IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received September 11, 2019, accepted September 30, 2019, date of publication October 8, 2019, date of current version October 22, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2946240

Early Imaging-Based Predictive Modeling of Cognitive Performance Following Therapy for Childhood ALL

RAKIB AL-FAHAD[1], MOHAMMED YEASIN[1], JOHN O. GLASS[2], HEATHER M. CONKLIN[2], LISA M. JACOLA[2], AND WILBURN E. REDDICK[2]

[1]Electrical and Computer Engineering, The University of Memphis, Memphis, TN 38152, USA
[2]St. Jude Children's Research Hospital, Memphis, TN 38105, USA

Corresponding author: Rakib Al-Fahad (ralfahad@memphis.edu)

This work was supported in part by the Electrical and Computer Engineering Department and the Institute for Intelligent Systems, the University of Memphis, and the Departments of Diagnostic Imaging and Psychology, St. Jude Children's Research Hospital, and in part by the National Cancer Institute under Grant R01 CA090246 (PI: Reddick).

ABSTRACT In the United States, Acute Lymphoblastic Leukemia (ALL), the most common child and adolescent malignancy, accounts for roughly 25% of childhood cancers diagnosed annually with a 5-year survival rate as high as 94%. This improved survival rate comes with an increased risk for delayed neurocognitive effects in attention, working memory, and processing speed. Predictive modeling and characterization of neurocognitive effects are critical to inform the family and also to identify patients for interventions targeting. Current state-of-the-art methods mainly use hypothesis-driven statistical testing methods to characterize and model such cognitive events. While these techniques have proven to be useful in understanding cognitive abilities, they are inadequate in explaining causal relationships, as well as individuality and variations. In this study, we developed multivariate data-driven models to measure the late neurocognitive effects of ALL patients using behavioral phenotypes, Diffusion Tensor Magnetic Resonance Imaging (DTI) based tractography data, morphometry statistics, tractography measures, behavioral, and demographic variables. Alongside conventional machine learning and graph mining, we adopted "Stability Selection" to select the most relevant features and choose models that are consistent over a range of parameters. The proposed approach demonstrated substantially improved accuracy (13% – 26%) over existing models and also yielded relevant features that were verified by domain experts.

INDEX TERMS Feature selection, stability selection and control, neurocognitive late effect, graph mining, predictive modeling.

I. INTRODUCTION

Neurotoxicity associated with cancer, radiation therapy, or chemotherapy plays a major role in neurocognitive impairments among survivors due to a disruption of developing neural circuitry. Commonly affected biological mechanisms include: atrophy of grey matter (GM) and/or demyelination of the white matter (WM), suppression of neural progenitor proliferation, microvascular damage, dysregulation of proinflammatory cytokine cascades, oxidative stress, and general vulnerabilities [3]–[6]. A study of Acute Lymphoblastic Leukemia (ALL) survivors and controls revealed reduced gray matter volumes in cortical regions associated with central executive and salience networks, as well as bilateral reductions in the periventricular and subcortical WM volumes [7]. The most relevant study conducted, analyzed 31 ALL survivors and 39 matched healthy controls with a graph metric analysis of the diffusion based structural connectome to demonstrate that ALL survivors had significantly lower small-worldness and network cluster coefficient [8]. Reductions in WM volume in the frontal lobes and significant bilateral reduction in prefrontal cortices have been shown to correspond with lower performances on tests of attention and short-term memory [9]. Diffusion Tensor Imaging (DTI) studies have shown that fractional anisotropy (FA) in right frontal, fronto-parietal, and temporal areas are associated with processing speed [10], [11] and working memory [12]. The differential in FA between patients and controls

The associate editor coordinating the review of this manuscript and approving it for publication was Yongming Li.

was proportional to both IQ and processing speed. Another study of ALL survivors 15 years off therapy and controls, demonstrated higher FA on the left but not the right and worse performance in processing speed and academics [2]. Taken together, data from the broader research literature and our studies suggest that ALL survivors have reduced WM volumes that correspond to decreased structural and functional connectivity within regions of the central executive and salience networks; this decreased connectivity may be associated with deficits in cognitive performance in the domains of processing speed, attention, and working memory.

In this study, our main goal was to develop a multivariate data-driven model of ''cognitive abilities'' of ALL patients from MRI-based volumetric measures, morphometry statistics (e.g., surface area and cortical thickness) from diffusion tensor imaging (DTI), and behavioral as well as demographic variables. We used (i) Wechsler Intelligence Scale: Digit Span Backwards (DSB), (ii) Woodcock-Johnson Tests of Cognitive Abilities: Processing Speed (PS), and (iii) The Behavior Rating Inventory of Executive Function (BRIEF-Working Memory) to measure cognitive abilities (executive functions and processing speed) and categorized them in two categories: below-average and average. Performance below one standard deviation of the mean was considered to be below-average. The DSB and PS scores are performance measures completed by the patients, whereas the BRIEF-Working Memory scores are based on parent reports and were inversely coded relative to the DSB and PS scores such that higher scores indicated more difficulties. All scores were normalized for patient age. For a patient to participate in the neurocognitive testing, the patient must be English speaking or English language dominant.

A plethora of studies (e.g., [8], [13]–[15]) used DTI tractography data and graph-theoretic approach to construct structural networks and compare the topological parameters of the network between ALL patients and healthy controls. Commonly used network properties (features) are clustering coefficient, small-worldness index, characteristic path length, modularity, and nodal clustering. These properties are used to evaluate cognitive abilities based upon the p-value or correlation. For example, Zou *et al.* [13], reported significantly lower small-worldness and network clustering coefficient, in addition to greater cognitive impairments in the ALL subjects.

The statistical testing methods provide concurrence of a fixed hypothesis with the available data points but fail to evaluate all possible hypotheses. Significance testing does not describe how strongly two variables were related. Instead, they say more about how large our supporting sample was. However, parameter estimation from an experiment consists of more useful information, than hypothesis testing [16]. Filter based multivariate feature selection methods use the weighted schemes to select relevant features. But the selection of features is inconsistent and changes with the type of estimator, range of model parameters such as regularization, threshold selection, and hyperparameter tuning. Hence, most of the wrapper-based feature selection methods fail to select consistent and relevant features that are invariant and stable over a range of model parameters.

To understand the non-linear embedding of the data, we performed t-SNE visualization (t-distributed Stochastic Neighbor Embedding [17]) with LDA projection of high-dimensional DTI connectome data (i.e., cortical thickness, average length of all fibers that interconnect Region of Interests (ROIs), and demographic measurements) as shown in the Fig. 2. It is easy to note that data exhibit complex and linearly separable distributions. However, popular machine learning algorithm shows poor performance (accuracy: PS (74%), DSB (62%), BRIEF-Working Memory (62%)). Also, the AUC and F1 score for below-average group remains nearly random guess (from 0.5 to 0). Presence of noise, high correlation among variables, reparation, small number of positive samples, and unbalanced distribution in ALL connectome data prevent further improvements. It was also observed that even the structural connectivity based network features are less informative, discriminative, and unable to describe the variability and structures inherent in connectome data.

To overcome such limitations, we adopted a multivariate, wrapper-based feature selection method called stability selection [18]. It not only works efficiently in the high-dimensional data but also provides finite sample control for some error rates of false discoveries in structure estimation. Besides the error control approach, we also applied Randomized Lasso for feature selection. The stability selection not only significantly improved the model performances (PS (13%), DSB (26%), BRIEF-Working Memory (23%) improvement in accuracy), it effectively reduce the feature dimension and selected features that were verified by the domain expert. It was observed that few demographic variables, morphometry statistics and up-to 8% structural connectivity among ROIs are consistent and relevant features that are invariant and stable over range of model parameters. Besides, stability section ranks the importance of features, hence helps us to interpret the relation between structural brain connectivity and cognitive abilities.

The rest of the paper is organized as follows: In Section II, III we discuss MRI data collection protocol, visualization, and processing, respectively. Subsequently, conventional graph mining approach on DTI connectome are presented in Section IV. Following this, Section V presents detail features selection approach, importance of stability selection and its mathematical interpretation. Modeling, performance evaluation, and empirical analysis are discussed in Section VI, and VII. We discuss the findings from the empirical analysis in the discussion section VIII. Finally, Section IX concludes the paper with lessons learned and a few remarks on future direction.

II. DATA COLLECTION

Survivors of childhood ALL treated on a chemotherapy-only protocol (Total XVI([*NCT*00549848]), were prospectively evaluated ($N = 200$; age on protocol 7.2 ± 4.4 years;

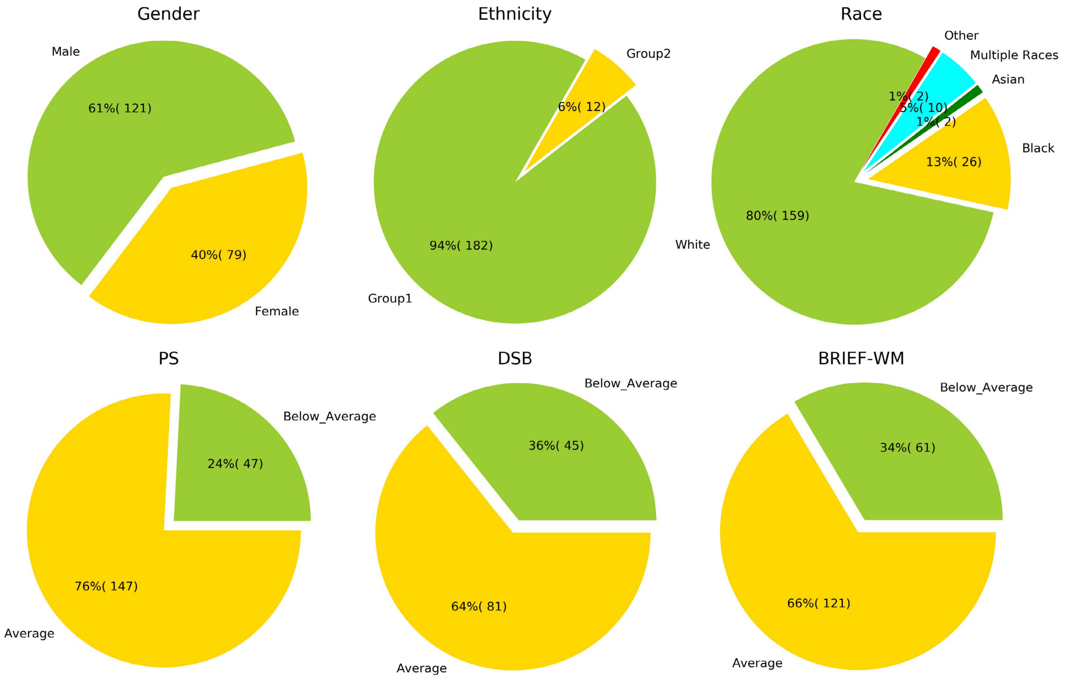


FIGURE 1. **Pie plot shows diversity in the dataset. The dataset has demographic measures of different sex, ethnicity, race and age group. The sample size of the average and below-average group of DSB, PS, and BRIEF-Working Memory measures are relatively unbalanced. Here, ethnicity Group1 represents Non-Hispanic and Group 2 represents not otherwise specified Spanish, Hispanic, Latino group respectively.**

61% male; 96 low-risk, 104 standard/high-risk). Subjects underwent MRI within six months after beginning treatment and neurocognitive testing two years later at the end of all protocol directed therapy. Working memory and decision speed were assessed using:

1) Wechsler Intelligence Scale: Digit Span Backwards (DSB),
2) Woodcock-Johnson Tests of Cognitive Abilities: Processing Speed (PS),
3) The Behavior Rating Inventory of Executive Function (BRIEF-Working Memory).

All MR examinations were performed on a Siemens 3T scanner. A T1-weighted imaging set was acquired with a 3D MPRAGE sequence which provides excellent tissue contrasts among white matter, gray matter and CSF as well as high spatial resolution (1x1x1 mm). Diffusion tensor imaging was acquired with 1.8x1.8x3.0 mm resolution with 12 directions, $a, b = 700$, and 4 averages to increase signal-to-noise. For each scan, the anatomic imaging set was processed using FreeSurfer [19] to obtain the 82 cortical and sub-cortical regions. Cortical thickness measures were evaluated for each of these regions. DTI processing was performed using the FSL FMRIB Toolbox [20]. To establish a reproducible network graph for each exam, probabilistic fiber tracking was then performed using FSL with 500 permutations from each voxel of the anatomic structures. The connection pathway between two nodes, which was the volume in image space that the connection fibers passed through, was extracted for each valid connection using a previously developed adaptation of the probabilistic fiber tracing technique [20]. The mean fractional anisotropy (FA) values of the connection pathway served as the quantitative measure for each edge. All processing was performed in the patient's native space. Overall, the following three types of features were used for model development:

1) Thickness of cortical regions (e.g., thickness of Left Cuneus),
2) DTI measures (undirected, weighted ROI connectivity),
3) Demographic and clinical variables (e.g., Sex, race, ethnicity).

III. DATA VISUALIZATION

The dataset we used was 61% male, 94% non-Hispanic, with the racial categories white, black, Asian, and multiple being 81%, 13%, 1%, and 5% respectively. The feature matrix has 186, 126 and 182 samples for PS, DSB and BRIEF-Working Memory measures. The number of samples in the below-average group is relatively low in our dataset (24 to 36%). Hence, it is relatively unbalanced. Pie plots in Fig. 1 show the data distribution for the different demographic variables and cognitive measures. It should also be noted that while the prevalence of below-average performance is relatively low for this application, it is substantially greater than normative expectation (16%).

Our dataset has 1019 variables overall. Before applying any machine learning algorithm, it is expected to check the assumptions required for model fitting and hypothesis testing. The t-distributed stochastic neighbor embedding or t-SNE [17] is a widely used unsupervised learning algorithm used to visualize high-dimensional data. It converts

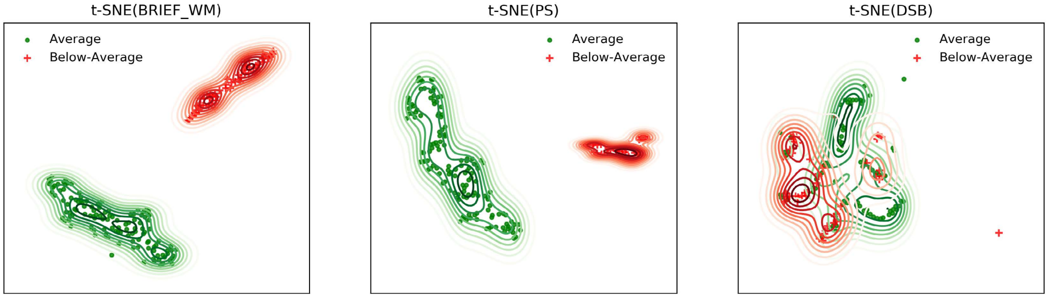


FIGURE 2. The t-SNE embedded higher dimensional features are represented by 2-dimensional scatter and kernel density estimation (KDE) plot. The green lines with dots and red lines with '+' sign represents average and below-average group data, respectively.

similarities between higher dimensional data points to joint probabilities. Thus, this method provides a faithful representation of those data points in a lower-dimensional human interpretable 2D or 3D plane. Such a projection brings insight on whether the data is separable, the data lies in multiple different clusters or inspecting the nature of those clusters. We applied LDA on our two-class dataset and considered 50 dimensions for t-SNE visualization. The LDA based t-SNE approach shows two distinct clusters for average and below-average groups in cognitive measurements. Fig. 2 shows the t-SNE embedded scatter and kernel density estimation (KDE) plot of our data distribution. KDE plot is a non-parametric way to represent the probability density function. Besides, the scatter plot, the KDE plot is used here to visualize the trend of data distribution. The green dot and red '+' sign represent data points for average and below-average groups respectively. It is evident that the distribution of DSB is complex and linearly non-separable. This necessitates the use of stability selection and control to choose features that are relevant and stable.

IV. GRAPH MINING

Cognitive function is supported by distributed neural networks with highly segregated and integrated "small-world" organizations or clusters [21]–[24]. More specifically, those organizations of neurons are densely intra-connected and sparsely inter-connected. We applied graph theory to construct and analyze the brain connectome from DTI data. The $82 * 82$ undirected and weighted adjacency connectivity matrix from DTI FA data is used to calculate 7 basic global network features using BCT tools [25]: (i) Characteristics path, (ii) Global efficiency, (iii) Average clustering coefficient, (iv) Transitivity, (v) Small-worldness, (vi) Assortativity coefficient, and (vii) Modularity (see Appendix for mathematical definitions and interpretation of these network features).

We computed the Wilcoxon rank-sum statistic on network features to find significance across DSB, BRIEF-Working Memory, and PS. There is no significant difference in measurement across groups (except DSB: Transitivity ($p < 0.048$), Global efficiency ($p < 0.030$), Characteristics path ($p < 0.046$)). Those global measurements are based on normalized or averaged versions of clustering and community structure. Therefore, we found a strong correlation between the measurements except for modularity.

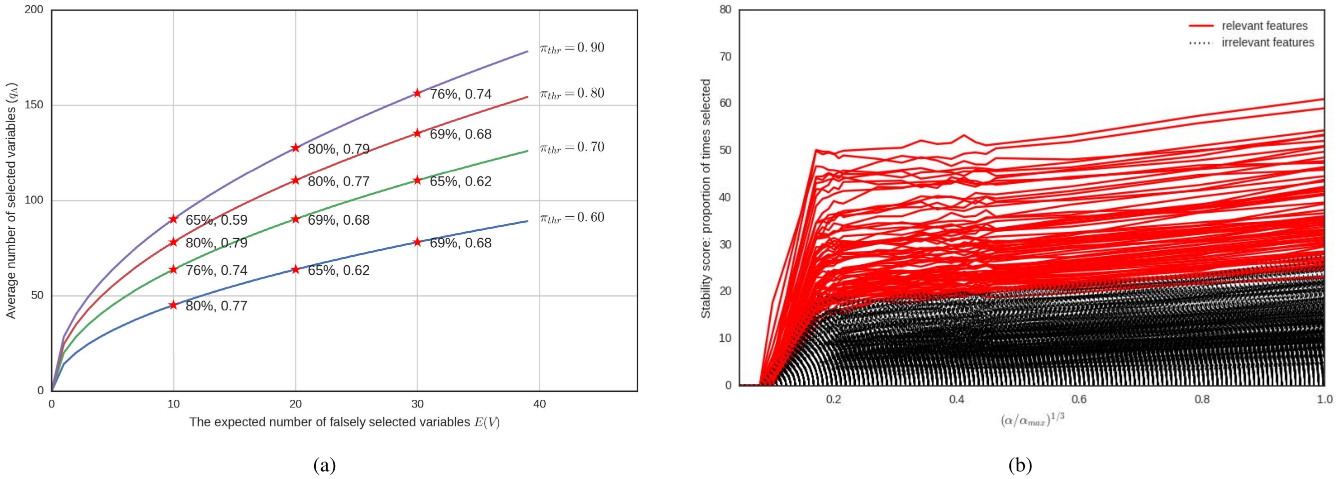
In the next step, we concatenated network features, applied different classifiers and observed their performance (details of parameter tuning and model fitting are explained in the appendix). We also evaluated the performance of the combination of network features and demographic variables (named as ND). The combination of network features and demographic variables with cortical thickness (named as NDI) were also similarly assessed. Summary results of overall empirical analysis are listed in Table 1. It was observed that best classification accuracy among NI and NID features are 74%, 69% and 62% for PS, DSB, and BRIEF-Working Memory respectively. However, AUC scores ($0.5 \sim 0.62$) of this models indicates performances are not better than the random guess. The mathematical definition of network features, p-values, high correlation among features, and poor model performances indicate overfitting, the presence of noise, and repetition. Overall, the model's performances differ from previous studies [8] because of :

1) Less number of trials for average group rather than below-average,
2) Connectivity matrix is highly sparse,
3) Network measurements are in average form. Average value over space matrix with lots of outliers making the features less discriminative,
4) Network features represent global properties rather than local,
5) Few numbers of highly correlated features are used for modeling. Hence there is scope for improvement using multimodal features, feature fusion or decision fusion.

Therefore, we applied conventional machine learning on weighted connectivity matrix with a stability selection. The details of this approach are discussed below.

V. FEATURE SELECTION

Feature selection is used to reduce the dimensionality, improve the estimator's accuracy, and enhance generalizations by reducing overfitting in high-dimensional



(a)

(b)

FIGURE 3. Expected number of falsely selected variable $E(V)$ VS q_Λ graph and stability path for DSB class. Left side of the plot (a) show the variation of $E(V)$ and pi_{thr} on model accuracy. Red solid lines of plot (b) show the relevant features (63) for best $E(V)$ and pi_{thr} and black dotted lines represents stability path for irrelevant features (956) over a range of regularization parameter.

datasets [26]–[29]. Stability selection is a combination of subsampling and high-dimensional feature selection algorithms. Despite its simplicity, it is consistent for variable selection. The main advantages of this algorithm are:

1) It works efficiently in the high-dimensional data with less number of samples,
2) Stability selection provides finite sample control for some error rates of false discoveries and hence a transparent principle to choose a proper amount of regularization for structure estimation,
3) The method is extremely general and has an extensive range of applicability.

A. STABILITY SELECTION WITH ERROR CONTROL

With stability selection, data are perturbed many times and features are selected that occur in a large fraction of the resulting selection sets. Those variables are called the set of stable variables (see Appendix for mathematical definitions). The $L1$ penalized logistic regression was used to discard the feature with zero coefficient and consider the non-zero coefficient over a range of regularization parameters. This process is iterated over multiple times (e.g. 10,000 times). The stability selection with error control (EC) provides an upper bound to the expected number of falsely selected variables $E(V)$. The boundary equation can be defined as [18]:

$$E(V) \leq \frac{1}{2\pi_{thr} - 1} \frac{q_\Lambda^2}{p}. \qquad (1)$$

where, $E(V)$ is the expected number of falsely selected variables, p is the number of variables, q_Λ is the average number of selected variables over a range of regularization parameter Λ. The threshold value π_{thr} in the equation 1 is a tuning parameter.

How to select $E(V)$, π_{thr} and q_Λ?: The influence of π_{thr} parameter in equation 1 is negligible [18], [30]. For a value, ranges from 0.6 to 0.9 results tend to be very similar. The value of $E(V)$ is a design specification and can be controlled at the desired level. For a specific the value of $E(V)$, π_{thr} and regularization parameter λ, the amount of stable features q can be calculated from the equation 1. The stable features are those which enters the regularization path first. For $E(V) = 10$, $\pi_{thr} = 0.70$, the average selected feature is $= 63$. The solid red lines of Fig. 3b are the 63 features that come first in the regularization path, therefore, they can be considered as the most relevant features. The rest of the black dotted lines are irrelevant features. We can see some black dotted lines are mixed with red solid lines, but those can be treated as false positives for error tolerance ($E(V)$).

To find the effect of average number of selected variables (and size of feature matrix) in modeling, we tuned different combination of $E(V)$ and π_{thr} and observed the model performance (The modeling and performance evaluation process is briefly described in modeling section). The points indicated by * in Fig. 3a are accuracy and AUC for different values of q. It is apparent that for a fixed value of $E(V)$ the impact of π_{thr} is similar. On the other hand, variations of $E(V)$ does not change model performance significantly. Though $E(V)$ gives stability selection more freedom or error tolerance, after a certain level, the stability selection starts selecting more noise features, hence degrading the model performance. Though this algorithm allows freedom for error control, the bound has some drawbacks. First, it applies to the population version of the subsampling process. For a data set with small sample size, it is unrealistic to use it in practice. Second, the bound is derived under a very strong exchangeability assumption on the selection of noise variables and a weak assumption upon the quality of the original selection procedure. Shah and Samworth [30] claims that this process is worse than random guessing.

B. STABILITY SELECTION WITH RANDOMIZED LASSO

Stability selection with Randomized Lasso (RL) as an alternative solution of EC [18]. The RL is a straightforward two-step approach. Instead of applying a specific algorithm to the whole data set to determine the selected set of variables based upon the weight of coefficient, RL is applied several times to random subsamples of the data of size $n/2$ and chose those variables that are selected consistently across subsamples. By performing this double randomization several times (e.g., 10,000 times), the method assigns high scores to features that are repeatedly selected across randomizations. In short, features selected more often are considered good features even though the "irrepresentable condition" [31] is violated. This approach is similar to the concept of the bagging [32] and sub-bagging [33] algorithms. RL assigns feature scores between 0 and 1 based on the frequency of selection over 10,000 iterations. We need to specify the score above which features should be selected to find out the best representative stable features. Threshold selection is a design parameter. We varied different the selection threshold (i.e., the number of selected features) and observed the effect on model performance.

Fig. 4 shows the effect of different selection threshold on modeling. The histogram illustrates the distribution of the score. The first line of x label shows the bin ranges of scores (0 to 1), second and third line shows the amount and percent of features that have a nearly same score for a specific bin. It was observed that 53% features have the score of 0 to 0.1. That means, out of 10,000 iterations they were selected between 0 to 10% of the time. For a specific selection threshold, e.g., 0.46, this algorithm selected 29 features. We built a model using those 29 features, which then gave us 89% accuracy (best model performance) with AUC = 0.87 for BRIEF-Working Memory class. The bell-shaped solid black and red dotted lines shows the Accuracy and AUC curves for different selection thresholds. It was observed that the selection threshold higher than the optimal value (0.46) allowed the model to consider more noise variables. Hence, degrading model performance significantly.

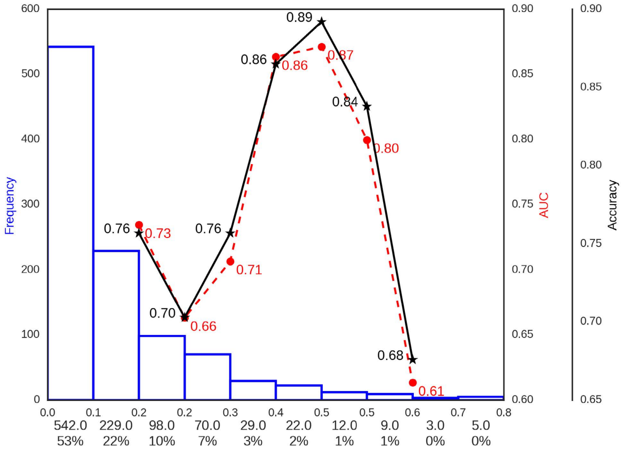


FIGURE 4. Effect on section threshold over model performance for BRIEF-Working Memory prediction. Three lines of x-label represent the range of each bin of features score (range: 0 to 1), number and percent of feature fall in each bin.

VI. MODELING

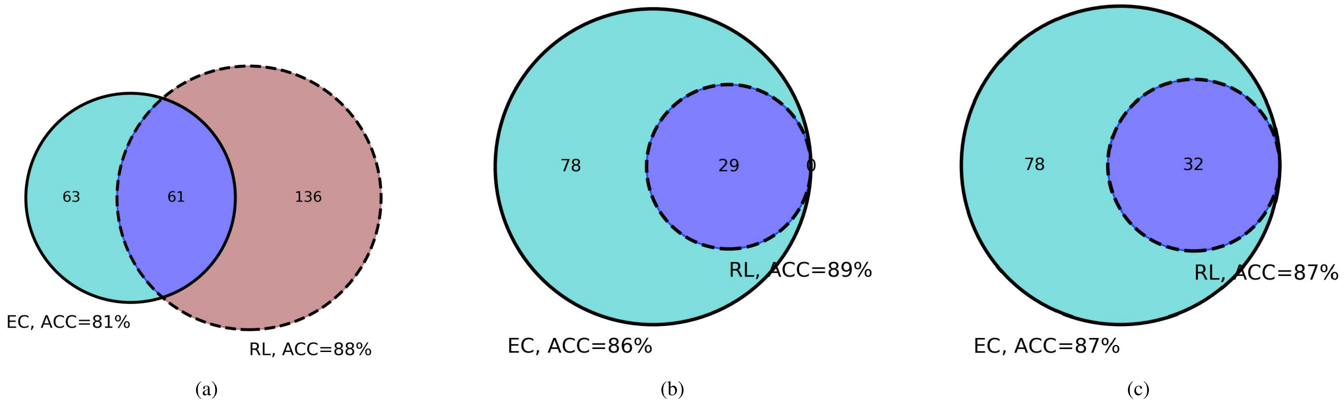
Modeling from selected features has several steps including (i) Use mean imputation to remove missing or NaN values (ii) Apply z-score normalization (Center to the mean and component-wise scale to unit variance) to normalize the data (iii) Test train splitting (80% for training and validation, 20% for testing), (iv) Besides SVM, Random Forest (RF) and Bagging (BAG) classifier are used as estimator. (v) Hyper parameter tuning and model fitting using best estimator and (vi) Performance evaluation. More about modeling are explained in the appendix.

VII. EMPIRICAL ANALYSIS

In this section, we will discuss the processing pipeline and results from the different experiment. The first step was data preprocessing. Missing values (NaN) of the feature matrix were replaced using the mean imputation along the column, and z-score normalization was used for data standardization. Preprocessed and standardized features matrix are then randomly shuffled and split into 80% training and 20% test examples. This testing data was kept unseen and used only for the final model evaluation.

The condition number of a matrix X is defined as the norm of X times the norm of the inverse of X [34]. In short $ConditionNumber = |X| \left|X^{-1}\right|$. The condition number was computed using singular value decomposition and $L2$ normalization. If the condition number is less than infinity, the matrix is invertible. There is no hardbound; the higher the condition number (ill-condition matrix), the greater the error in the calculation. The condition number of our feature matrix is moderately higher for PS and BRIEF-Working Memory than the DSB class (PS: 61.17, DSB: 22.63, BRIEF-Working Memory: 56.11). However, stability selection has the ability to work perfectly on ill-conditioned feature matrix, so we applied it with EC and RL on this training data. For EC, $L1$ penalized logistic regression with 10,000 iterations with 22 regularization parameters (ranges from $10^{-2} \sim 10^{2}$) was used to get the stability path for each class label. In this study, we did not specify the tolerance of error. Hence, we let the empirical analysis to find optimality. Different combination of $E(V)$ and π_{thr} of equation 1 was evaluated to get the best accuracy, AUC, and minimum amount of as discussed in the method section. Our such grid search approach indicates, the optimal $E(V) = 10$ and $\pi_{thr} = 0.8, 0.7, 0.8$ with accuracy 87%, 81% and 86% for PS, DSB and BRIEF-Working Memory, respectively. Overall, 78, 63 and 78 features can be considered stable. Besides EC, RL approach was also evaluated on training data using the same range of regularizations parameter (22 continuous values) over 10,000 iterations. It was observed that RL selected a small subset of feature (except PS class) compared to EC. Overall 32, 136 and 29 features are the optimal number of feature for PS, DSB, and BRIEF-Working Memory class respectively.

Though accuracy and AUC vary for the two-selection method, there is a significant commonality between selection. The Venn diagram of Fig. 5 shows the set of selected and



(a)

(b)

(c)

FIGURE 5. **Vin diagram of EC and RL selected features for (a): DSB, (b): BRIEF-Working Memory, (c): PS class. Cyan, Brown and blue colored circle represent the number of stable features selected by EC, RL and common features among methods. Prediction accuracy and number of selected features are relatively better for RL method. Here ACC represents accuracy.**

common features among methods. The cyan, brown and blue circles represent ES, RL and common features among two methods. As we allowed some errors in selection (E(V) = 10), EC method selected more features (except for PS) than RL method. This method selects nearly 8% features from the feature matrix as stable features. However, RL method selected nearly 4%. Those selected variables are then used to train estimators.

Estimator learning has three steps: (i) Reshape the feature matrix with stability selection (reduce the dimension), (ii) Random shuffle and split the selected feature matrix into 80-20% tainting and validation set and (iii) Iterative grid search approach was used to find the model with the best accuracy.

Same steps are applied for ensemble methods (RF and BAG). Best tuned model is then evaluated on test data. Data processing and modeling pipeline are shown in Fig. 6. Table 1 shows the performance of different methods. It was observed that the performance of RL method is not only better than EC but also selected less number of stable and robust features. On the other hand, SVM shows better performance than ensemble methods.

The best F1 score for average category is greater than or equal to 0.90 for all three estimators, that means SVM with RL has less false negative. On the other hand, the best F1 score for the below-average category is 0.71, 0.86 and 0.87 for PS, DSB, and BRIEF-Working Memory respectively. Though this score for DSB and BRIEF-Working Memory class is at a satisfactory level (less than false positive), as well as the score for the PS class (0.71) with an accuracy 87% has scope for further improvement. The main reason for this poor performance is a fewer number of negative examples. The PS class has only 24% negative examples. The estimators got very few (only 35) training examples after 80-20% split. Therefore, we need more negative instances for further improvement.

Besides directionality redaction and model improvement, Stability selection can be used for interpreting important features and their rank. Class label correlated shadow features get high score even though necessary conditions (regularization parameter or estimator) change. This illustrates their strong and stable relationship with the response (class label).

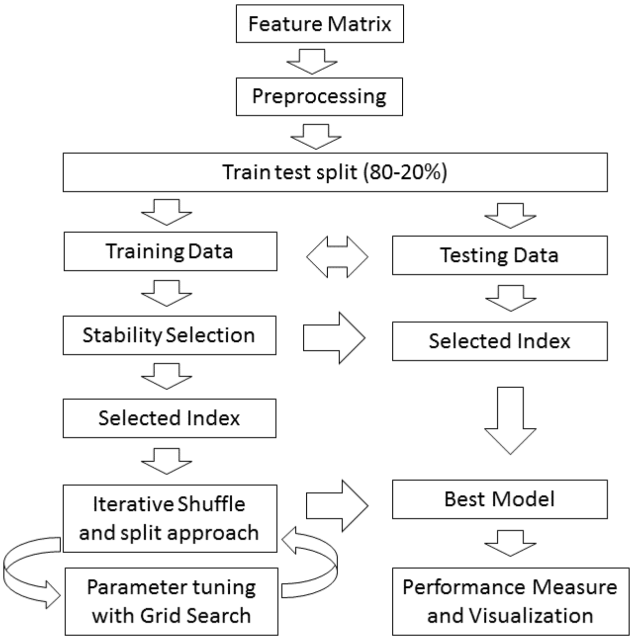
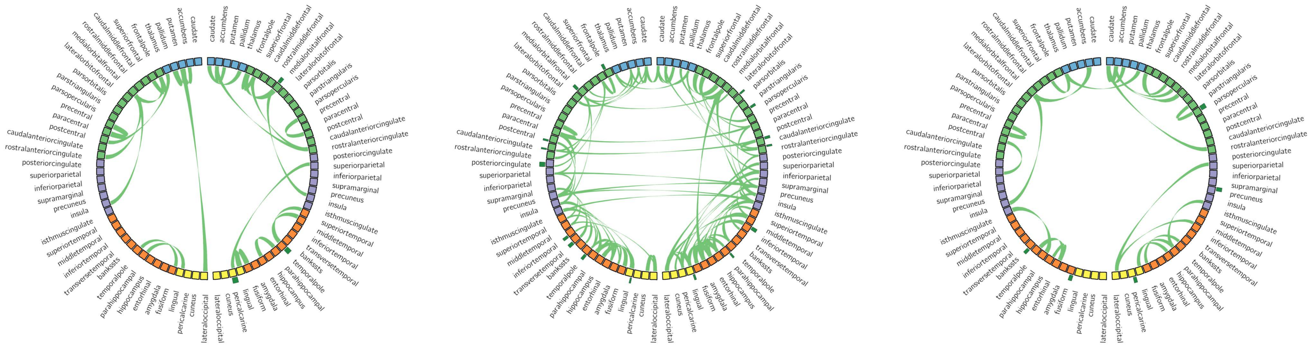

Feature Matrix
Preprocessing
Train test split (80-20%)
Training Data
Testing Data
Stability Selection
Selected Index
Selected Index
Iterative Shuffle and split approach
Best Model
Parameter tuning with Grid Search
Performance Measure and Visualization

FIGURE 6. **Schematic diagram of the processing pipeline. Feature matrix is randomly shuffled and split into 80 and 20% as training and testing data. Feature selection methods (EC and RL) are applied on training data to find the stable features. Those selected features were used to tune and estimator learning using shuffle-split grid search approach, and finally, models are evaluated on test data.**

Fig. 7a, Fig. 7b and Fig. 7c shows the circular visualization of anatomical connectivity between different ROIs that are closely related class labels. Those connectivities are selected by RL method. Left and right sides represent left and right hemisphere accordingly. The width of connection varies with the rank of importance. Similarly, the outer green square represents the related cortical thickness. The width of those

TABLE 1. Overall result of empirical analysis, here All: whole dataset without feature extraction and selection, Net: Network features, ND: Network features and demographic variables and NDI: Network features, demographic variables and cortical thickness, EC: Stability selection with Error Control, RL: Stability selection with Randomized Lasso, AUC: Area Under the Curve, ACC: Accuracy, PS: Processing Speed Cognitive Abilities, DSB: Digit Span Backwards, BRIEF-Working Memory: Behavior Rating Inventory Executive Function class.

Class	Selection Method	Number of Variables	SVM				RF				BAG			
			ACC	AUC	F1		ACC	AUC	F1		ACC	AUC	F1	
					Average	Below-Average			Average	Below-Average			Average	Below-Average
PS	All	1019	74%	0.50	0.85	0.00	71%	0.48	0.83	0.00	66%	0.45	0.79	0.00
	Net	7	74%	0.50	0.85	0.00	74%	0.50	0.85	0.00	74%	0.60	0.83	0.37
	ND	12	74%	0.50	0.85	0.00	74%	0.50	0.85	0.00	71%	0.71	0.83	0.15
	NDI	93	74%	0.50	0.85	0.00	59%	0.49	0.74	0.12	71%	0.48	0.83	0.00
	EC	78	87%	0.78	0.92	0.71	66%	0.45	0.79	0.00	71%	0.50	0.83	0.00
	RL	32	87%	0.78	0.92	0.71	74%	0.50	0.85	0.00	74%	0.50	0.85	0.00
DSB	All	1019	62%	0.50	0.76	0.00	65%	0.57	0.77	0.31	58%	0.49	0.72	0.15
	Net	7	62%	0.50	0.76	0.00	62%	0.54	0.74	0.29	69%	0.62	0.79	0.43
	ND	12	62%	0.50	0.76	0.00	65%	0.59	0.76	0.40	65%	0.55	0.78	0.18
	NDI	93	62%	0.50	0.76	0.00	65%	0.57	0.77	0.31	65%	0.65	0.76	0.40
	EC	63	81%	0.79	0.85	0.74	62%	0.52	0.75	0.17	62%	0.54	0.74	0.29
	RL	136	88%	0.89	0.90	0.86	65%	0.59	0.76	0.40	62%	0.59	0.69	0.50
BRIEF-Working Memory	All	1019	62%	0.50	0.77	0.00	62%	0.53	0.75	0.22	57%	0.48	0.70	0.20
	Net	7	57%	0.46	0.72	0.00	57%	0.46	0.72	0.00	51%	0.43	0.67	0.10
	ND	12	62%	0.50	0.77	0.00	35%	0.31	0.48	0.14	35%	0.31	0.48	0.14
	NDI	93	62%	0.50	0.77	0.00	59%	0.49	0.74	0.12	57%	0.47	0.71	0.11
	EC	78	86%	0.82	0.90	0.78	68%	0.59	0.79	0.33	70%	0.64	0.79	0.48
	RL	29	89%	0.87	0.92	0.85	68%	0.59	0.79	0.33	65%	0.55	0.77	0.24



(a) (b) (c)

FIGURE 7. Circular brain connectivity graph for a): BRIEF-Working Memory, (b): DSB, and (c): PS class using RL method. Left and right side of the circle represents left and right hemisphere. The inner squires, outer squires and green connected lined indicate selected ROIs, cortical thickness of ROIs and connectivity among ROIs, respectively. Shape and the size of the outer square varies with rank (importance) in predicting impairment.

squares varies with their rank as well. It was observed that among 311 possible connectivity among ROIs 29.90% 7.3%, 7.71% anatomical connectivity is important in modeling PS, DSB, and BRIEF-Working Memory respectively.

VIII. DISCUSSION

The hippocampus serves a critical function in long-term memory (LTM), navigation, cognition, and working memory maintenance. An increasing amount of evidence shows that the hippocampus is involved during the processing of spatial and spatiotemporal discontinuity, and relational memory [35], [36]. Specifically, CA1 neurons in the hippocampus are critical for autobiographical memory, autonoetic consciousness, and mental time travel [37]. The medial orbitofrontal cortex is necessary for the coordination of working memory, manipulation, maintenance, and monitoring processes [38]. Stability selection ranked the volume of left and right CA1, left CA2 of hippocampus, right Medial Orbitofrontal (RMO) and right hippocampus as a very important features for working memory classification. Significant lesion ($p < 0.0001$) on those areas are detected in below-average BRIEF-Working Memory group.

On the other hand, memory span is the longest list of items that a person can repeat back in correct order immediately after the presentation. It is a standard measure of short-term memory. Once digit sequence is presented, the participant is asked to recall the sequence in reverse order in DSB related task (to assess working memory). The Posterior Cingulate Cortex (PCC) has a central role in supporting internally directed cognition [39]. The PCC shows increased activity when individuals retrieve autobiographical memories, plan for the future, and regulate the focus of attention [40], [41]. We found the volume of left PCC is a highly ranked feature in DSB classification and significantly lower ($p < 0.02$) for the below-average group. Hence, working memory is strongly related to the volume of the left PCC.

However, CA4 neurons of hippocampus in the perikaryon area, and dendritic branching of both CA4 and CA1 neurons are less in autistic children [42]. We observed that right CA4 and dentate gyrus of hippocampus (CA4-DG) is an essential, highly ranked, significantly distinguishable features to predict PS and BRIEF-Working Memory. The decreased volume in CA4-DG volume is observed in below-average group (PS, DSB, and BRIEF-Working Memory).

Global efficiency is used to find how cost-efficient and fault-tolerant a particular network construction is. We found Global efficiency of DSB related network is significantly ($p < 0.03$) lower in the below-average group. Which indicates patients with below-average working memory are unable to use brain connectivity effectively or adequately. Besides Global efficiency, below-average groups exhibited significantly ($p < 0.04$) reduced transitivity. Lower transitivity indicates the loose connectivity and less potential for integration among nodes to create a clique or complete graph. Hence, the group with the below-average working memory is less likely to have a complex, highly segregated, and densely integrated structural network.

Selected network edges presented in figure 7 shows remarkable connectivity pattern. Highly distinguishable inter-hemispheric connectivity is evident among patient with below-average working memory (DSB network). However, cognitive abilities and working memory-related distinct structural brain network are more intra-hemisphere centric.

Networks presented in Fig. 7a and Fig. 7b contain many of the brain regions known to be associated with executive functioning, including working memory, fluency, and attention. Involvement of the superior and middle frontal regions, the ventrolateral frontal regions of parstriangularis and parsopercularis, anterior cingulate, insula, and superior parietal are critical components of the central executive and salience networks. While these networks form the most reproducible basis for these functions in both fMRI and cognitive neuroscience, the involvement of the temporal lobe regions is also consistent with short-term storage of information for manipulation in the frontal lobes during the working memory tasks.

On the other hand, networks presented in Fig. 7c contain many brain regions, which would be associated with processing speed. Involvement of the superior and middle frontal regions, the orbital frontal regions, anterior cingulate, insula, and superior parietal regions are consistent with regions, which would potentially be engaged during the processing speed tasks. Some of the regions such as the insula and anterior cingulate would be engaged in active switching between tasks such as surveillance and response. While these networks form the most reproducible basis for these functions in both fMRI and cognitive neuroscience, the involvement of the temporal lobe regions is also consistent with short-term storage of information for manipulation during the evaluation process.

Besides cortical thickness and structural connectivity, it was observed that sex, race, and ethnicity are important demographic variables in modeling those cognitive functions.

IX. CONCLUSION

The aim of this study was to develop a data-driven multivariate approach to accurately classify cognitive abilities in ALL patients at the end of therapy. The state-of-the-art cognitive neuroscience mainly uses hypothesis-driven statistical testing to characterize and model neural disorders and diseases, and while these methods provide concurrence of a fixed hypothesis with the available data points, they fail to evaluate all possible hypotheses. In this study, we developed models with stability selection using MRI-based volumetric measures, morphometry statistics and behavioral as well as demographic variables. Stability selection not only reduced feature dimension and improved model accuracy but it selected consistent and relevant connectome features that were invariant and stable over a range of model parameters. This approach also discovered brain regions and structural connectivity which were strongly associated with processing speed and executive functions including working memory, fluency, and attention. The findings of this study suggest that the performance and generalization capability of stability selection based models are superior compared to the classical machine learning and graph mining approach. Since this study was limited to DTI based structural connectivity, it is inadequate in explaining causal relationships among brain regions as well as individuality and variations. Furthermore, a number of possible fMRI based future studies using the same experimental set up with a larger population are necessary for further improvement.

APPENDIX

A. GRAPH MINING

1) CHARACTERISTICS PATH

The characteristic path length is the average shortest path length in the network [43]. Hence, high characteristics path value implies dense connectivity among nodes.

2) GLOBAL EFFICIENCY

Global efficiency is the average of inverse shortest path length, hence inversely related to the average characteristic path length. Global efficiency is used to find, how cost-efficient a particular network construction and how fault tolerant the network is. Hence, high global efficiency implying the excellent use of resources. In brain connectivity analysis, structural and effective networks are similarly organized and share high global efficiency. On the other hand, functional networks have weaker connections and consequently share lower global efficiency [23].

3) AVERAGE CLUSTERING COEFFICIENT

The clustering coefficient of a node is defined as the fraction of triangles around a node [43].The mean clustering coefficient for the network reflects, how close its neighbors are to being a clique or complete graph.

4) TRANSITIVITY

Transitivity is a classical variant of average clustering coefficient. The value of the average clustering coefficient can be influenced by nodes with a low degree. But transitivity is normalized collectively and consequently hence, does not have such problem [24].

5) SMALL-WORLDNESS

Small-world network (S) is formally defined as networks that are significantly densely clustered and have larger characteristic path length than random networks [43]. Mathematically S can be expressed as:

$$S = \frac{\frac{C}{C_random}}{\frac{L}{L_random}}.$$

where C and C_{random} are the clustering coefficients, and L and L_{random} are the characteristic path lengths of the test network and an equivalent random network with the same degree on average respectively. For a small world network $S > 1$, $C \gg C_{random}$ and $L \gg L_{random}$. Such network tends to contain more densely connected cliques/ near-cliques/ sub-networks than random network. Those sub-networks are interconnected by one or more edge.

6) ASSORTATIVITY COEFFICIENT

The assortativity coefficient is a correlation coefficient between the degrees of all nodes on two opposite ends of an edge. A positive assortativity coefficient indicates that nodes tend to link to other nodes with the same or similar degree, on the other hand, negative values indicate relationships between nodes of different degree. Biological networks typically show negative assortativity coefficient as high degree nodes tend to attach to low degree nodes [44].

7) MODULARITY

Modularity refers to the ability of subdivision the network into non-overlapping groups of nodes (known as modules or community) in a way that maximizes the number of within-group edges. Networks with high modularity have dense connections between the nodes within the modules but sparse connections between nodes in different modules. Hence, modularity quantifies the community strength of a test network by comparing the fraction of edges within the community with respect to a random network [45]. It is widely used to discover anatomical modules corresponding to groups of specialized functional area which is previously determined by physiological recordings. Usually, anatomical, effective and functional modules in brain connectivity show extensive overlap [25].

B. FEATURE SELECTION

1) STABILITY PATH

The concept of ‘stability path’ comes from regularization path of regression analysis. A regularization path is defined as the coefficient value β_k^λ of each features of a feature matrix over a range of regularization parameter ($\lambda \in \Lambda$). Let I be a random subsample of $n * p$ feature matrix ($n =$ number of sample and p is the dimension of features) of size $\frac{n}{2}$ is drown without replacement. The random sample size of $\frac{n}{2}$ resembles most closely to the bootstrap [33], [46] which is not worse than random guessing [18]. For every set $K \in (1, \cdots, p)$, the probability of being in the selected set $S^\lambda(I)$ is:

$$\Pi_k = P^*(S^\lambda(I)).$$

For range of regularization parameter Λ and a cutoff threshold π_{thr} with $0 < \pi_{thr} < 1$, the set of stable variables is defined as:

$$S_{stable} = \{\max_{\lambda \in \Lambda} \Pi_k^\lambda \geq \pi_{thr}\}.$$

Fig. 8 shows the stability path of the feature matrix for the DSB class. Each of the black dotted lines represents a stability path of one feature out of 1,019 features, and each dot represents the percent of times it was selected out of all iterations. A red broken vertical line was drawn for the regularization parameter $\alpha = 7$. The * marked point of Fig. 8 indicates the most stable feature that was selected 57% times out of 10,000 iterations.

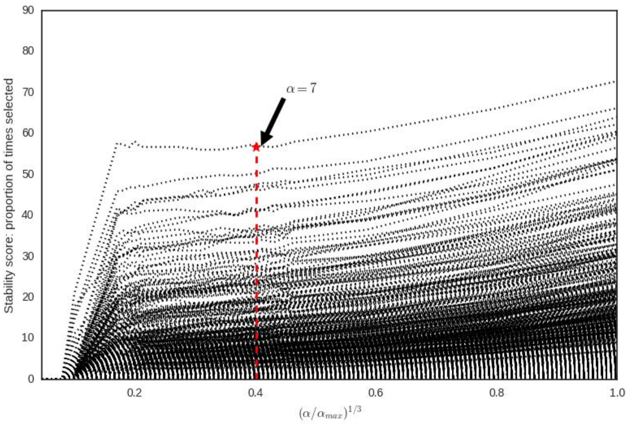


FIGURE 8. **Stability path of features matrix (for DSB) with a range of regularization parameter ($\alpha = 0.01 \sim 100$) as a function of $(\alpha/\alpha_{max})^{1/3}$. The power 1/3 scales the path and enables to visualize the progression along the path.**

2) STABILITY SELECTION WITH RANDOMIZED LASSO

We know Lasso has sparse solutions. For higher dimensional data, many estimated coefficients of variables become zero. Removing the variables can be used to reduce the dimensionality of the data. The limitations of Lasso for feature selection are:

1) Lasso tends to select an individual variable out of a group of highly correlated features,

2) When the correlation between features is not too high, the performance of Lasso is restrictive.

Lasso penalizes the absolute value of coefficients $| \ \beta \ |_k$ of every component with a penalty term proportional to the regularization parameter $\lambda \in \mathbb{R}$. On the other hand, Randomized Lasso penalizes using randomly chosen values in a range $[\lambda, \lambda/\alpha]$ where, $\alpha \in (0, 1)$ is the weakness parameter. The concept of weakness parameter is closely related to weak greedy algorithms. Let W_k be an i.i.d. random variable in a range from $(\alpha, 1)$ for $K \in (1, \cdots, p)$. The estimator of Randomized Lasso can be written as:

$$\check{\beta}^{\lambda,W} = \underset{\beta\in\mathbb{R}^p}{arg\,min} \ \|Y - X\beta\|_2^2 + \lambda \sum_{k=1}^{p} \frac{| \ \beta_k \ |}{W_k}. \qquad (2)$$

Here, Y and X is the class label and feature matrix respectively. Implementation of equation 2 is a straightforward two-stage process:

1) Re-scaling of the feature variables (with scale factor W_k for the k^{th} variable),
2) LARS algorithm is applied on re-scaled variables [47].

In this approach, the re-weighting is simply chosen at random. It is not sensible to expect improvements from randomization with one random perturbation.

For stability selection with RL, we used Randomized Logistic Regression. It works by subsampling the training data and fitting a $L1$-penalized Logistic Regression model where the penalty of a random subset of coefficients has been scaled. We considered sample fraction = 0.75, number of resampling = 10,000 with tolerance = 0.001.

C. MODELING

1) TRAIN-TEST DATA SPLITTING

Learning the parameters of a model (training) and testing it on the same data is a methodological mistake. Such a model would have perfect performance but would fail to predict on unseen data. To avoid such overfitting, the whole data set is divided into (i) training, (ii) validation and (iii) testing. The hyperparameter of the estimator is tuned on the training set and tested on the validation set on a grid-based approach. The best performing tuned estimator is then selected as the final model. The final evaluation can be done on the unseen-test set. However, splitting a small dataset into three parts is a bad idea because it reduces the sample size in each set. On the other hand, without randomization and shuffling each set may contain highly unbalanced samples. A solution to this problem is a procedure called cross-validation (CV). In this approach, test set should still be held out for final evaluation, but the validation set is no longer needed. In the basic approach, called k-fold CV, the training set is split into k smaller sets and an estimator is trained using k-1 of the folds as training data. The resulting model is validated on the remaining part of the data. The hyperparameter is tuned on a grid-based approach. The performance reported by k-fold cross-validation is the average accuracy of all folds.

2) ESTIMATORS

In this study, we mainly used SVM with RBF kernel. However, our data set is unbalanced. The performance of SVM is limited in such a condition, but some machine learning packages use optimization algorithms that can overcome such problems [27], [28], [48]. Ensemble method (EM) is often used to get better predictive performance, generalization or robustness for unbalanced data. Besides SVM, we adopted EM for robust modeling. The most popular ensemble methods are known as 3B or Bagging, Boosting, and Blending. Bagging or Bootstrap Aggregating is an ensemble method that divides data set into smaller parts and build classifier on that dataset. Results of those models are then combined using average or majority voting to get the outcome. Besides SVM we decided to adopt Random Forest (RF), Bagging (BAG) classifier as estimator.

3) PARAMETER TUNING

The performance of an estimator depends on proper hyperparameter tuning. For example, hyperparameters for SVM are C and gamma. Similarly, RF has the number of estimators, maximum depth, minimum samples split, minimum samples leaf, and criterion. On the other hand, BAG has the number of estimators as hyperparameter. In this study, we adopted the iterative shuffle-split approach to overcome overfitting. In each iteration training data is randomly shuffled and split into 80% training and 20% validation data. A pre-specified estimator was trained using different combination of hyperparameters and tested on validation data. The best performing estimator was used as final model. In this study, we used grid based parameter tuning and 80-20% shuffle-split with 10 iterations on training data to find the best predictive model. For example, we tuned 169 (13*13) combination of C and gamma (C: $1e^{-9}$ to $1e^{+3}$, gamma: $1e^{-2}$ to $1e^{+10}$) in 10 iterations. Therefore, the grid search approach finds the best predictive model out of 1690 fit. The Figure 9 shows the learning curve of SVM with RBF kernel for DSB class. For a specific value of C and gamma, the red and green solid line

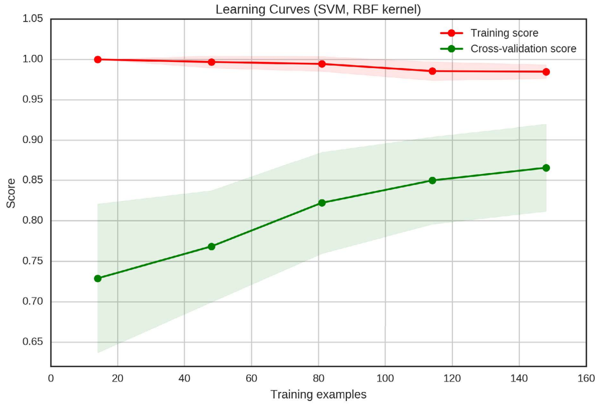


FIGURE 9. **Learning curve for DSB class using random shuffle-split approach. This method reduces the chance of over or under fitting. Green and red lines indicate that the amount of training example increase cross validation accuracy and decrease the training accuracy.**

shows the average accuracy of 10 iterations for different size of training example. The red dotted line indicated the training score where the estimator learned parameter and evaluated on the same dataset. This curve starts with 100% accuracy. With the increase of training example, it starts to degrade. The solid green line represents the shuffle-split (80-20% split and 10 iterations) and cross-validation score for different size of training example. With a small amount of sample, the model accuracy was poor, but an increase of the training sample, decreased not only the width (variance of accuracy) but increased model accuracy significantly.

ACKNOWLEDGMENT

Any opinions, findings, and conclusions recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding institution.

REFERENCES

[1] C.-H. Pui *et al.*, "Treating childhood acute lymphoblastic leukemia without cranial irradiation," *New England J. Med.*, vol. 360, no. 26, pp. 2730–2741, 2009.

[2] M. N. Edelmann, K. R. Krull, W. Liu, J. O. Glass, Q. Ji, R. J. Ogg, N. D. Sabin, D. K. Srivastava, L. L. Robison, W. E. Reddick, and M. M. Hudson, "Diffusion tensor imaging and neurocognition in survivors of childhood acute lymphoblastic leukaemia," *Brain*, vol. 137, no. 11, pp. 2973–2983, 2014.

[3] K. R. Krull, D. Bhojwani, H. M. Conklin, D. Pei, C. Cheng, W. E. Reddick, J. T. Sandlund, and C.-H. Pui, "Genetic mediators of neurocognitive outcomes in survivors of childhood acute lymphoblastic leukemia," *J. Clin. Oncol.*, vol. 31, no. 17, p. 2182, 2013.

[4] R. Seigers and J. E. Fardell, "Neurobiological basis of chemotherapy-induced cognitive impairment: A review of rodent research," *Neurosci. Biobehav. Rev.*, vol. 35, no. 3, pp. 729–741, 2011.

[5] M. Monje and J. Dietrich, "Cognitive side effects of cancer therapy demonstrate a functional role for adult neurogenesis," *Behav. Brain Res.*, vol. 227, no. 2, pp. 376–379, 2012.

[6] F. Altiparmak, M. Gen, L. Lin, and T. Paksoy, "A genetic algorithm approach for multi-objective optimization of supply chain networks," *Comput. Ind. Eng.*, vol. 51, no. 1, pp. 196–215, 2006.

[7] L. Porto, C. Preibisch, E. Hattingen, M. Bartels, T. Lehrnbecher, R. Dewitz, F. Zanella, C. Good, H. Lanfermann, M. Kieslich, and R. DuMesnil, "Voxel-based morphometry and diffusion-tensor mr imaging of the brain in long-term survivors of childhood leukemia," *Eur. Radiol.*, vol. 18, no. 11, p. 2691, 2008.

[8] S. R. Kesler, M. Gugel, E. Huston-Warren, and C. Watson, "Atypical structural connectome organization and cognitive impairment in young survivors of acute lymphoblastic leukemia," *Brain Connectivity*, vol. 6, no. 4, pp. 273–282, 2016.

[9] M. E. Carey, M. W. Haut, S. L. Reminger, J. J. Hutter, R. Theilmann, and K. Kaemingk, "Reduced frontal white matter volume in long-term childhood leukemia survivors: A voxel-based morphometry study," *Amer. J. Neuroradiol.*, vol. 29, no. 4, pp. 792–797, 2008.

[10] D. J. Mabbott, M. Noseworthy, E. Bouffet, S. Laughlin, and C. Rockel, "White matter growth as a mechanism of cognitive development in children," *NeuroImage*, vol. 33, no. 3, pp. 936–946, 2006.

[11] D. J. Mabbott, M. D. Noseworthy, E. Bouffet, C. Rockel, and S. Laughlin, "Diffusion tensor imaging of white matter after cranial radiation in children for medulloblastoma: Correlation with IQ," *Neuro-Oncol.*, vol. 8, no. 3, pp. 244–252, 2006.

[12] D. J. Mabbott, J. Rovet, M. D. Noseworthy, M. L. Smith, and C. Rockel, "The relations between white matter and declarative memory in older children and adolescents," *Brain Res.*, vol. 1294, pp. 80–90, Oct. 2009.

[13] L. Zou, L. Su, R. Qi, F. Bao, X. Fang, L. Wang, Z. Zhai, D. Li, and S. Zheng, "Abnormal topological organization in white matter structural networks in survivors of acute lymphoblastic leukaemia with chemotherapy treatment," *Oncotarget*, vol. 8, no. 36, pp. 60568–60575, 2017.

[14] A. Amidi, S. Hosseini, A. Leemans, S. R. Kesler, M. Agerbæk, L. M. Wu, and R. Zachariae, "Changes in brain structural networks and cognitive functions in testicular cancer patients receiving cisplatin-based chemotherapy," *J. Nat. Cancer Inst.*, vol. 109, no. 12, 2017.

[15] R. K. Chikara, E. C. Chang, Y.-C. Lu, D.-S. Lin, C.-T. Lin, and L.-W. Ko, "Monetary reward and punishment to response inhibition modulate activation and synchronization within the inhibitory brain network," *Frontiers Hum. Neurosci.*, vol. 12, p. 27, Mar. 2018.

[16] H. A. Simon, "How big is a chunk?: By combining data from several experiments, a basic human memory unit can be identified and measured," *Science*, vol. 183, no. 4124, pp. 482–488, 1974.

[17] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.

[18] N. Meinshausen and P. Bühlmann, "Stability selection," *J. Roy. Stat. Soc. B, Stat. Methodol.*, vol. 72, no. 4, pp. 417–473, 2010.

[19] B. Fischl, A. van der Kouwe, C. Destrieux, E. Halgren, F. Ségonne, D. H. Salat, E. Busa, L. J. Seidman, J. Goldstein, D. Kennedy, V. Caviness, N. Makris, B. Rosen, and A. M. Dale, "Automatically parcellating the human cerebral cortex," *Cerebral Cortex*, vol. 14, no. 1, pp. 11–22, 2004.

[20] T. E. J. Behrens, H. J. Berg, S. Jbabdi, M. F. S. Rushworth, and M. W. Woolrich, "Probabilistic diffusion tractography with multiple fibre orientations: What can we gain?" *NeuroImage*, vol. 34, no. 1, pp. 144–155, Jan. 2007.

[21] G. Tononi, O. Sporns, and G. M. Edelman, "A measure for brain complexity: Relating functional segregation and integration in the nervous system," *Proc. Nat. Acad. Sci. USA*, vol. 91, no. 11, pp. 5033–5037, 1994.

[22] D. S. Bassett and E. D. Bullmore, "Small-world brain networks," *Neuroscientist*, vol. 12, no. 6, pp. 512–523, 2006.

[23] C. J. Honey, R. Kötter, M. Breakspear, and O. Sporns, "Network structure of cerebral cortex shapes functional connectivity on multiple time scales," *Proc. Nat. Acad. Sci. USA*, vol. 104, no. 24, pp. 10240–10245, 2007.

[24] M. E. J. Newman, "The structure and function of complex networks," *SIAM Rev.*, vol. 45, no. 2, pp. 167–256, 2003.

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

[26] M. L. Bermingham, R. Pong-Wong, A. Spiliopoulou, C. Hayward, I. Rudan, H. Campbell, A. F. Wright, J. F. Wilson, F. Agakov, C. S. Haley, and P. Navarro, "Application of high-dimensional feature selection: Evaluation for genomic prediction in man," *Sci. Rep.*, vol. 5, May 2015, Art. no. 10312.

[27] R. Al-Fahad, M. Yeasin, A. S. M. I. Anam, and B. Elahian, "Selection of stable features for modeling 4-d affective space from EEG recording," in *Proc. IEEE Int. Joint Conf. Neural Netw. (IJCNN)*, May 2017, pp. 1202–1209.

[28] R. Al-Fahad and M. Yeasin, "Robust modeling of continuous 4-D affective space from EEG recording," in *Proc. 15th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA)*, Dec. 2016, pp. 1040–1045.

[29] R. Al-Fahad, M. Yeasin, and G. M. Bidelman, "Unsupervised decoding of single-trial EEG reveals unique states of functional brain connectivity that drive rapid speech categorization decisions," *bioRxiv*, Jan. 2019, Art. no. 686048.

[30] R. D. Shah and R. J. Samworth, "Variable selection with error control: Another look at stability selection," *J. Roy. Stat. Soc. B, Stat. Methodol.*, vol. 75, no. 1, pp. 55–80, 2013.

[31] P. Zhao and B. Yu, "On model selection consistency of lasso," *J. Mach. Learn. Res.*, vol. 7, pp. 2541–2563, Nov. 2006.

[32] L. Breiman, "Using adaptive bagging to debias regressions," Dept. Statist., Univ. California Berkeley, Berkeley, CA, USA, Tech. Rep. 547, 1999.

[33] P. Bühlmann and B. Yu, "Analyzing bagging," *Ann. Statist.*, vol. 30, no. 4, pp. 927–961, 2002.

[34] N. J. Rose, "Linear algebra and its applications (Gilbert Strang)," *SIAM Rev.*, vol. 24, no. 4, pp. 499–501, 1982.

[35] B. P. Staresina and L. Davachi, "Mind the gap: Binding experiences across space and time in the human hippocampus," *Neuron*, vol. 63, no. 2, pp. 267–276, 2009.

[36] H. Eichenbaum, "Hippocampus: Cognitive processes and neural representations that underlie declarative memory," *Neuron*, vol. 44, no. 1, pp. 109–120, 2004.

[37] T. Bartsch, J. Döhring, A. Rohr, O. Jansen, and G. Deuschl, "CA1 neurons in the human hippocampus are critical for autobiographical memory, mental time travel, and autonoetic consciousness," *Proc. Nat. Acad. Sci. USA*, vol. 108, pp. 17562–17567, Oct. 2011.

[38] A. K. Barbey, M. Koenigs, and J. Grafman, "Orbitofrontal contributions to human working memory," *Cerebral Cortex*, vol. 21, pp. 789–795, Apr. 2011.

[39] M. E. Raichle, A. M. MacLeod, A. Z. Snyder, W. J. Powers, D. A. Gusnard, and G. L. Shulman, "A default mode of brain function," *Proc. Nat. Acad. Sci. USA*, vol. 98, no. 2, pp. 676–682, 2001.

[40] D. A. Gusnard and M. E. Raichle, "Searching for a baseline: Functional imaging and the resting human brain," *Nature Rev. Neurosci.*, vol. 2, no. 10, pp. 685–694, 2001.

[41] B. Hahn, T. J. Ross, and E. A. Stein, "Cingulate activation increases dynamically with response speed under stimulus unpredictability," *Cerebral Cortex*, vol. 17, no. 7, pp. 1664–1671, 2006.

[42] G. V. Raymond, M. L. Bauman, and T. L. Kemper, "Hippocampus in autism: A Golgi analysis," *Acta Neuropathol.*, vol. 91, no. 1, pp. 117–119, Dec. 1995.

[43] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, no. 6684, pp. 440–442, 1998.

[44] M. Piraveenan, M. Prokopenko, and A. Zomaya, "Assortative mixing in directed biological networks," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 9, no. 1, pp. 66–78, Jan./Feb. 2012.

[45] M. Chen, K. Kuzmin, and B. K. Szymanski, "Community detection via maximization of modularity and its variants," *IEEE Trans. Comput. Social Syst.*, vol. 1, no. 1, pp. 46–65, Mar. 2014.

[46] D. Freedman, "A remark on the difference between sampling with and without replacement," *J. Amer. Stat. Assoc.*, vol. 72, no. 359, p. 681, 1977.

[47] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *Ann. Statist.*, vol. 32, no. 2, pp. 407–499, 2004.

[48] R. Akbani, S. Kwek, and N. Japkowicz, "Applying support vector machines to imbalanced datasets," in *Proc. Eur. Conf. Mach. Learn.* Berlin, Germany: Springer, 2004, pp. 39–50.

RAKIB AL-FAHAD is PhD candidate and Research Associate in the Computer Vision, Perception and Image Analysis (CVPIA) Laboratory at The University of Memphis.

MOHAMMED YEASIN, PhD is Professor, Electrical Computer Engineering at the University of Memphis. He leads the CVPIA laboratory. Main thrust of research in the CVPIA lab is in the general areas of computer vision, data mining, bio informatics/computational biology, pattern recognition and human computer interfaces (HCI).

JOHN O. GLASS, MS is the Supervisor of the Structural and Connectivity Imaging Research lab in the Department of Diagnostic Imaging at St. Jude Children's Research Hospital, with a mission to assess the impact of cancer on children by analyzing brain structure and function and its correlation to neurocognitive performance.

HEATHER CONKLIN, PhD Member, St. Jude Faculty and Chief, Section of Neuropsychology. The goal of her research program is to mitigate cognitive deficits following treatment for childhood cancer by identifying risk/resiliency factors and developing empirically supported interventions that ameliorate cognitive late effects.

LISA JACOLA, PhD, ABPP-CN is a board-certified clinical neuropsychologist and an Assistant Member, St. Jude faculty. Her research program seeks to improve neurobehavioral outcomes in childhood cancer survivors. Current projects aim to characterize neurobehavioral outcomes and underlying bio-behavioral mechanisms, identify risk and resiliency factors, and develop and evaluate cognitive interventions.

WILBURN E. REDDICK, PhD Member, St. Jude Faculty in Department of Diagnostic Imaging and Director of Structural and Connectivity Imaging Research. His lab focuses on using MRI to understand the impact of cancer and its treatment on the developing brain by quantifying changes in brain structure and function and its impact on neurocognitive performance.