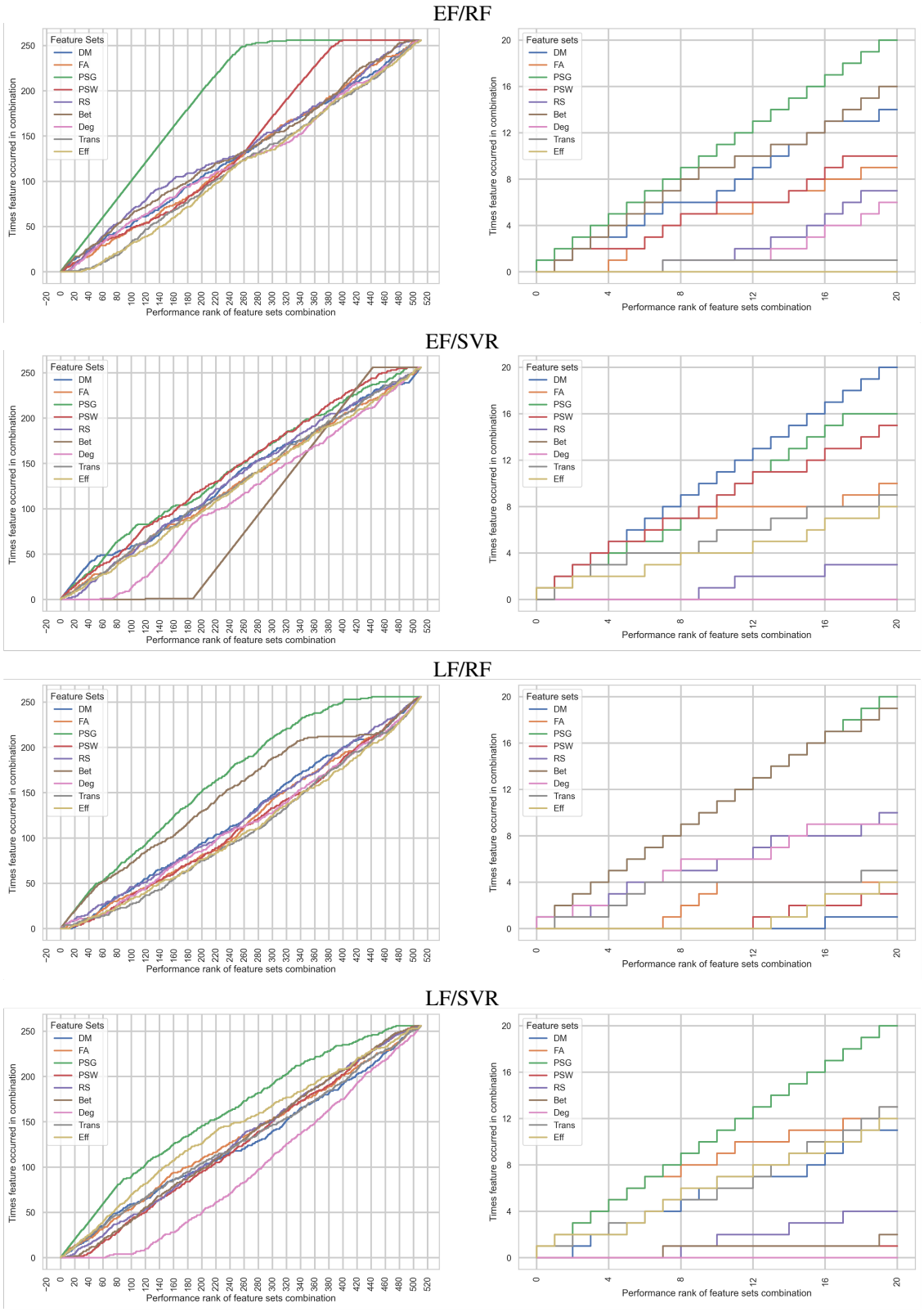

Figure 3. Feature set cumulative occurrence plots Early (EF) and Late (LF) Fusion and RF/SVR models. The performance rank is computed based on Mean-fold RMSE

combinations in both LF and EF approaches.

Early Fusion versus Late Fusion: To compare both these approaches, we look at their mean-fold-RMSE performances. The Late Fusion approach has 17.45 and 16.85 as the best mean-fold RMSE values for RF and SVR correspondingly (see tables 7 and 8), while Early Fusion approach has 17.41 and 16.72 as best mean-fold RMSE values for RF and SVR (see tables 5 and 6). This pattern of the Early Fusion approach having visibly better results than the Late Fusion approach holds true among the top-20 feature set combinations. But the std dev of RMSE values among folds being high, we cannot claim any of these approaches is significantly better than the other. A further analysis by checking for statistical significance among the top-feature set combinations via Wilcoxon-signed ranked t-test on MSE values (omitted here due to limited space) indicated that there are no significant differences among the top-performing feature set combinations. This suggests that there is no clear winner among both these approaches. Results of the t-test also indicated that the top-performing feature set combinations are significant over the individual feature sets and all the feature sets combined.

8. Conclusions

We proposed a supervised feature selection algorithm and two fusion approaches for predicting post-stroke aphasia recovery outcomes. Despite relatively low performance values, our results show that feature reduction and the use of multimodal features combinations are advantageous strategies and increase the overall interpretability of the results. Through our experiments with an exhaustive number of multimodal feature combinations, we identified feature combinations that are useful as predictors of recovery outcomes. Combinations that include the Percent Spared Gray Matter feature set are particularly effective.

References

[1] A.M.O. Bakheit, S. Carrington, S Griffiths, and K. Searle. High scores on the Western Aphasia Battery correlate with good functional communication skills (as measured with the Communicative Effectiveness Index) in aphasic stroke patients. *Disability and Rehabilitation*, 27(6):287—291, 2005. doi:10.1080/09638280400009006. 2

[2] Anne Billot, Sha Lai, Maria Varkanitsa, Emily J Braun, Brenda Rapp, Todd B Parrish, James Higgins, Ajay S Kurani, David Caplan, Cynthia K Thompson, Prakash Ishwar, Margrit Betke, and Swathi Kiran. Multimodal neural and behavioral data predict response to rehabilitation in chronic poststroke aphasia. *Stroke*, 53(5):1606–1614, Jan. 2022. 1, 2, 3

[3] Leo Breiman. Stacked regressions. *Machine Learning*, 24(1):49–64, Jul 1996. 4

[4] Ed Bullmore and Olaf Sporns. Complex brain networks: graph theoretical analysis of structural and functional systems. *Nat Rev Neurosci*, 10(3):186–198, Feb. 2009. 3

[5] Bruce Crosson, Amy D Rodriguez, David Copland, Julius Fridriksson, Lisa C Krishnamurthy, Marcus Meinzer, Anastasia M Raymer, Venkatagiri Krishnamurthy, and Alexander P Leff. Neuroplasticity and aphasia treatments: new approaches for an old problem. *J Neurol Neurosurg Psychiatry*, 90(10):1147–1155, May 2019. 2

[6] Madeline Cruice, Linda Worrall, and Louise Hickson. Reporting on psychological well-being of older adults with chronic aphasia in the context of unaffected peers. *Disability and rehabilitation*, 33(3):219–28, 2011. 2

[7] Harris Drucker, Chris J. C. Burges, Linda Kaufman, Alex Smola, and Vladimir Vapnik. Support vector regression machines. In *Proceedings of the 9th International Conference on Neural Information Processing Systems*, NIPS'96, page 155–161, Cambridge, MA, USA, 1996. MIT Press. 5

[8] Julius Fridriksson, Dirk-Bart den Ouden, Argye E Hillis, Gregory Hickok, Chris Rorden, Alexandra Basilakos, Grigori Yourganov, and Leonardo Bonilha. Anatomy of aphasia revisited. *Brain*, 141(3):848–862, Mar. 2018. 2

[9] Yiwen Gu, Murtadha Bahrani, Anne Billot, Sha Lai, Emily J. Braun, Maria Varkanitsa, Julia Bighetto, Brenda Rapp, Todd B. Parrish, David Caplan, Cynthia K. Thompson, Swathi Kiran, and Margrit Betke. A machine learning approach for predicting post-stroke aphasia recovery: A pilot study. In *Proceedings of the 13th ACM International Conference on PErvasive Technologies Related to Assistive Environments*, PETRA '20, New York, NY, USA, 2020. Association for Computing Machinery. 1, 2

[10] Ajay D Halai, Anna M Woollams, and Matthew A Lambon Ralph. Investigating the effect of changing parameters when building prediction models for post-stroke aphasia. *Nat Hum Behav*, 4(7):725–735, Apr. 2020. 2

[11] Tin Kam Ho. Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, volume 1, pages 278–282. IEEE, 1995. 5

[12] Gregory Holste, Savannah C. Partridge, Habib Rahbar, Debosmita Biswas, Christoph I. Lee, and Adam M.

Alessio. End-to-end learning of fused image and non-image features for improved breast cancer classification from MRI. In *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 3287–3296, 2021. 2

[13] Thomas M H Hope, Alex P Leff, and Cathy J Price. Predicting language outcomes after stroke: Is structural disconnection a useful predictor? *Neuroimage Clin*, 19:22–29, Mar. 2018. 2

[14] Shih-Cheng Huang, Anuj Pareek, Saeed Seyyedi, Imon Banerjee, and Matthew P. Lungren. Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. *npj Digital Medicine*, 3(136), Oct 2020. 9 pp. 2

[15] M.L. Kauhanen, JT Korpelainen, P Hiltunen, R. Määttä, Mononen H., Brusin E., Sotaniemi K.A., and VV. Myllylä. Aphasia, depression, and non-verbal cognitive impairment in ischaemic stroke. *Cerebrovasc Dis*, 10:455—461, 2000. doi: 10.1159/000016107. 1

[16] Andrew Kertesz. *WAB-R : Western Aphasia Battery-Revised*. Pearson Clinical Assessment, PsychCorp, San Antonio, TX, 2007. 2

[17] Swathi Kiran and Cynthia K Thompson. Neuroplasticity of language networks in aphasia: Advances, updates, and future challenges. *Front Neurol*, 10:295, Apr. 2019. 2

[18] Sigfus Kristinsson, Wanfang Zhang, Chris Rorden, Roger Newman-Norlund, Alexandra Basilakos, Leonardo Bonilha, Grigori Yourganov, Feifei Xiao, Argye Hillis, and Julius Fridriksson. Machine learning-based multimodal prediction of language outcomes in chronic aphasia. *Human Brain Mapping*, 42(6):1682–1698, 2021. 1, 2, 3

[19] Sha Lai, Anne Billot, Maria Varkanitsa, Emily Braun, Brenda Rapp, Todd Parrish, Ajay Kurani, James Higgins, David Caplan, Cynthia Thompson, Swathi Kiran, Margrit Betke, and Prakash Ishwar. An exploration of machine learning methods for predicting post-stroke aphasia recovery. In *The 14th PErvasive Technologies Related to Assistive Environments Conference*, PETRA 2021, pages 556—564, New York, NY, USA, 2021. Association for Computing Machinery. 1, 2, 3

[20] Erin L Meier, Jeffrey P Johnson, Yue Pan, and Swathi Kiran. The utility of lesion classification in predicting language and treatment outcomes in chronic stroke-induced aphasia. *Brain Imaging Behav*, 13(6):1510–1525, Dec. 2019. 2

[21] M M Mesulam. Large-scale neurocognitive networks and distributed processing for attention, language, and memory. *Ann Neurol*, 28(5):597–613, Nov. 1990. 2

[22] 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. 5

[23] 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. *Hum Brain Mapp*, 38(11):5603–5615, Aug. 2017. 1, 2, 3

[24] Mikail Rubinov and Olaf Sporns. Complex network measures of brain connectivity: Uses and interpretations. *NeuroImage*, 52(3):1059–1069, 2010. Computational Models of the Brain. 3

[25] Dorothee Saur, Björn W Kreher, Susanne Schnell, Dorothee Kümmerer, Philipp Kellmeyer, Magnus-Sebastian Vry, Roza Umarova, Mariacristina Musso, Volkmar Glauche, Stefanie Abel, Walter Huber, Michel Rijntjes, Jürgen Hennig, and Cornelius Weiller. Ventral and dorsal pathways for language. *Proc Natl Acad Sci U S A*, 105(46):18035–18040, Nov. 2008. 2

[26] Olaf Sporns and Richard F Betzel. Modular brain networks. *Annu Rev Psychol*, 67:613–640, Sept. 2015. 2

[27] Anika Stockert, Michael Schwartze, David Poeppel, Alfred Anwander, and Sonja A Kotz. Temporo-cerebellar connectivity underlies timing constraints in audition. *eLife*, 10:e67303, sep 2021. 2

[28] Taiji Ueno, Satoru Saito, Timothy T Rogers, and Matthew A Lambon Ralph. Lichtheim 2: synthesizing aphasia and the neural basis of language in a neurocomputational model of the dual dorsal-ventral language pathways. *Neuron*, 72(2):385–396, Oct. 2011. 2

[29] David H. Wolpert. Stacked generalization. *Neural Networks*, 5(2):241–259, 1992. 4

[30] 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. *J Neurosci*, 36(25):6668–6679, June 2016. 2